Interactive Human-in-the-Loop Topic Modeling

Laura Ham and Luis A. Leiva

Abstract Broadcasting companies produce large amounts of text and audiovisual content. Extracting meaningful insights from these sources requires efficient analysis methods, which are often only palatable to data scientists. Even in large organizations, there is a critical knowledge gap: media experts manually curate work to derive insights, which is very time consuming, while engineers can use advanced data science methods but lack the domain expertise to derive key insights from the data. We propose to bridge this knowledge gap with INTEX, a human-in-the-loop interactive topic modeling application. We designed INTEX considering non-technical media experts as the main stakeholders of the application. A user evaluation shows that INTEX enables domain experts to extract and explore topics in an intuitive and efficient manner. Our work illustrates how complex applications can be made more accessible by hiding low-level details and linking these to high-level interpretations. INTEX overcomes past challenges in topic modeling, representing the future of interactive applications in this domain.

Keywords Interactive machine learning · Human-in-the-loop · Topic modelling · Human–computer interaction · Exploratory data analysis

1 Introduction

We are entering a new era where Artificial Intelligence (AI) is incorporated to every computerized system. It is, therefore, more important than ever to understand the interplay of AI systems and humans, such that they can cooperate toward a common goal, acknowledging their weaknesses and leveraging their strengths. This can be framed as a Human–Computer Interaction (HCI) problem, i.e., instead of focusing on

L. Ham
Independent Researcher, Espoo, Finland

L. A. Leiva (✉)
Department of Computer Science, University of Luxembourg, 6, Avenue de La Fonte, 4364 Esch-Sur-Alzette, Luxembourg
e-mail: luis.leiva@uni.lu

S. Sayago et al. (eds.), *Future Perspectives on Human-Computer Interaction Research*,
https://doi.org/10.1007/978-3-031-71697-3_5

maximizing a target metric, such as classification accuracy, we should think instead about how to best support the users of the AI systems. In this regard, a human-in-the-loop approach seems one of the most promising and effective solutions, since HCI has been (and still is) an enabler of human-in-the-loop AI systems. In a nutshell, human-in-the-loop approaches put humans at the center of the AI systems, aiming at handling this aforementioned interplay between humans and systems as optimally as possible.

This book chapter discusses a real-world use case of human-in-the-loop AI, to support broadcasting companies in the production of audiovisual contents. Broadcasting media companies produce large amounts of text and audiovisual content worth of analysis. For example, the Washington Post produces about 500 stories per day [29] and Netflix has 2.2 million minutes of content, or over 50,000 titles, only in the US [31]. Tapping these sources helps to uncover hidden patterns and gain insights to support data-driven business decisions. This requires efficient analysis methods and modeling techniques for automatic theme discovery, among which topic modeling is the most popular one. In a nutshell, topic modeling infers latent structures of large document collections by automatically coding them into a smaller number of semantically meaningful categories.

A shortcoming of classic topic models is that the discovered topics can be hard to interpret [10, 17, 23]. Likewise, extracting too many or too few topics leads to either too general or too specific results [17, 33]. Interactive Topic Modeling (ITM) was introduced to solve these issues, incorporating human expertise in the modeling process [20]. ITM applications allow users to refine extracted topics by, e.g., keyword and document source. These applications are typically used by data scientists, who are experienced in Natural Language Processing (NLP). However, these NLP experts often lack domain knowledge about the data and its high-level interpretation in a business context. At the same time, domain experts in the broadcasting media, like journalists and data analysts, have this broader knowledge about the produced and consumed media content, but usually lack data science skills to develop and use complex topic models.

This knowledge gap between data scientists and domain experts is excruciating, because strategies for thinking and problem solving differ significantly [39] and also because domain experts find it hard to articulate their problems [37]. We propose to fill this knowledge gap with INTEX (INteractive Topic EXplorer), a human-in-the-loop ITM application designed according to Human-Centered Design (HCD) principles [16] in collaboration with non-technical end-users. This is a unique approach; no previous work has considered the stakeholders in designing such an interactive application for the broadcasting media. Our work illustrates how complex applications can be made more accessible by hiding low-level details and linking these to high-level interpretations.

INTEX's user interface follows five steps covering the users' mental model, which we identified in formative user studies: (1) data selection, (2) model configuration, (3) model output evaluation, (4) model refinement, and (5) exploratory data

analysis. The user interface makes high-level model interactions accessible to non-technical domain experts by means of an intuitive design and exploratory visualizations, while low-level complex model details are hidden in the background. Latent theme discovery and further exploratory data analysis of media content are finally accessible to broadcasting end-users like journalists and media planners.

2 Related Work

Topic models aim to reduce the dimensionality of a set of words in a set of documents into a smaller set of interpretable and meaningful themes (most commonly known as topics). As shown in Fig. 1, documents may cover different topics whereas words can be associated with many of those topics.

Classic approaches to topic modeling include Latent Semantic Analysis (LSA) [14] and variations thereof, such as pLSA [18]. While these approaches may create compact semantic representations [45], they are not attractive for real-world use cases because the discovered topics and keywords are hard to interpret. More recently, Latent Dirichlet Allocation (LDA) was shown to discover more descriptive topics [7], however. However, LDA is suboptimal in terms of consistency and convergence [11]. Both model consistency and convergence are important from the user's point of view, as low consistency and slow model convergence lead to bad user experience. Non-negative Matrix Factorization (NMF) overcomes these aforementioned problems [11], leading to outcomes that are naturally interpretable [3, 34] in a computationally efficient way [45], so it is preferred over other topic models in practice.

NMF decomposes the document-term matrix into two matrices: one consisting of n words by k topics and other consisting of k topics by m documents. NMF is often applied in topic modeling applications not only because it leads to the naturally interpretable topics, but also because it is computationally efficient way [45]. Interestingly, NMF is deterministic so user interactions beyond changing static parameters can be easily incorporated via forms or as part of semi-supervised methods.

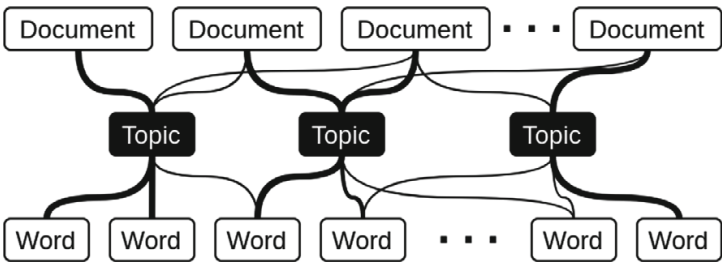

Fig. 1 Topic modeling overview. Documents are associated with one or more topics represented by multiple words. Line thickness indicates the degree to which a topic is represented in each document and word in each topic

2.1 Visualization Techniques

An intuitive representation of the extracted topics, as well as the underlying model, is desired to promote understanding, since in ITM applications the user can control modeling results by direct manipulation. Previous work presented results as word lists [9, 46], word clouds [15], bubble charts [12, 32, 38], and Sankey diagrams [40, 43]. Simple visualizations like word lists and word clouds support a quick initial understanding of topics, while more complex visualizations like Sankey diagrams take longer to understand but reveal better relationships between topics. Smith et al. [42] concluded that there is no 'best' visualization technique for every use case, although, for efficiency and simplicity, simple word lists are the best choice. Lee et al. [25] concluded that topics may be misinterpreted because of the words representing them, and recommended that topic refinement should be focused on topics with low coherence [8].

iVisClustering [24] uses LDA for clustering large document corpora. User revisions include deleting, merging, sub-clustering, and word refinements. It also includes extensive visualizations that allow the user to explore topic modeling results, but requires high cognitive effort and prior knowledge in natural language processing, which makes it less applicable to the broadcasting media.

UTOPIAN [11] uses semi-supervised NMF as topic modeling method, to incorporate user feedback in the matrix factorization process. Quantitative experiments showed that this method outperformed LDA regarding consistency and convergence time. A deterministic (high consistency) and low running time (empirical convergence) of this model are important factors in achieving high user experience. Topic clusters and their relations are visualized using node-link graphs and the t-distributed Stochastic Neighbor Embedding (t-SNE) [27] algorithm. This framework also overshoots the goal of our use case because of the complex visualizations and model refinement possibilities.

ConVisIT [19] allows users to interactively extract topics from asynchronous online conversations. Although the use case (focused on topic modeling of small texts) is different from our current setting, the rich and interactive visualization platform, which allows the user to explore and revise topics on a high level, can be considered as another key prior work in the context of our research.

Finally, ITMViz [36] allows users to incorporate revisions to the LDA model via must-link and cannot-link constraints, a unique revision system not seen in other frameworks. Although these revision possibilities are limited, they showed that constraining topic models by domain knowledge contributes to extracting more meaningful topics.

2.2 Progress Beyond State-of-the-Art ITM

Effective collaboration in ITM requires transparency and predictability [1, 22]. However, there is often a trade-off between the two since high transparency, where model outcomes are easy to validate, expects predictable outcomes and makes it difficult to provide users with suitable controls [44]. Therefore, ITM applications must balance user controls and truly model the data to promote trust in the application [4].

To the best of our knowledge, there is no ITM application for broadcasting media companies. Further, since domain and user expertise largely impact how a topic model is perceived and used, they should be considered in the design of any ITM application. On the one hand, machine learning experts or advanced data scientists most likely have a rough understanding of how a model works, so they can assert its flaws, like unexpected results or instability. On the other hand, domain experts without this background, who have less technical understanding and thus a different perception of the model, are more likely to become frustrated if the model is not adherent, stable, or fast.

Smith et al. [41] noted that future ITM applications should provide a history of actions and model results, support 'undo' actions, have a saving option with reminders to save, allow topic freezing, and support multi-word refinements. In a follow-up user study, Smith et al. [44] found that users dislike latency the most. A lack of adherence, whether the user's input is applied as expected, came out as the second most prevalent dislike. They conclude with four recommendations: users want to be in control, users want speed, (unexpected) model output changes should be explained, and parts of the model should be lockable. With these ideas in mind, we elaborate on the design, implementation, and subsequent evaluation of INTEX.

3 System

Following previous work and the HCD principles, a formative user study with stakeholders was conducted to gather business requirements. Primary, secondary, and tertiary stakeholders were identified; see Appendix 1. Primary users are the potential end-users of the application, in our case domain experts in the broadcasting media. Secondary users are journalists, content producers, and media planners interested in using topic modeling in their decision making processes. Finally, tertiary users are managers, system administrators, and product owners, who are not considered in our research.

Later on, participatory design methods were applied, where the users interacted with a product prototype for evaluation purposes. See Appendix 2 for more details. We performed a 'Wants and Needs' analysis: a fast brainstorming method to gather requirements from multiple users simultaneously [6]. Then, recurring one-hour sessions with focus groups were organized, using three user study techniques: Concept testing, Desirability studies, and Participatory design. The three techniques

are attitudinal, mostly qualitative, and all incorporate hybrid prototype usage during data collection [35].

3.1 Design Choices and Interactions

INTEX's interface is designed according to the mental model of a non-technical end-user. Some screenshots are presented later in Figs. 2, 3, 4, 5, 6 and 7. The workflow of the application is summarized as follows.

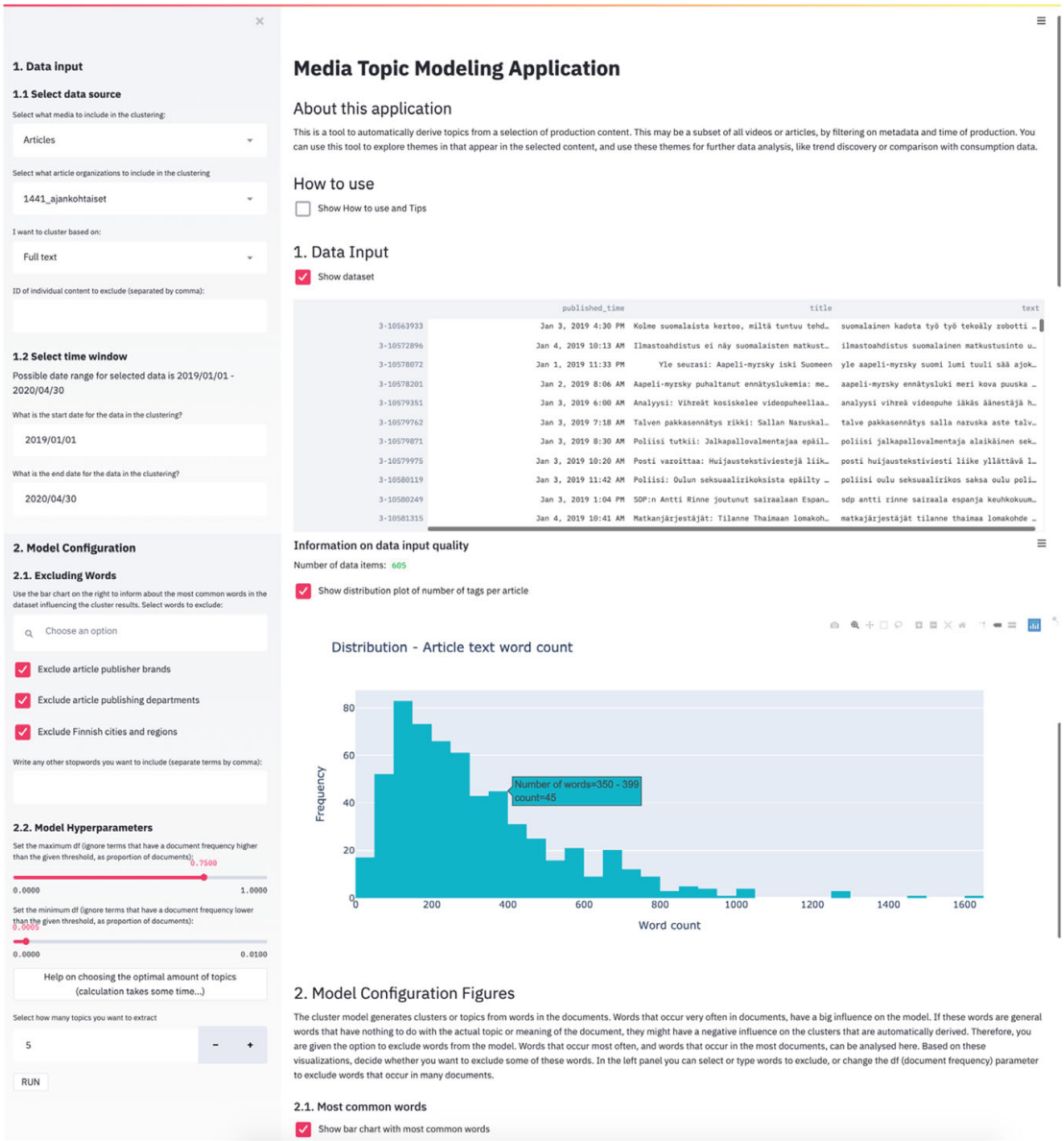

Fig. 2 Screenshot of INTEX's data input and filtering screen

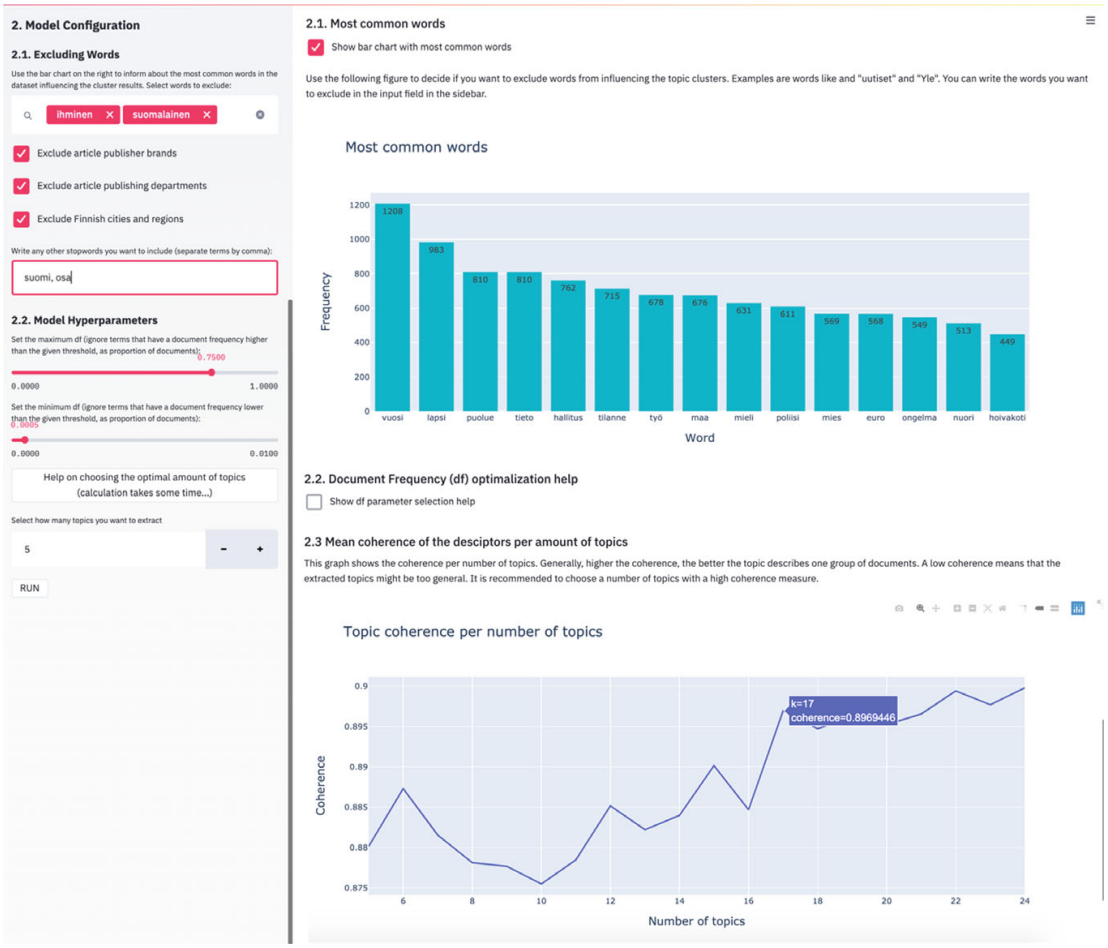

Fig. 3 Screenshot of INTEX's model configuration screen

(i) **Data input selection**. Users can select sources, filter by metadata, and choose a time window. Immediate feedback to user's input is presented either in graphical or tabular form.

(ii) **Model configuration**. The only hyperparameter in INTEX is the number of topics to extract. Since choosing the optimal number of topics beforehand can be challenging, a topic suggestion of 15 topics is provided initially. Topic coherence is computed based on the Word2Vec model [30].

(iii) **Model interpretation and assessment**. Model output is shown as topic-term and document-topic tables, to provide a quick overview of the generated topics. Users can notice what topics are reflected in the input set of documents and can see suggestions on what topics need refinement.

(iv) **Model refinement**. Iterations with focus group sessions resulted in the following set of options: merge, split, and remove keywords from a topic, and rename a topic. INTEX's visualizations reflect the results of these refinements in real time.

(v) **Exploratory data analysis (EDA) in wider context**. Users can see topic development over time and compare topic content production with consumption by

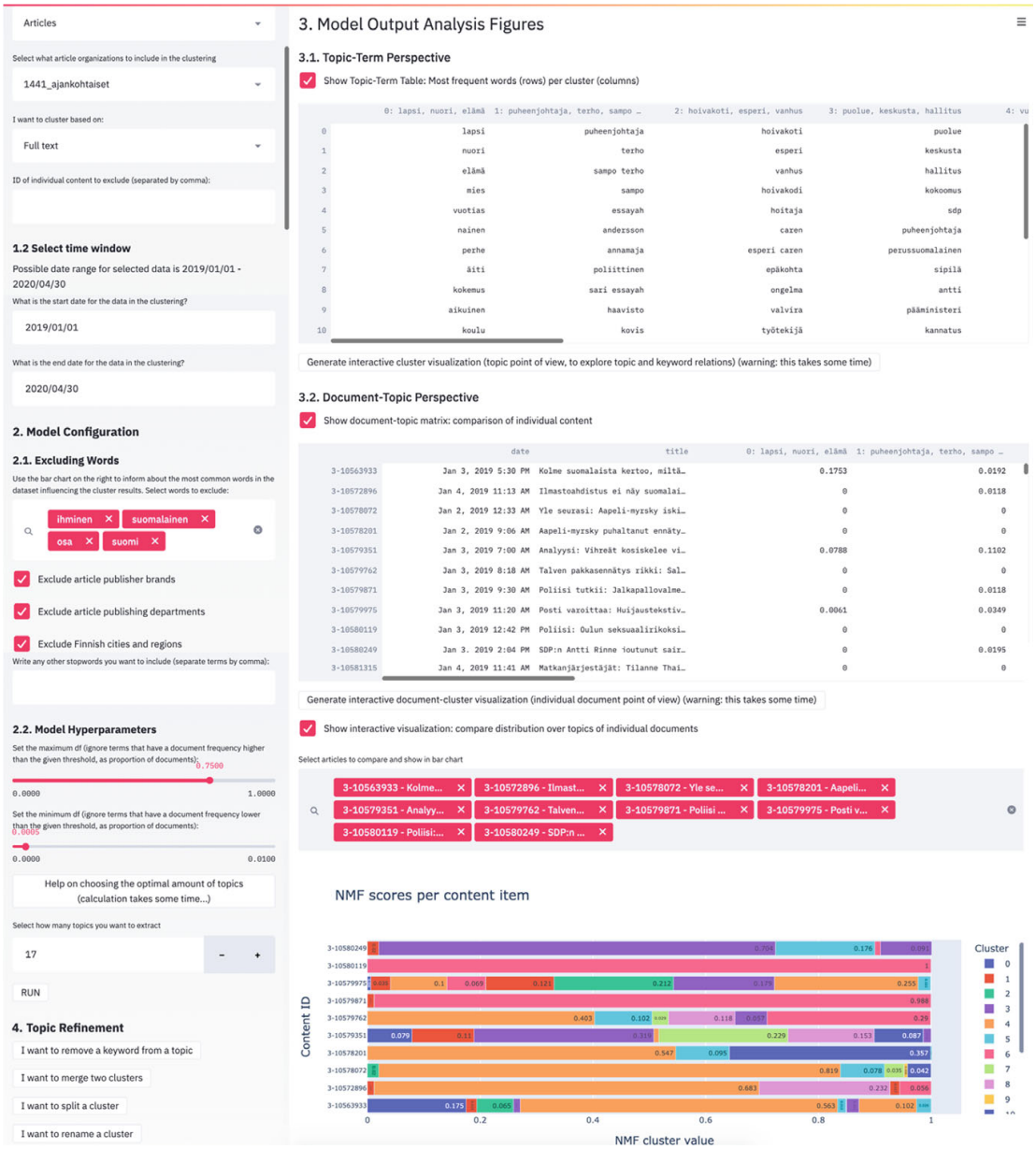

Fig. 4 Screenshot of INTEX's model output screen

different age groups. Data export, including intermediate modeling steps, is also available.

As noted, the workflow in INTEX covers the whole 'user journey', from data selection to exploratory data analysis and exporting the results. The user can follow the five steps explained above both in sequential order or can go back to any earlier step at their own will, such as changing the data input or model configuration after model refinement.

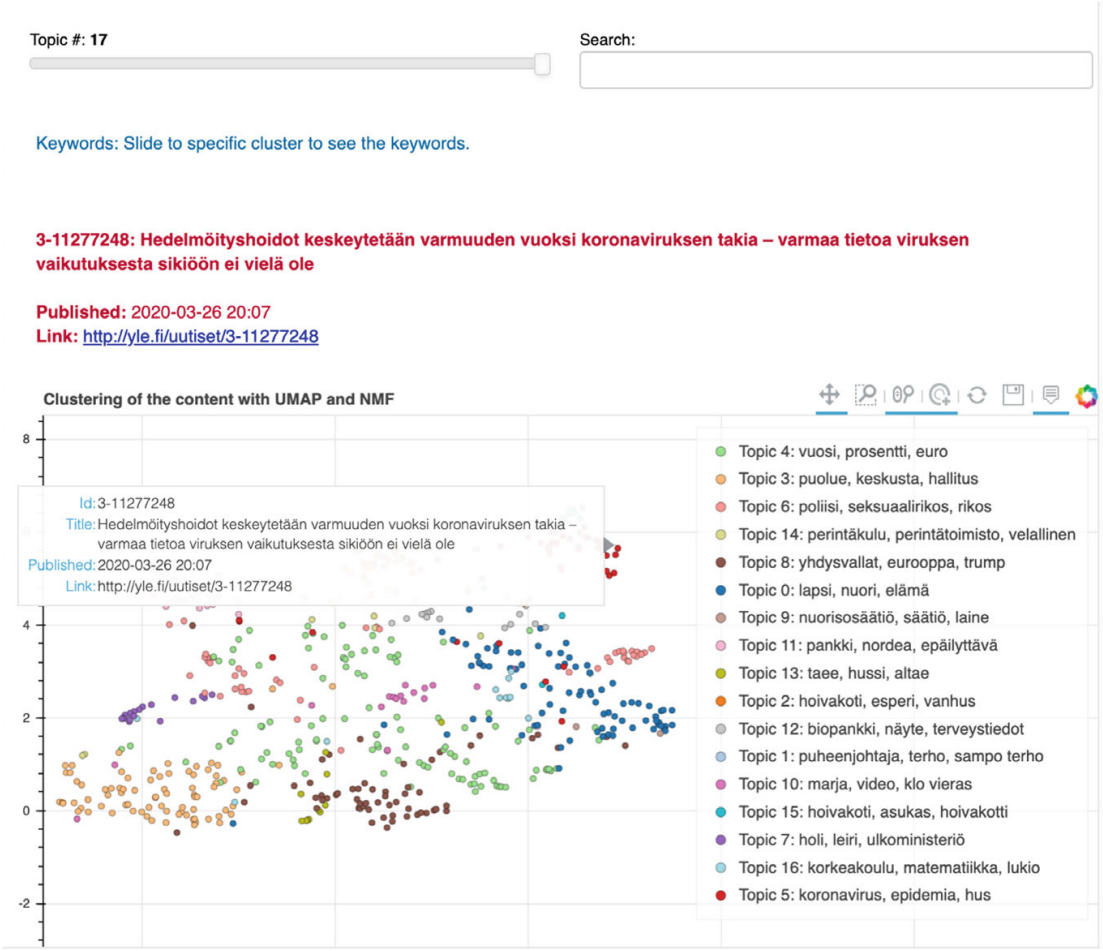

Fig. 5 Screenshot of INTEX's interactive visualizations screen

Fig. 6 Screenshot of INTEX's topic quality screen

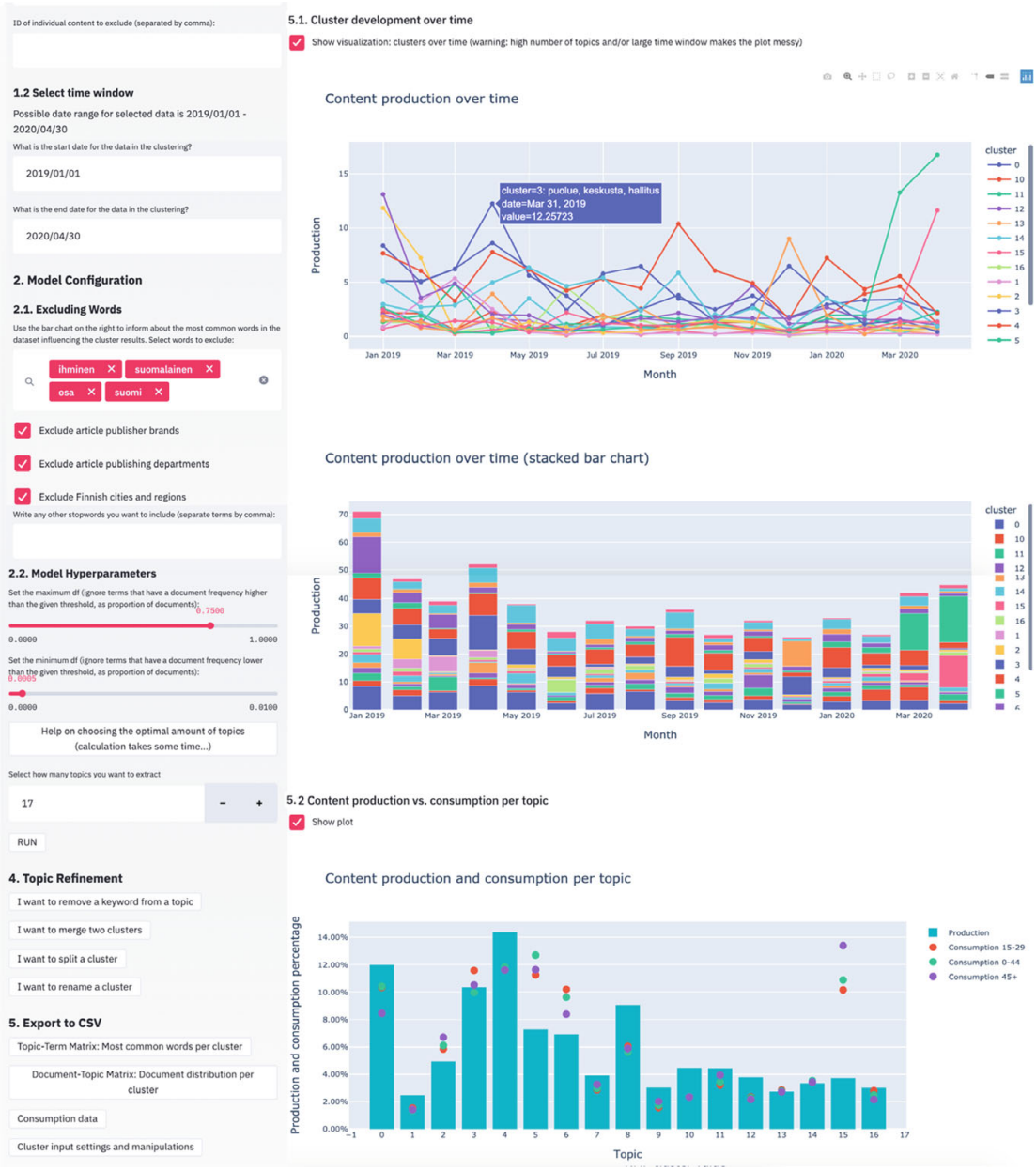

Fig. 7 Screenshot of INTEX's exploratory visualizations screen

3.2 Implementation

INTEX is delivered as a web-based single-page application that interacts with the user by dynamically rewriting the current page with new data from a web server, instead of the default method of the browser loading entire new pages. Fast and smooth transitions between pages make the web application feel like a native desktop-based app.

The backend of INTEX uses the Python library Stanza[1] on top of SpaCy[2] to preprocess text documents and apply NMF for topic modeling. The frontend of INTEX is built with the open-source framework Streamlit.[3] Like Jupyter Notebooks, Python scripts for data modeling are the only assets required to build the user interface; the Streamlit framework creates an interactive and user-friendly interface on top of those. But unlike Jupyter notebooks, which require users to run blocks of Python code, Streamlit hides the scripts in the background and presents an interactive interface to the user.

The user interface has two main modules (Fig. 2). A panel on the right displays information in text, tables, and interactive visualizations. A panel on the left takes in the user actions that influence the topic model. Regarding model configuration (Fig. 3), INTEX includes a bar chart of the 15 most prevalent keywords in the document corpus and a table showing the document frequency of each word. Users can control feature extraction by excluding words via keyword filters and sliders. The left panel allows controlling feature extraction by excluding words, either by selecting a popular word using a drop-down menu with the most common words, typing additional words, clicking on checkboxes to exclude predefined sets of words, or changing the minimum or maximum document frequency using sliders.

The user can also select the number of topics to extract and click on a button to initialize the model (Fig. 4). As soon as the topics are derived, the interface allows the user to interpret the model, refine it, explore the resulting data, and export them. Checkboxes are implemented to toggle tables and figures, and buttons are shown to generate additional, more computationally demanding interactive visualizations. The interactive topic visualization to explore relationships between topics and keywords is only generated on demand. This visualization is made using LDAvis[4] which generates an interactive HTML file from the output of a topic model; see Fig. 5. Relations between individual documents and their topics are visualized with the Bokeh library[5] and the UMAP dimensionality reduction algorithm [28].

INTEX allows the user to get an estimation of the topics quality (Fig. 6), by comparing model residuals per topic, and refine topics on demand. Depending on the refinement action, topics and words can be selected from dynamically generated drop-down menus or typing. Note that topics are usually independent from geographical location. Finally, exploratory data analysis and data export options are also available (Fig. 7). The figures for EDA can be accessed after the model has generated the initial set of topics, but the figures are not displayed in detail initially, since the user should focus on topic evaluation and refinement before diving into how the topics can be utilized in wider context.

---

[1] https://stanfordnlp.github.io/stanza/

[2] https://spacy.io/

[3] https://www.streamlit.io/

[4] https://github.com/bmabey/pyLDAvis.

[5] https://bokeh.org/

4 Evaluation

To derive useful insights about how users perceive and feel about using INTEX, we analyzed the perceived usability (efficiency and satisfaction) and user experience of INTEX via rating scales that were complemented with semi-structured interviews at post-task.

4.1 Dataset

A dataset of 605 news articles were provided by Yle, the Finnish national broadcasting company.[6] Yle's department of Current Affairs ('Ajankohtaiset') selected articles from January 1st 2019 until April 30th 2020. Each article comprises M = 318 words (SD = 231) after text preprocessing. Most articles were familiar to all the participants, thus lowering the barrier to getting introduced to topic modeling.

4.2 Participants

Ten participants (6 female, 4 male) aged 30–39 were recruited from Yle. All participants were Finnish and spoke English fluently. They had various backgrounds regarding data analytics and data science. Seven participants were confident in their ability to use data analytics tools, and the remaining three were self-perceived as neutral. Five participants indicated to have been aware or used some clustering techniques before, but none had done topic modeling.

4.3 Procedure

We conducted individual evaluation sessions that took up to one hour per participant. Each session was conducted remotely with audio and screen-capture recording. Each session started with a walk-through of INTEX. Then, the following task scenario was presented: "You want to make a report about the most important news articles published by Yle's Current Affairs department in the last year. Use INTEX to derive a set of topics that would help you and your target audience to understand the contents that have been covered by such news articles". Participants were instructed to think-aloud during this task. Afterward, a short semi-structured interview was conducted. After the interview, participants completed a survey containing closed questions, addressing their feeling on usability, user experience, and user perception of their interaction with INTEX. See Appendix C for more details.

[6] https://yle.fi/

4.4 Evaluation Measures

We logged task completion time along with the aforementioned post-test survey. Perceived usability was measured in a 1–5 Likert scale (1: 'strongly disagree', …, 5: 'strongly agree') following the System Usability Scale (SUS) [21]. User experience was measured with nine questions adopted from Smith et al. [44]; see Appendix C. The first four questions relate to frustration, trust, task ease, and confidence. The next five questions relate to model adherence, instability, latency, quality, and improvement. Participants answered these questions again on a 1–5 point Likert scale. Finally, the outcomes of the think-aloud protocol were coded thematically.

4.5 Results

Evaluation results are reported in Figs. 8 and 9. Regarding task completion time, participants spent $M = 24$ min (SD $= 7$ min) on the task. This represents a significant improvement of over 50% with regards to manual work, which all participants estimated to be more than one hour at the very least. We should point out that none of the participants had conducted any topic modeling task before, but quickly understood and saw the value of INTEX for their daily work. Regarding system usability and user experience, the average SUS score is 81, which is well above the benchmarked average for websites and web applications [5, 26]. It is worth noting that SUS scores below 50 imply serious usability issues [5], so we can conclude that INTEX is perceived as highly usable.

Participants found the task easy ($M = 4.2$, SD $= 0.8$), trusted INTEX ($M = 4.1$, SD $= 0.7$), felt confident using it ($M = 3.9$, SD $= 0.9$), and did not experience frustration (negative statement, lower is better, $M = 1.8$, SD $= 0.8$). Participants mentioned that INTEX adhered to their input ($M = 4.2$, SD $= 0.9$), had low latency ($M = 4.1$, SD $= 0.9$), and was not unstable (negative statement, $M = 1.8$, SD $= 0.6$).

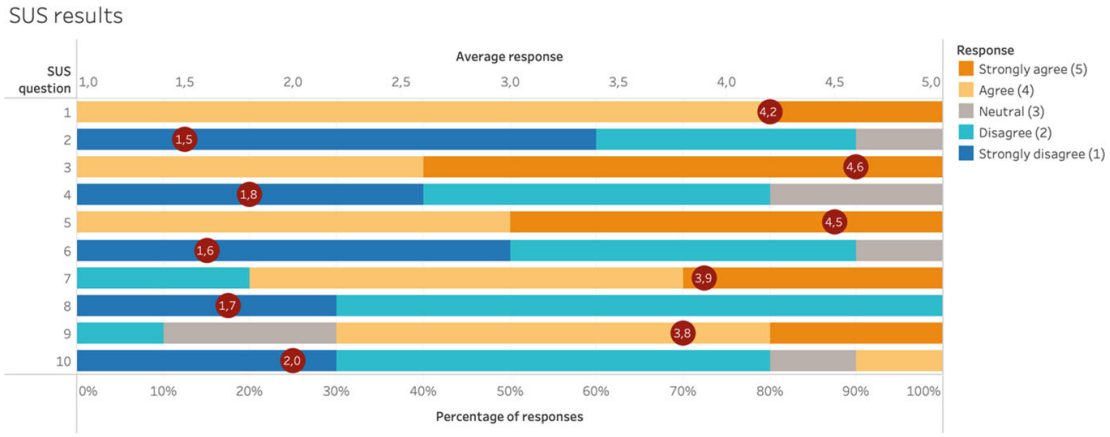

Fig. 8 Results of the SUS questionnaire

Fig. 9 Results of the post-task questionnaire

In addition, participants argued that the final topics were substantially improved over the initial topics ($M = 3.8$, SD $= 0.4$) and most participants were satisfied with the results ($M = 4.2$, SD $= 0.4$).

4.6 Research Findings

In the following, we distil the most relevant findings from the post-task interviews. We hope these will inform other researchers interested in creating ITM applications.

1. **Users do not change the model configuration before an initial run**. Only two participants made changes to the default model configuration before they ran the model for the first time. Three participants made no changes at all. P3 mentioned that they did not change any stopwords or hyperparameters because she wanted to *"see what the model comes up with before manipulating it, so that I can get an idea of how it works and nothing important will be left out"*.
2. **Users are uncertain about the number of topics to extract, but only initially**. Three participants decided to run the model with the default number of 5 and four participants specified from 10 to 20 topics. P5, who chose 10 topics, argued that *"just extracting 5 topics might be too few on this dataset"*, and P6, who chose 20, mentioned that he *"loves to explore a lot of data first to get a general idea on what to specify in the longer run"*.
3. **Users like to explore and try out functionality multiple times**. Seven participants ran the model with at least two different numbers of topics to extract. This suggests that users engaged with INTEX instead of just following the study instructions blindly.
4. **Aligning expectations with system implementation is key, as users' motivation can be quite diverse**. Post-task interviews revealed that individual participants had different interests and use cases in mind, which resulted in differences in user actions and findings. For example, P8 focused a lot on excluding stopwords (*"I don't want common words, that can be part of any subject, to influence my modelling results."*), while P7 used multiple topic refinement operations to *"ensure topics have high quality so I can safely use them later on"*, and P6 spent most of his time exploring the topic results in the EDA plots with topic development over time and consumption rates, which *"helps in micro-segment discovery"*.
5. **Exploratory data visualizations do not only serve as an 'extra' analysis, but also support the users in topic model understanding, evaluation, and refinement**. Four participants have used the visualizations to evaluate individual topic results, and made model refinements based on analyzing these plots. P10 mentioned that she merged two topics on the coronavirus as well as topics on politics, based on trend peaks in the plot of topic development over time: *"The production of articles about politics and elections have a similar development over time, both showing a peak in the spring, so I will merge these two clusters together"*.
6. **Users like the variety and interactivity of the visualizations**. All participants pointed out that they liked the fact that the visualizations are interactive. For example, P6 mentioned: *"When I hover over data in the charts, I see more information on the content and the topic clusters, that is really good"*.
7. **INTEX triggers users' curiosity and encourages them to use the application with other data sources**. When exploring the last plot including consumption data per topic, P5 stated that *"I want to do some analysis on questions that come to my mind. It is great that I can do this analysis with just a few clicks."*
8. **Users like to be in control of the model, but would like to see suggestions and get a sense of its quality**. Users got excited when they see that manual

refinements to the model are implemented and reflected as expected. Users also appreciate recommendations made by the application about which topics to refine. This also holds for customizing model configuration settings, like adding custom stopwords. P8 mentioned that *"it is really handy to see the most common words in the dataset, this makes it intuitive and fast to decide which words to exclude"*.

9. **Users like to see model refinements reflected immediately**. Multiple participants verbally appreciated the results of user and system actions to be directly visible. P3 indicated that *"the value of INTEX is that you get really quickly information on the data"*.
10. **Users without previous modeling knowledge feel confident using INTEX**. Participants indicated that they are amazed by how much they can achieve with this application without having any technical experience. They did not feel the need to completely understand the underlying model, but indicated their confidence in their actions and the final results. P2 said that *"it is fascinating how much you can do with machine learning, although I don't even know how it works"*.
11. **Users like INTEX's user interface**. P5 indicated that *"the left-hand panel stays with the model settings, that is really consistent and really helpful"*. P6 said that he really liked that the interface was designed from a user experience point of view: *"It does not only look good, but it is also very intuitive and functional, it gives a lot of details which are all very understandable"*.
12. **Users like to explore functionalities rather than reading any user manual first**. None of the participants went through the 'how to use' tutorial before they started the modeling task. P10 said: *"I know what the application should be roughly capable of, so I will just explore the functionalities"*. P6 proposed to introduce question marks or info icons with explanations: *"that would help me a lot, because then I do not have to search for the information elsewhere in the application"*.
13. **Users would like to see an 'undo' option**. P1 wanted to revert an action, but was not able to. *"I would like to see an undo option, to revert an action if I made a mistake and want to go back to the previous model results"*. Previous work also flagged the importance of having an 'undo' operation [41], however this is challenging to implement mainly because of the memory cost of keeping track of previous model computations, which are typically very expensive.

5 Discussion

Our results indicate that INTEX indeed helps non-technical end-users with domain expertise perform topic modeling and intervene on the process, without the help of data scientists of NLP experts. According to the Curved Grading Scale of SUS, a score of 81 indicates that the application has a high usability, falling in the top 10% of scores (90th percentile), suggesting that INTEX has 'excellent' performance [26] in

terms of effectiveness, efficiency, overall ease of use, and learnability. Results from the additional questionnaires suggest that INTEX provides a high user experience and that users are satisfied with the underlying model and overall functionality of the application. For example, participants indicated that being able to control the model results by the offered refinement options does not only influence their user experience but also has a positive impact on how they perceived the final results.

INTEX's design follows the HCD framework as well as recommendations from previous research. For example, Smith et al. [44] noticed that users want to be in control, they dislike latency, and that model results should be easily interpretable. Our evaluation results indicate that participants indeed perceived INTEX supporting these aspects. In particular, latency has been mentioned as a major user experience limitation in previous applications [2, 25]. While the overall latency of functions in INTEX is very low, some functions such as calculating model coherence or generating visualizations may take a higher computation time. Since these time-consuming computations are explicitly stated in the interface, participants did not mention latency as a problem. We can conclude that the model internals should be transparent to the user, and the actions and computations needed should be made explicit in the user interface.

Previous ITM applications used complicated plots to visualize topic model output. In INTEX, a combination of both simple and advanced plots is implemented, to give the user the option to explore results in detail. We found that participants demand different visualizations, depending on their background and personal interests. None of the participants indicated that there were too many visualizations, or that visualizations were too complicated. Although our participants had mixed expertise, the user experience and perception results are quite stable. The average standard deviation over all questions is 0.71 on a five-point Likert scale, with the highest standard deviation of 0.9 for statements on user confidence, perceived model adherence, and model latency. Observations of user-model interactions during the evaluation task show that participants use simple topic refinement options and avoid changing complex model settings. Complex model refinement options such as changing keyword order or changing their weights were not implemented because they would be hard to understand for novice users, as suggested by Lee et al. [25]. An unexpected insight is that the EDA visualizations, which were implemented as 'extra' analysis figures, supported users in topic model understanding. This finding suggests that users can improve the model and link the results to data that they are familiar with.

5.1 Limitations

INTEX is unique in its combination of use context, data, users, interface design, topic model, and visualization techniques. Currently there are no existing ITM applications for the broadcasting domain, and previous work has not designed with and for domain experts, so we cannot compare INTEX against any competing ITM application. In addition, previous ITM applications have not been evaluated with real users under

a similar setting like ours, so it is difficult to compare results across domains. This implies that the results of our user evaluation are indicators of the usability and user experience with INTEX only. INTEX is designed to extract and visualize a relatively small number of topics from a set of documents. From our focus group sessions, we found that users would not be interested in extracting more than 15 topics, but later in our evaluation we found that users would benefit from extracting a larger number of topics, up to 50. Currently, visualizations can easily represent results of up to 15 topics, but may become difficult to interpret or slow to generate when used on a larger number of topics.

5.2 The Past, Present, and Future of ITM

The Past. Topic modeling requires efficient analysis methods and techniques for automatic theme discovery, to automatically group multiple documents into a smaller number of semantically meaningful categories. Classic topic models had two important shortcomings that prevented these models for being usable in practice: (1) the discovered topics were hard to interpret and (2) the models were prone to extracting too many or too few topics, leading to too general or too specific results.

The Present. ITM applications address the previously mentioned key shortcomings of classic Topic Models, incorporating human expertise in the process in a way that users can refine extracted topics and do exploratory data analysis. However, most ITM applications are designed for data scientists and NLP experts, who often lack domain knowledge about the data and its high-level interpretation in a business context.

The Future. INTEX represents the future of ITM applications, as it bridges the gap between NLP experts and domain experts in the broadcasting media, like journalists and data analysts, who have a broader knowledge than NLP experts about the produced and consumed media content, but often lack NLP expertise. All in all, any human-in-the-loop application should be designed according to HCD principles in collaboration with non-technical end-users.

6 Conclusion

INTEX is an interactive topic modeling application for media content production analysis that bridges the gap between domain experts and data scientists. As a practical proof, INTEX has enabled professionals from the Finnish broadcasting company Yle to perform topic modeling on their published media content. Based on a formal user evaluation, we can conclude that INTEX is highly usable, promotes an adequate user experience, and is easy to learn by non-technical domain experts. Our findings suggest that INTEX represents the future of Interactive Topic Modeling applications. INTEX is publicly available at https://github.com/laura-ham/INTEX.

Acknowledgements We thank Yle for their support. This work was supported by the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001) and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

Appendix 1: Stakeholders

The following stakeholders were identified:

1. **End-users** of INTEX: domain experts within the broadcasting company. They are data analysts and describe themselves as data-literate, but do not have an engineering or scientific background. Part of their daily job is analyzing data and making statistical reports, with the help of spreadsheets and interactive data visualization tools like Microsoft Excel and Tableau. The national broadcasting company Yle is the domain of all end-users, which means that they interact with content (article, audio and video) production and consumption across different types of media platforms (television, web pages, mobile applications, social media, etc.), and across various types of audiences (with different demographics, media consumption intentions and behavior, etc.).
2. **Content producers and journalists**: data-literate domain experts who are interested in learning on what topics represent their produced media. They do exploratory data analysis on various sources of data and will consume the output of the topic models, but will not directly interact with the interactive topic modeling application itself.
3. **Production decision-makers and media planners**: they mostly want to be informed about new insights and trends that may influence their decision making process. They will not interact with the application itself, but will consume the results of the subsequent data analysis.
4. **Managers** of the direct users: they play a role from a business point of view. They make decisions based on data reports, and thus benefit from the results of INTEX, given that new insights can be gained. In addition, managers may decide on where their teams spend time on, and thus whether an ITM application may be used in the first place. It is important to assure buy-in from this group of stakeholders.
5. **System administrators and IT**: they offer the infrastructure for deploying and maintaining the product.
6. **Data scientists and engineers**: they are responsible for the technical model, implementation and improvements of the application. Technical questions from all stakeholders about the product will be taken by this group of people.
7. **Product owners**: they are responsible for the final product, its maintenance, and marketing operations.

The first group, the direct end-users of the application, is the only primary stakeholder group. The data-literate producers and journalists mentioned in the second

group are secondary users, since they do not interact frequently with the product itself. This group rather uses the product through an intermediary, which is the first user group mentioned. Then, the other stakeholder groups are tertiary users. These users are affected by the system and/or decision-makers. Note that, in this context, individual people can belong to multiple stakeholder groups, since people may have multiple roles in the company. For example, a content producer (group 2) may also be a production decision maker or media planner (group 3). Designing the application is done with taking the different stakeholder groups in mind, rather than different individual persons, since their roles, and thus desires and needs, can be ambiguous.

Appendix 2: Focus Group Sessions

2.1. First Focus Group Session

The first focus group consisted of two potential end-users and one secondary user. One end-user is an analyst at the Yle News Lab,[7] a testing laboratory for data journalism. Another primary user is head of Yle's general data analytics team. The secondary user is a web producer at Yle's department of current affairs. None of the participants have worked with any topic modeling application tool before. The session took one hour, and was held in-person.

After an initial, open discussion on opportunities and wishes from the participants, various models, interaction, visualization and design possibilities were proposed. This informed the participants what was possible from a technical point of view. Participants indicated their preferences among the options provided, which led to a richer set of requirements.

User requirements were extracted from the first focus group session, and also from iterations and theoretical background [13]. Requirements were categorized as application architecture (business), technical topic model implementation (system), and user interface and interactions (design). In addition, we identified the following set of attitudes and desires that should be considered in the design of ITM applications:

- End-users would like to perform the topic modeling 'every few weeks to months', but multiple times with various subsets of a bigger dataset of articles. They want to filter on metadata to achieve this, for example the type of the articles, publishing department of the articles, etc.
- End-users are able to spend time on the total modeling process, from opening to closing the application after completion. As long as the end-user is able to extract meaningful and accurate topics, it is worth the time to explore and refine results.
- The preferred number of topics or themes to extract lies between 5 and 15. This was later updated to 5–50 topics, after conducting interviews with other stakeholders.

[7] https://newslab.yle.fi/

- End-users would like to see some visualization to quickly determine what stop-words to exclude. It was mentioned that it is hard to come up with words to exclude from modeling without seeing which words influence the modeling process.
- End-users and secondary users (the 'data-literate producers') would like to see preliminary visualizations on how the extracted topics can be used. This helps them determine if the extracted topics are meaningful. A visualization was proposed on how topics are represented in the set of input documents over time, and participants would like to see it included in the final application.
- End-users are willing to spend time and effort to interact with the model for the best results, but pointed out that they would rather refine settings that are proposed by the model than manually set parameters from the beginning. This applies to model parameters as well as topic labeling. As a result, the application should propose settings and output labels, and give the user the opportunity to edit these proposed settings and labels.
- The design of a product is not limited to the requirements of the end-users. It became clear that the results of a topic modeling application are likely to be used by the secondary users. Although this user group is not interacting with the modeling application itself, the analysts are using the results of the model, the extracted topics, in exploratory data analysis. Since the primary goal of this group is to gather new insights from this analysis of media topics combined with other data sources, it is important that extracted topics are accurate, meaningful and easy to understand. Additionally, the learned topics should be extracted to a data format that is easy to handle by other data analysis software.

2.2. Recurring Focus Group Sessions

Recurring sessions with two focus groups followed the first focus group session and the first design iteration:

1. The same focus group as the initial user requirement research, with two end-users and one secondary user. Sessions with this group were held every one or two weeks for the course of the two-month design and development cycle. This group had a very active role in the participatory design process. Also concept testing and desirability study methods were applied in some sessions.
2. A focus group with three end-users, one secondary user, and two tertiary users. Sessions with this focus group were less frequent, and started later in the design process. Sessions were less oriented to the design of the platform, but more toward concept and usability testing. Although only a small amount of new requirements was extracted from this group of participants, they confirmed findings from the other focus group sessions. In addition, addressing a second focus group leads to a design for a wider target group and avoids design fixation.

No strict protocol was followed during these sessions. Study methods that were applied during these sessions varied, depending on the design phase in the HCD cycle. In the beginning of the design and development process, sessions were more

focused toward participatory design, whereas later more desirability studies were conducted.

An example of a **participatory design** session is finding out whether and which model revision techniques are most desired. A prototype of INTEX was presented to the participants. Participants were asked to interact with a high-fidelity prototype of the first part of INTEX, in which no revision or interaction was possible. The users were thus presented with a topic modeling output, and were asked whether they would like to make any changes to the result, and which changes. After an initial discussion on the users' wishes, the range of possibilities were presented, to give the participants the opportunity to think about those as well.

An example of applying the **concept-testing method** in a focus group session is introducing possible topic model visualizations to the end-user. In early focus group sessions, users were presented with different types of (graphical and textual) visualizations that represent (part of) the model output. Participants were asked to give their preference on visualizations, which they would most use and benefit from during the interactive modeling process. This method was used to understand if users would want or need specific visualizations.

Finally, **desirability study** techniques were applied in some focus group sessions as well. Although desirability studies are principally carried out to find out what visual design alternative is preferred, the method was applied here to find out the most desired option in a set of presented designs. An example is on interactive versus non-interactive visualizations. Data plots that represent the same data but differ in whether they provide interactivity (for example selecting a subset of data to display), were both presented to users, after which they were asked to indicate their preference. Additionally, the following was observed:

- From exploratory data analysis with topic modeling results during testing sessions, the stakeholders showed most interest in two data analysis visualizations. First, how different topics in produced media content relates to its consumption over different age groups. Second, how the representation of topics in produced media develops over time.
- Although wishes, desires, and requirements from both focus groups were mostly complementary, sometimes they conflicted. An example is the number of topics to extract. Group 1 indicated that extracting up to 15 topics would suffice, while group 2 indicated interest in extracting a larger number of topics (close to 50). The possible number of topics influences how the model should be configured for optimal results and also influences the visualizations. For example, increasing the number of extracted topics to 50 makes the visualizations that plot topic development over time a bit messy, while up to 15 topics are much more clear to distinguish in a single plot. Conflicting requirements like this were solved by discussing the concerning requirement with both focus groups, and a decision was made while keeping the technical requirements and capabilities in mind.
- The iterative nature of the design and development process also led to revisions of requirements within the groups. While this is the advantage of HCD, this may also lead to conflicting desires, and requires a well thought out decision. This may

slow down the development, but we believe that this will only benefit the final product.

Appendix 3: Evaluation Study Questionnaires

In the following we provide the list of questions presented to the participants regarding the user experience evaluation and the post-task interviews.

3.1. User Experience and User Perception Questionnaire

1. Using this application to perform the task was frustrating.
2. I trusted that the application would update the clusters of the articles well.
3. It was easy to use this application to perform the task.
4. I was confident in my specified changes to the tool.
5. How satisfied are you with the final topics?
6. How do you think the final topics compare to the initial suggested topics?
7. After my changes, the application updated fast enough.
8. The tool made the changes I asked it to make.
9. The tool made unexpected changes beyond what I asked to make.

3.2. Post-task Semi-structured Interview Questions

1. Why did you choose this number of topics?
2. What do you think about the on-page explanations?
3. Did the visualizations provide you with sufficient information?
4. Was every function and visualization clear?
5. Does this application help you in discovering new insights or new use cases?
6. Are there functions, visualizations, or data that where not implemented but you would like to have seen or used?

References

1. Abdollahi B, Nasraoui O (2018) Transparency in fair machine learning: the case of explainable recommender systems. In: Human and machine learning
2. Arapakis I, Bai X, Barla Cambazoglu B (2014) Impact of response latency on user behavior in web search. In: Proceedings of SIGIR
3. Arora S, Ge R, Moitra A (2012) Learning topic models–going beyond SVD. In: Proceedings of FOCS
4. Bakharia A, Bruza P, Watters J, Narayan B, Sitbon L (2016) Interactive topic modeling for aiding qualitative content analysis. In: Proceeding of the CHIIR
5. Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 24(6)
6. Baxter K, Courage C, Caine K (2015) Understanding your users: a practical guide to user research methods. Morgan Kaufmann

7. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3
8. Bouma G (2009) Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the GSCL
9. June-Barlow Chaney A, Blei DM (2012) Visualizing topic models. In: Proceedings of the AAAI
10. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D (2009) Reading tea leaves: how humans interpret topic models. In: Proceedings of the NeurIPS
11. Choo J, Lee C, Reddy CK, Park H (2013) Utopian: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans Vis Comput Graph 19(12)
12. Chuang J, Manning CD, Heer J (2012) Termite: visualization techniques for assessing textual topic models. In: Proceedings of the ACL workshops
13. Clegg D, Barker R (1994) Case method fast-track: a RAD approach. Addison-Wesley Longman Publishing Co., Inc.
14. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Assoc Inf Sci Technol 41(6)
15. Gardner MJ, Lutes J, Lund J, Hansen J, Walker D, Ringger E, Seppi K (2010) The topic browser: an interactive tool for browsing topic models. In: Proceedings of the NeurIPS workshops
16. Goodwin K (2009) Designing for the digital age: how to create human-centered products and services. Wiley
17. Greene D, O'Callaghan D, Cunningham P (2014) How many topics? Stability analysis for topic models. In: Proceedings of the ECML-PKDD
18. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the UAI
19. Hoque E, Carenini G (2015) Convisit: interactive topic modeling for exploring asynchronous online conversations. In: Proceedings of the IUI
20. Hu Y, Boyd-Graber J, Satinoff B, Smith A (2014) Interactive topic modeling. Mach Learn 95(3)
21. Jordan PW, Thomas B, McClelland IL, Weerdmeester B (1996) Usability evaluation in industry. CRC Press
22. Kulesza T, Stumpf S, Burnett M, Wong WK, Riche Y, Moore T, Oberst I, Shinsel A, McIntosh K (2010) Explanatory debugging: supporting end-user debugging of machine-learned programs. In: Proceedings of the IEEE VL/HCC
23. Lau JH, Newman D, Baldwin T (2014) Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: Proceedings of the EACL
24. Lee H, Kihm J, Choo J, Stasko J, Park H (2012) iVisClustering: an interactive visual document clustering via topic modeling. In: Computer graphics forum, vol 31
25. Lee TY, Smith A, Seppi K, Elmqvist N, Boyd-Graber J, Findlater L (2017) The human touch: How non-expert users perceive, interpret, and fix topic models. Int J Hum Comput Stud 105
26. Lewis JR, Sauro J (2018) Item benchmarks for the System Usability Scale. J Usabil Stud 13(3)
27. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9
28. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426
29. Meyer R (2016) How many stories do newspapers publish per day? Available at theatlantic.com. Accessed Jan 2022
30. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proceedings of the ICLR workshops
31. Moore K (2020) How long would it take to watch all of Netflix? Available at whats-on-netflix.com. Accessed Jan 2022
32. Murdock J, Allen C (2015) Visualization techniques for topic model checking. In: Proceedings of the AAAI
33. O'callaghan D, Greene D, Carthy J, Cunningham P (2015) An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 42(13)
34. Paatero P, Tapper U (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5(2)
35. Rohrer C (2014) When to use which user-experience research methods

36. Saeidi AM, Hage J, Khadka R, Jansen S (2015) ITMViz: interactive topic modeling for source code analysis. In: Proceedings of the ICPC
37. Sedlmair M, Meyer M, Munzner T (2012) Design study methodology: reflections from the trenches and the stacks. IEEE Trans Vis Comput Graph 18(12)
38. Sievert C, Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the ACL workshops
39. Simon S, Mittelstädt S, Keim DA, Sedlmair M (2015) Bridging the gap of domain and visualization experts with a liaison. In: Proceedings of the EuroVis, vol 2015
40. Smith A, Chuang J, Hu Y, Boyd-Graber J, Findlater L (2014) Concurrent visualization of relationships between words and topics in topic models. In: Proceedings of the ACL workshops
41. Smith A, Kumar V, Boyd-Graber J, Seppi K, Findlater L (2018) Closing the loop: user-centered design and evaluation of a human-in-the-loop topic modeling system. In: Proceedings of the IUI
42. Smith A, Lee TY, Poursabzi-Sangdeh F, Boyd-Graber J, Elmqvist N, Findlater L (2017) Evaluating visual representations for topic understanding and their effects on manually generated topic labels. Trans Assoc Comput Linguist 5
43. Smith A, Malik S, Shneiderman B (2015) Visual analysis of topical evolution in unstructured text: design and evaluation of TopicFlow. In: Applications of social media and social network analysis
44. Smith-Renner A, Kumar V, Boyd-Graber J, Seppi K, Findlater L (2020) Digging into user control: perceptions of adherence and instability in transparent models. In: Proceedings of the IUI
45. Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D (2012) Exploring topic coherence over many models and many topics. In: Proceedings of the EMNLP
46. Zou C, Hou D (2014) LDA analyzer: a tool for exploring topic models. In: Proceedings of ICSME