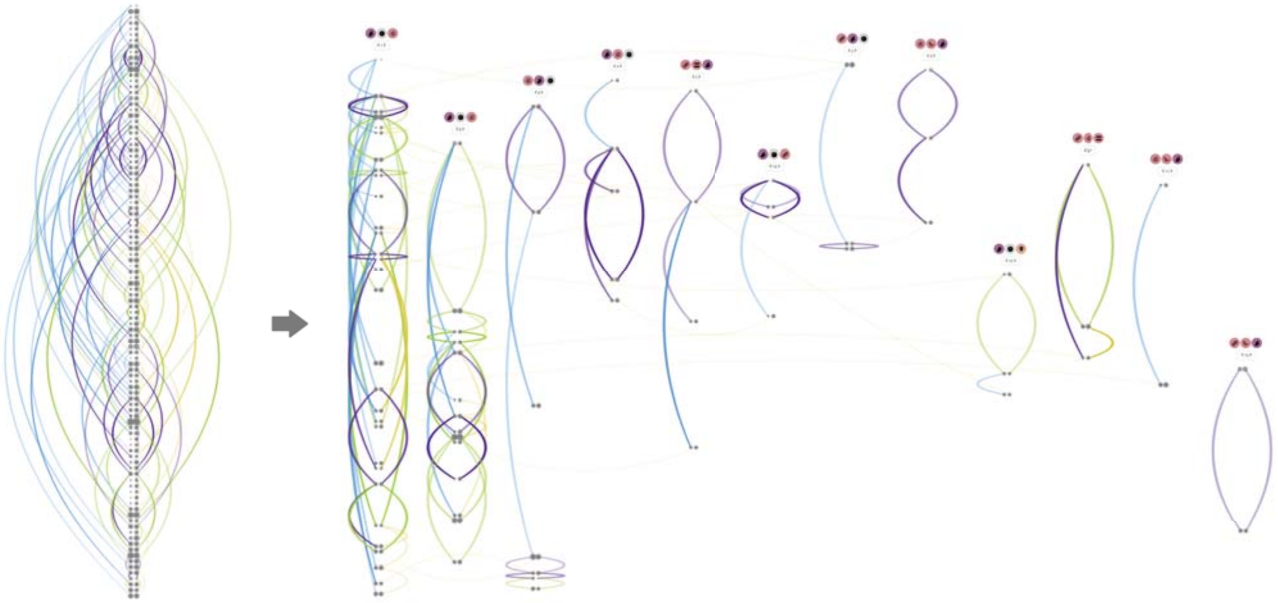
Erschienen in: Computer Graphics Forum ; 37 (2018), 3. - S. 351-365
https://dx.doi.org/10.1111/cgf.13425

ThreadReconstructor: Modeling Reply-Chains to Untangle Conversational Text through Visual Analytics

Mennatallah El-Assady[1,2], Rita Sevastjanova[1], Daniel Keim[1], and Christopher Collins[2]

[1]University of Konstanz, Germany
[2]University of Ontario Institute of Technology, Canada



(a) *Reply-Relation View*

(b) *Thematic-Forest View showing each connected component as a separate tree, sorted by the number of posts.*

Figure 1: *Thematic-Forest (1b) of untangled reply-chains from a full-conversation (1a) according to a content-focused query (**left arcs**) compared to a random-forest model trained on 13 features (**right arcs**). Model **agreement** and **match to ground truth** are shown using color.*

Abstract

We present ThreadReconstructor, a visual analytics approach for detecting and analyzing the implicit conversational structure of discussions, e.g., in political debates and forums. Our work is motivated by the need to reveal and understand single threads in massive online conversations and verbatim text transcripts. We combine supervised and unsupervised machine learning models to generate a basic structure that is enriched by user-defined queries and rule-based heuristics. Depending on the data and tasks, users can modify and create various reconstruction models that are presented and compared in the visualization interface. Our tool enables the exploration of the generated threaded structures and the analysis of the untangled reply-chains, comparing different models and their agreement. To understand the inner-workings of the models, we visualize their decision spaces, including all considered candidate relations. In addition to a quantitative evaluation, we report qualitative feedback from an expert user study with four forum moderators and one machine learning expert, showing the effectiveness of our approach.

1. Introduction

Massive online conversations, such as forums and comment sections, or real-world discussions, such as political debates, produce lengthy, *verbatim* transcripts of people's viewpoints on different issues. These texts result from the *social interaction* between a discussion's participants; however, explicit reply-relations (i.e., threading) are not always available, nor (if present) do they always represent all relevant aspects of connection between contributions in a discussion. Due to their *implicit conversational structure*, information contained in these mediums is not readily available for analysis. Therefore, to understand stances, arguments, and opinions in conversations, it is crucial to gather this data to structure and analyze its content.

To counteract this, some forums, e.g., Reddit [reda], provide their users with the option to create a nested reply, producing threaded conversations. However, not many forums maintain such a logical reply-structure as a publicly accessible interface. Most commonly, only the temporally-ordered sequence of posts is provided for usage and further analysis. Hence, to observe some existing patterns in the data, analysts rely on automatic techniques to reconstruct the reply-relation structure or manually go through the whole dataset. Moreover, even when the threaded structure is given, analysts often observe multiple intertwined discussions in supposedly coherent threads. This is due to drifts into side-discussions, participants not making use of the provided "reply" functionality, or posts not strictly replying to previous messages but rather generally commenting on all previous text. In addition, different reply structures might be valid, depending on the semantic context of the analysis.

Forum moderators and political analysts often have to go through such lengthy transcripts on a daily basis. In our interviews with professional forum moderators we learned that they sometimes have to resolve legal or policy issues arising from users' misconduct. Following the flow of a discussion to understand all the relevant exchange of opinions and information is crucial to fulfilling such a task. They sometimes need to ensure that they have captured all related reply-chains – favoring recall over precision. While for other tasks, they are usually interested in precisely reconstructing reply-chains with a high accuracy. Forum moderators, therefore, describe this as being one of the most time-consuming parts of their job. On the other hand, our political scientist colleagues use statistical models to test different theories of communication and argumentation on real-world debates [EAHJG*17]. These models may be improved by automatically separating debates into thematically-coherent threads.

Facing these challenges, we identified five requirements for an effective solution, namely; (1) Accurate Reconstruction (supporting different analytical tasks); (2) Untangling Conversational Threads (into distinct connected-components); (3) Understanding Relations (e.g., branching-out of posts into conversations); (4) Comparing and Understanding Reconstruction-Models; (5) Optimizing Reconstruction-Models (to the given semantics of data and tasks).

Addressing these requirements, we present *ThreadReconstructor*, a visual analytics approach for the semi-automatic reconstruction of discussions into threaded conversations. Our framework is designed to support (1) Thread Reconstruction, (2) Untangling Conversations, and (3) Model Diagnostics. It enables users to answer questions, such as: Which utterances are related in a transcript? What are the main discussion topics and how are they related? Which aspects of the text did different reconstruction models favor? How do different models compare to each other and how well did they perform?

Our research is motivated by the need for a semi-automatic thread reconstruction technique to assist researchers and practitioners, alike, in analyzing their data, extracting conversational structures in lengthy transcripts, and understanding the underlying models to gain trust in using them. Through teaming-up with domain experts, we identified three stakeholder groups, namely; (1) Analysts (e.g., forum moderators, political scientists); (2) Creators (e.g., machine learning experts developing automatic thread-reconstruction models); (3) Participants (in a conversation).

Our proposed technique aims at tightening the human-in-the-loop process through enabling users to participate in the reconstruction of threaded conversation structures, using visual analytics. To focus the scope of this paper, we are reporting the usage of our approach to forum data, more specifically, Reddit data [reda]. This is because it is a rich data-source that provides a ground-truth structure of threaded conversations. In addition, we have worked with professional forum moderators and machine learning experts from a company focused on forum management. Our unique access to the company allowed us to study the usability of our approach in a realistic setting involving a variety of different stakeholders. However, our tool is developed with a wider audience in mind, e.g., political science researchers, scholars of society and culture, journalists, and analysts.

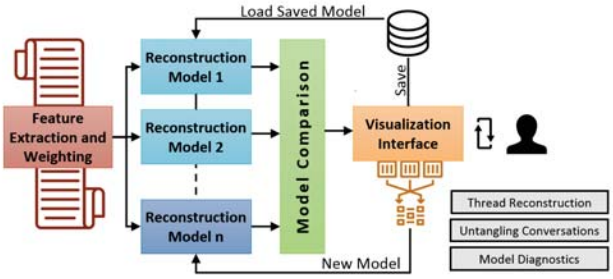


Figure 2: *The architecture of ThreadReconstructor.*

The architecture of our system is described in Figure 2, and it is comprised of several stages. First, features are extracted from incoming conversation data, weighted, and fed to a selection of thread reconstruction algorithms, including machine learning and query-based models. The resulting thread structures are compared, and their overlaps and differences are presented in a visual interface, along with the features and details of the evidence used by the model to generate each thread relation. Based on the results of this process, the user can choose to save some or all relations from the model, or adjust the parameters of the reconstruction algorithms and re-run the process. Each stage will be described in the following sections.

This paper makes the following contributions in the domain of thread reconstruction: (1) An approach for task-driven *feature-selection* and *interactive model-generation*. (2) A visual interface for *model comparison* for thread reconstruction models. (3) A visualization of *connected components* in threaded conversations.

Furthermore, we contribute a visual analytics approach for analyzing *model decision spaces*, which we demonstrate on the task of thread reconstruction but could generalize to other domains. Such an interaction paradigm contributes to ensuring trust in model building through enabling a high-level model understanding without requiring in-depth knowledge about complex machine learning processes. We evaluate the impact of our system with quantitative metrics and through a qualitative study conducted with five domain experts.

2. Related Work

We focus our exploration of the related research to approaches for thread reconstruction (which are generally not visual approaches), and techniques for visualizing conversation data (which do not address the thread reconstruction task). Our work, to our knowledge, is unique in its combination of visualizations for exploring the model space of thread reconstruction algorithms.

2.1. Thread Reconstruction

In general, the reconstruction of thread structure is done based on two approaches: supervised and unsupervised techniques. Both of these approaches require a set of features, which describe the likelihood that two posts in a thread are linked through a reply-relation. Frequently, these features are categorized as content (also called "textual" or "intrinsic") features (e.g., cosine similarity) or meta-data (also called "non-textual", "extrinsic") features (e.g., time-distance) features [AC, SCS09, BFAD13].

Most of the related work uses the cosine similarity function as an indicator of a reply-relation existence. For example, Wang et al. [WJCR07] use a graph-based representation where connections between messages are determined based on inter-message similarity, which is calculated using the cosine similarity function of TF-IDF vectors. They use multiple penalization functions to remove edges between nodes which do not satisfy the selected function (e.g., *time-distance* between messages).

A slightly different approach is presented by Lin et al. [LYC*09], who use a sparse coding-based model named SMSS (Simultaneously Model Semantics and Structure). The model projects each message into a topic space, and approximates it by a linear combination of previous messages in the same discussion. For a new unlabeled post they compute the similarity between itself and each of its previous posts, rank the similarities, and then choose the top ranked post as a candidate parent.

Most of the related work on the thread structure's reconstruction uses supervised methods. Multiple machine learning algorithms have been applied for this task, such as decision trees [SMdR07, AC], support vector machines (SVM) [AC], and ranking SVM [SCS09, BFAD13]. Wang et al. [WLK*11] propose a probabilistic model, *threadCRF* to predict the reply structure for threaded discussions. They use two groups of features: node and edge features. Node features depend on the observed attributes in a post. Edge features are defined over the relation between the attributes of two nodes. Liu et al. [LCC13] use the *threadCRF* to extract a reply-relation structure from patient forums. In patient forums the personal relationships are critical to understand discussion's context. Therefore, features such as matching between the *address*, the *signature*, or the *role* of the person (e.g., my daughter) are important.

Our work relates to these past works in that we provide functions such as cosine similarity as inputs to an unsupervised query-based thread reconstruction model. In addition, we have trained two classes of supervised machine learning algorithms (random forest and decision tree) using features inspired by the literature.

Most of the supervised models are trained and tested on a relatively short dataset. Some authors set a limitation that the thread should be at least three messages long to be used for training the model [SCS09]. None of the mentioned works explicitly test on long threads (e.g., 100 messages per discussion and more). Some authors evaluate threads of different lengths and show that the performance of the classification model is always better for shorter threads. Wang et al. [WLK*11] emphasize that the size of a thread influences the model's performance; if the thread length increases, the performance of the model is reduced. This problem is often reduced by restricting the reply relation distance to a fixed window. However, this has the negative effect of eliminating the possibility to model long distance relations that do exist in real data.

In addition, most of the related work deals with the thread structure's reconstruction task on only one specific discussion type (e.g. emails), using features tuned to that data to create a model. Different models are needed for different conversational text datasets. Our generalized approach presents a flexible query model that can be tuned to different reconstruction goals (topic, author, quotes, etc.).

Often real-world datasets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples [CBHK02]. That is known as the *imbalanced classes problem*. Classification of message reply-relations belonging to one relatively long thread is a representative example of this issue. If a thread discussion contains n messages, then at most $n-1$ reply-relations may exist in data, under the assumption that one message can have at most one single parent. At the same time, $\frac{n(n-1)}{2}$ false reply-relation candidates exist. If n is relatively large, then the two reply-relation classes (*positive* and *negative*) are highly imbalanced. For 100 messages, $\approx 2\%$ reply-relations are of the class *positive* and $\approx 98\%$ of the class *negative*. The performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced, as the accuracy of the model could simply reflect the underlying class distribution [WMZ07]. The model is very likely to predict the majority class regardless of the input features.

Usually, classification algorithms (for two-class or multi-class problem) require the data to be balanced, meaning, that there should be the same (or similar) number of instances representing different classes. This requirement influences the classifier's performance significantly, but reduces applicability of the classifier in real-world scenarios. Our approach includes sampling methods used to balance the training instances for our machine learning models. We allow for arbitrary lengths of threads and do not restrict the reply-relation distance by default, however, we optionally apply heuristics found in the related work to allow for easier comparison of the performance of our tools to past approaches.

Our research demonstrates an approach that works for real-world data, and provides mechanisms for human-in-the-loop analytics when working with low precision and recall possible with classifiers running on imbalanced datasets.

2.2. Visualization of Conversational Data

Conversational data has been visualized using an approach called thread arcs [Ker03], which represents a discussion as a tree, with messages as nodes and reply relations as edges. Hubmann created ThreadVis, a similar arc-diagram-based visualization of email threads [HH08]. The chronology of messages is combined with the branching tree structure of a conversational thread to create a form of temporal arc diagram. Fu et al. [FZCQ17] extend the basic thread arcs technique to present *thread river*, which can illustrate temporal and structural information of lengthy threaded discussions. In their tool, iForum, they offer a set of visualization designs for presenting the main interleaving aspects of MOOC forums at three different scales (posts, users, and threads). ForumReader [DWM04] is a tool combining visualization techniques with automatic topic extraction algorithms to help users explore *flash forums* (representing shallow threads). ForAVis [WRK11] allows to explore online forums using author level, post level and discussion level features, including sentiment analysis. Hoque et al. [HC14] have developed ConVis to support multi-faceted exploration of blog conversations, which contains multiple views to provide thread information at different granularities. Our approach borrows visualization inspiration from these works, which visualize the thread structure embedded in the data. We extend this to construct a threaded relation over non-threaded data. In contrast to works tailored to the content-exploration of threaded forum data [HSS10, Che15], this work is designed to enable the generation and comparison of thread reconstruction models.

Some authors use visualization techniques to show topic changes within a discussion, such as El-Assady et al. [EAGA*16] showing the speaker dynamics across topics and time, or visualizations which use a river metaphor to represent topic flows over time [TG10, LYW*16]. In the work of Trampus et al. [TG10] a timeline displays the evolution of forum topics, and a semantic "atlas" shows a thematic overview of discussed topics. Liu et al. [LYW*16] provide an overview of the evolving hierarchical topics by connecting the corresponding topics at different times.

To the best of our knowledge a tool which can be used to reconstruct thread structure and provides a visual evidence of the results does not exist. In the following sections, we will describe the back-end data modeling and front-end visualization interface that compose our ThreadReconstructor framework.

3. Problem Characterization

As motivated by Liu et al. [LWLZ17], an effective use of visual analytics in the context of machine learning is to address the analysis tasks of (1) Understanding, (2) Diagnosis, and (3) Refinement of the machine learning models. This work is designed to support such tasks, applied on the problem of *semi-automatic thread reconstruction in conversational text*. More precisely, our approach aims at allowing the users to **[T1] explore their data** and **[T2] get a better understanding of thread reconstruction models** (applied to their data); **[T3] generate** and **[T4] diagnose different models** through comparing their decision spaces; and **[T5] refining the classifiers** using a set of descriptive features for reply-relations. This section describes the research problem addressed by our work, including the targeted users, tasks, and data.

Thread Reconstruction Models As discussed in subsection 2.1, the reconstruction of reply-relations in multi-party conversations has been an active area of research. Machine learning approaches proposed to address this research challenge mostly rely on classifier models. In the context of our work, we define a model as a reconstruction method that takes-in a list of speaker-turns (without relation information) as input and generates an acyclic reply-relation graph as output. Our framework allows users to generate various models through combining pre-trained classifiers with logical queries and heuristics, as described in section 5. All relations are based on a rich set of features, as described in subsection 4.1.

Conversational Text Data Transcripts of multi-party discourse, such as in forums and debates are based on the interaction between speakers and, thus, comprise a structure of implicit reply-relations. Some forum interfaces show such a structure in the form of threaded discussions. However, in other domains, figuring out a possible reply-relation structure is a more challenging task. Moreover, based on our discussions with forum moderators, the analysis of such lengthy conversation transcripts often indicates multiple valid threaded structures, depending on the considered textual and semantic features. Therefore, our framework is designed to support the users in creating and refining thread reconstruction models to tailor them to their respective tasks and data.

Users and Analysis Tasks This approach is designed with three different user groups in mind. In the following, we describe each stakeholder and some of their potential analysis tasks, based on our interview with employees of a forum-management company. For the user study (described in section 9), we recruited participants from all of the three user profiles. The first group of users are **analysts**, such as professional forum moderators. In contrast to community forum moderators (who are concerned with controlling the content of discussions), these analysts are interested in resolving legal issues and claims arising in a forum community, reporting on the impact of topics/brands in different threads, and analyzing the potential substructures that might split discussion communities. To perform such tasks, multiple thread reconstruction models have to be utilized in order to allow for a focused analysis. The second stakeholder group is model **creators**, such as machine learning experts. These users are typically interested in comparing and diagnosing their models on real-world data to understand pitfalls and refine their algorithms. They are also interested in understanding the impact on textual features on the models to analyze their sensitivity and robustness. In contrast to the analysts (who are interested in creating multiple reply-chain structures depending on their data and tasks), machine learning experts often rely on an existing "ground-truth" structure to train and refine their models. Lastly, **participants** in conversations might be interested in exploring the evolution of certain discussion branches, analyzing new arguments and sub-communities in an active forum with a lively debate.

4. Data-Driven Feature Extraction and Weighting

In order to accurately reconstruct the threaded structure of a conversation we rely on a set of tailored linguistic and statistical features extracted from the text. These features are weighted based on their distribution in the dataset being analyzed. As described in this section, we extract features for relations between each pair of posts and categorize the posts into 10 different categories.

4.1. Reply-Relation Features

Reply-relation features are extracted from pairs of messages and capture patterns that can indicate parent-child thread relations. We organize them into three categories based on the **content** of the message, the **structure** of the message in relation to others, and **meta** features related to the dynamics of the conversation.

Content Features — Cosine Similarity measures the lexical overlap using the *cosine similarity* [Sin01] function. To improve the detection of similarity, a series of enrichment features expand messages' word vectors as follows: **WordNet Enrichment**

adds synonyms from WordNet. **URL Enrichment** adds words from linked URLs. **Topic Enrichment** adds words from topics represented in the document, using a topic model. **Author Enrichment** adds words used by the same author in other posts. Additionally, **Word Embedding** uses coreference resolution [LRC*12] to bring messages closer together lexically by replacing all mentions from one coreference-chain with the first mention (referent). After applying these enrichment techniques, the cosine similarity of the enriched message pairs is recalculated. Finally, messages are modeled using four topic model algorithms (LDA, IHTM [EA15], SWB [CSS06], and BTM [YGLC13]. **Topic Agreement** is a Boolean feature which is true if at least *k* (a threshold set by the user) models agree that the messages belong to the same topic.

Structure Features — All structural features are numeric counts of the occurrence of a relation between message pairs. **Quotes** counts the number of direct quotes across messages, as indicated by quotation marks or ">" (greater than) characters, and the body of the quote. The **Author Name Reference** feature similarly counts explicit references between messages, in this case the the number of times the parent message's author is explicitly mentioned in the child message. **Substring** counts the number of common substrings of at least N (default=4) tokens. Substrings may indicate a contextual connection. If an **N-Gram** is frequently mentioned in the discussion, it could be an indicator that these tokens are important for the conversation's primary subject. We extract common n-grams from the dataset we count the number of co-present n-grams across message pairs. The Stanford **Named-Entity** Recognizer [FGM05] is used to extract named entities. Shared named-entities are counted for each message pair. If the same word or n-gram appears densely in the discussion, it is a **Lexical Episode** and may indicate that messages where the sequence appears are part of one subtopic. Lexical episodes are extracted, then the count of occurrences of different episode-words in two messages becomes the value of this feature. Finally, the Stanford CoreNLP **Coreference** resolution [RLR*10] is used to extract and count *coreferences*, which indicate shared content between messages.

Meta Features — The **Distance** feature captures how many messages have been posted in the time between two messages, while the **Time Distance** represents the time-span between replying messages in a thread. The Boolean **Different Authors** feature is true if two messages have different authors. The intuition is that participants of a discussion do not usually reply to their own previously written messages.

4.2. Message Categorization

Different types of messages are more likely to be classified correctly by different models. Through categorizing messages, we are able to filter-out certain types of messages and sort the corpus during analysis, in order to see how discovered relations correlate with message categories. The following message categories were designed using a sample of 10 Reddit threads of 100–200 messages each to explore the characteristics.

Messages having less than 10 tokens are classified as **short**, messages having more than 40 tokens are **long**. **Quote** messages contain quotation marks, usually indicating a citation of a previous message. **Question** messages end with a question mark, which suggests searching for an answer in the following messages. Posts are classified as **active author** if the author has posted 3+ posts in the thread and **inactive author** if the author has only posted once.

Some messages have little useful content, such as messages containing only **URLs**, messages containing mainly **special characters**, and messages classified as **junk** [LYC*09] due to being off topic, containing a high proportion of banned words, or being less than 3 tokens long.

4.3. Data-Driven Feature Weighting

For each analyzed dataset, we pre-compute the distribution of the posts across categories and the presence of different reply relations. These are used as indicators for the users when constructing models. Additionally, these distributions are used for a data-driven weighting of the features to set tailored initial weights for every model. This pre-computation step has proven to be effective [EASS*18] for tuning models to changing data characteristics. All weights can be manually adjusted by the users in the visualization interface.

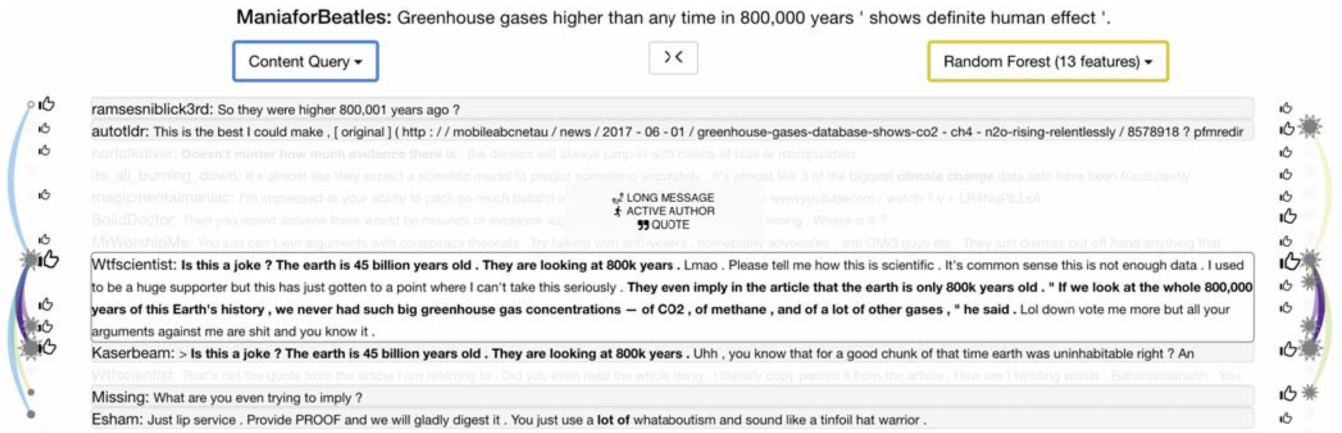
5. User-Driven Model Generation

As shown in Figure 2, the feature extraction step builds the foundation for generating the reconstruction models. Reconstruction models are built by combining three techniques: a classifier model; a user-defined query; and a rule-based heuristic. One or more techniques must be selected, and they are executed in order as later techniques act as filters on earlier ones. All models potentially produce multiple parent candidates per child message, therefore, as post-processing step, these candidates are ranked and the top parent is chosen to be displayed in the visualization interface.

5.1. Classifier Model

Decision Tree [SK16] and Random Forest [Bre01] algorithms were trained on Reddit data to create models for classifying any pair of messages as being in a reply-relation or not. Decision Trees were previously used by Aumayr and Chan [AC] and showed high performance. Random Forests are used as a comparison as they are more robust against overfitting. Two versions of each model were trained using WEKA [FHW16], one using 5 features defined by Aumayr and Chan [AC] and one with 13 features.

The training dataset contains 6926 threads, from which existing reply-relations were extracted and labeled as positive. All remaining pairs of messages were linked using a created relation labeled as negative. The numerical features (e.g. *cosine similarity*, *distance*) were calculated and added to the data set. Boolean features (e.g. *different author*) were represented with 1 or 0. Following Aumayr and Chan [AC], we artificially create a balanced training dataset using under-sampling, to reduce the number of instances representing



ManiaforBeatles: Greenhouse gases higher than any time in 800,000 years ' shows definite human effect '.

ramsesniblick3rd: So they were higher 800,001 years ago ?

autotldr: This is the best I could make , [original] (http : / / mobileabcnetau / news / 2017 - 06 - 01 / greenhouse-gases-database-shows-co2 - ch4 - n2o-rising-relentlessly / 8578918 ? pfmredir

norfolkdiver: **Doesn't matter how much evidence there is** . the deniers will always jump in with claims of bias or manipulation .

its_all_burning_down: It's almost like they expect a scientific model to predict something accurately . It's almost like 3 of the biggest **climate change** data sets have been fraudulently

magicmentalmaniac: I'm impressed at your ability to pack so much bullshit i / wwwyoutubecom / watch ? v = LR4NuPILLsA

SolidDoctor: Then you would assume there would be mounds of evidence sup wrong . Where is it ?

MrWorshipMe: You just can't win arguments with conspiracy theorists . Try talking with anti-vexers , homepathy advocates , anti GMO guys etc . They just dismiss out off hand anything that

Wtfscientist: **Is this a joke ? The earth is 45 billion years old . They are looking at 800k years .** Lmao . Please tell me how this is scientific . It's common sense this is not enough data . I used to be a huge supporter but this has just gotten to a point where I can't take this seriously . **They even imply in the article that the earth is only 800k years old . " If we look at the whole 800,000 years of this Earth's history , we never had such big greenhouse gas concentrations — of CO2 , of methane , and of a lot of other gases , " he said .** Lol down vote me more but all your arguments against me are shit and you know it .

Kaserbeam: > **Is this a joke ? The earth is 45 billion years old . They are looking at 800k years .** Uhh , you know that for a good chunk of that time earth was uninhabitable right ? An

Wtfscientist: That's not the quote from the article I am referring to . Did you even read the whole thing . I literally copy pasted it from the article . How am I twisting words . Bahahahaahahh . You

Missing: What are you even trying to imply ?

Esham: Just lip service . Provide PROOF and we will gladly digest it . You just use a **lot of** whataboutism and sound like a tinfoil hat warrior .

Figure 3: *Text-Level view showing all connections of a selected post, according to a content-focused query and a random forest model.*

the majority (negative) class. The final dataset has 110,038 positive instances, and 110,038 negative ones (in total 220,076 instances).

The user can apply one of the trained classifiers to extract the reply-relation structure from new datasets. Current models trained on the Reddit data have high recall, but poor precision (high false positives). As such, these models are suitable as a preprocessing step to determine parent candidates for the query-based model.

5.2. User-Defined Queries

The user-defined query is a set of criteria which are used to classify message relations as positive or negative. The query is generated in the following way: the user selects a feature subset to use for the reply-relation reconstruction and creates a logical expression. For each feature, s/he can set some parameters, such as the minimum similarity level (for content features), maximum distance between messages (for structural and meta features). S/he can weight each feature, specifying its importance in the decision making process. When the query is executed, for each possible reply relation the system checks if this relation satisfies the given query. All matching relations are seen as positive relation candidates. The final relation score is the weighted sum of comparison of all features. For each message, a list of parent candidates is stored, sorted to their scores. The first message (having the highest score) is seen as the most suitable parent message.

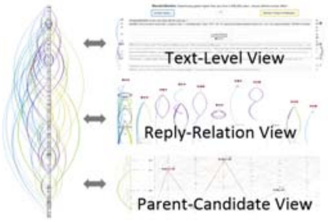
The query-based model has multiple advantages against the trained classifier. First of all, it does not overfit. The user can create different queries, and the system provides an evidence how reliable the extracted relations are, by showing the presence of features and the number of possible parent candidates for each of the child message. The user can use this knowledge to adapt the query and improve the certainty level of the extracted reply-relations. The query-based model is unsupervised. Hence, it can be applied on discussions where no ground truth information exists.

5.3. Rule-Based Heuristics

Sometimes it is likely that the most frequent parent message is the first message in the thread (the message contributing the discussion's title). For example, Balali et al. write that, "Commenters who post only one comment on a thread are more likely to reply to the root post." [BFAD13]. Thus our system will assign the title message as the parent to all messages where no appropriate parent message has been found. This heuristic is extended with the finding that an author usually does not reply to the title post in their subsequent message [BFAD13], so our rule doesn't apply to messages by the author of the title message.

6. Visual Analytics Interface of ThreadReconstructor

Our interactive visualization interface consists of four different views, addressing different requirements and analysis tasks. We use visual-anchoring and staged animations to transition from one view to the next. We use a consistent color scheme for the **left model**, the **right model**, the **model agreement**, and the **true relations**. All transitions from one view to the next happen through the *Reply-Relation View* (Figure 1a), which shows an overview of the complete conversation and all reply relations. In addition to this central view, we introduce three visualizations, each tailored to address one analysis task. The *Text-Level View* is used for close reading and detailed inspection. The *Thematic-Forest View* untangles a conversation into separate connected components. Lastly, the *Parent-Candidate View* is used for model diagnostics. While in any view, users can change settings and generate models using a side-bar, and all saved models are available for selection through drop down menus on top **left** and **right**, as shown in Figure 3. In addition, all views support a rich set of interactions, e.g., selection, sorting, filtering, linking, and brushing.



Visual Design Considerations Inspired by the technique of Thread Arcs [Ker03], we represent each utterance in a conversation as a node, while the reply structure is depicted as an arc. Hence, a reply-relation consists of two nodes and a link connecting them. To reduce edge crossings, the diameter of each arc is proportional to the distance between its endpoints.

Nodes are by default ordered by their timestamp. Each node is shown using a circle ●, with a radius proportional to its message's word count. As introduced by Vehlow et al. [VRW13], to indicate the certainty of a particular parent-child relation, each node encodes information about the number of parent candidates associated with it using a branching-out pattern, i.e., a node ✳ with fewer parent candidates is considered more certain than a node ✸ with more potential parents to choose from. The number of parent candidates is determined by the thread reconstruction model and is explicitly shown in the Parent-Candidate View, as described in section 7. *Junk* messages are colored black so that users can see their prevalence and remove them from view. Nodes with a reply-relation connecting directly to the title message are indicated with a white circle overlaying the node ✺, and edges from these nodes are omitted in the overview to reduce edge clutter. Lastly, additional information about each node and link are shown on-demand using tooltips.

Model Generation Interface Accessible through any view on-demand, the Model Generation Interface is located in the right side-bar of the system. This component allows users to generate custom-models, as described in section 5. The three model generation options are enumerated and individually selectable. First, the users have the choice to select one of the trained classifiers. Next, they can choose a query. This can be one of four default queries or created using a visual-querying interface, as described by El-Assady et al. [EASG*17]. This visual-querying interface enables the users to create nested logical expressions through dragging and dropping the feature-icons and logical operators. Next, the users can select to apply the rule-based heuristics in their model. Lastly, they have the choice to save a model by giving it a name. All created models immediately become accessible as selection for the **left** and/or **right** side of the visualization, as shown in Figure 3. In addition to the saved models, users can select to view the given ground-truth structure (if applicable), a model with all saved relations, as described in subsection 7.2, as well as provided default models. The remainder of this section will explain each view in more detail.

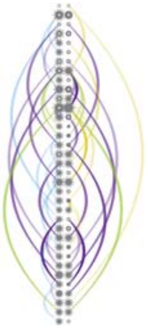
6.1. Text-Level View

This view, shown in Figure 3, is the first component users see. It is designed to facilitate the close-reading of texts. By default, all messages are ordered according to their posting-time, and the first line of each message is shown as one box in the center of the view. Hovering over a message reveals the remainder of the hidden text meta-information in tooltips and edge-highlights. Selecting a message pins it and all its relations in the view for closer inspection. Edges are the connections created by the chosen models on the left and right side, respectively.

This view also indicates influential posts. These are found using a score predicting up-voted content that correlates with the number of children. This score is model-dependent and is encoded in the size of a glyph 👍. Messages categorized as junk messages are mark with a black border to indicate that they have insufficient information for an accurate classification. After getting an impression of the dataset, users can switch to other views with the button in the upper center,which collapses the text through closing the two models on the side into the middle, creating the Reply-Relation View.

6.2. Reply-Relation View

As shown in Figure 1a, this visualization is the central overview component of the system. It displays the connections of the left and right models on zoomable canvas to allow for scalability. As the main overview, the Reply-Relation View the connecting element that enables transitioning from and to all three other visualization components. The side-figure shows an example of the connections displayed in this view. The thread diameter reflects the length of a connection and its color indicates the type of the relationship. On the left side of the tree, the tool displays a detail-bar (not depicted in the figures) with additional, on-demand information about nodes and edges in focus. In addition, when hovering over a relation, an informative tooltip is shown, depicting all feature values of the relation and common words and phrases shared by the parent and child of this relation. Based on this view, the users can select to transition to the untangled Thematic-Forest View, the Parent-Candidate View, or go back to read the text.



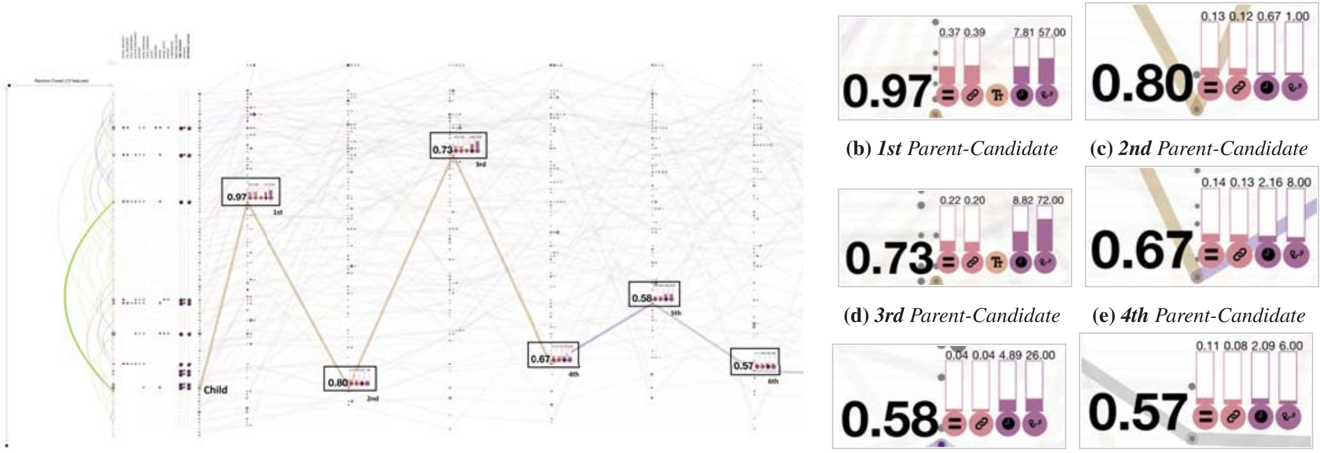
6.3. Thematic-Forest View

Figure 1b shows an example of a Thematic-Forest. This visualization shows the untangled reply-chains of a conversation. As the left and right models might not agree on the separation of a discussion into connected-components, layout of this view is based on one model. Using a staggered animation, this view pulls apart the different connected components from the Reply-Relation View, resulting in a forest of trees. To reduce clutter, posts that reply only to the title and have no children are joined in one separate connected component. Using the *sort* operation, the users can rearrange the trees to order them according to their number of posts, highlighting the largest trees. In order to get a more compact view, users can chose to *move* all trees vertically closer together, disregarding the temporal positioning, as shown in Figure 6. In addition, users can *rebase* the untangled structure from the left to the right model or vice versa, to compare their differences.

Each tree in the forest is annotated with the top three features its relations share. Common words and phrases as well as the particular features common in that tree, for example, if all reply-relations of a tree share a specific quote, are shown on demand. Features that are used in the current model are shown as colored icons, while the ones that are common in a tree but not used in the model are shown as a gray icon (e.g., Figure 6a). This is an indicator that certain features might be interesting to consider adding during model refinement steps. Each tree in the forest can be analyzed in the same manner as the overview. By clicking an expansion-button on top of the tree, a Text-Level view is opened in a separate window, as a new analysis sandbox showing only the nodes of the selected component.

7. Visual Model Diagnostics

Reviewing automatic and semi-automatic machine learning approaches has become an essential task to ensure model reliability and trust [KDS*17]. However, to perform an educated critique of different models, users have to narrow down the number of considered



(a) *Parent-Candidate View showing a path of all considered parents of a selected child-node.* **(b)** ***1st*** *Parent-Candidate* **(c)** ***2nd*** *Parent-Candidate* **(d)** ***3rd*** *Parent-Candidate* **(e)** ***4th*** *Parent-Candidate* **(f)** ***5th*** *Parent-Candidate* **(g)** ***6th*** *Parent-Candidate*

Figure 4: *Close-up views on the Parent-Candidate View shown in Figure 5b, using one selected child-node as an example.*

models or embark on an extensive educational journey to understand all relevant aspects of the internal workings of many models. In this section, we present a visual analytics technique that simplifies the task of model tuning through taking this action from the Model Space into the *Model Decision Space*. This approach maps all relevant information from each model into a comparable space, revealing the considerations that models took to reach a decision.

7.1. Parent-Candidate View

As an instance of a *Model Decision Space*, the Parent-Candidate View visualizes the inner-workings of thread reconstruction models. In particular, it shows all considered parent candidates for each child, ordered by the internal model certainty, as shown in Figure 4. For every node in the conversation, all parent candidate are ordered in columns according to their model-based ranking. In a representation reminiscent of a parallel-coordinate plot, each column can be treated as one dimension, with the first containing the child node followed by one dimension for each candidate ordered by their likelihood. The vertical position is determined based on the corpus order, e.g., sorted according to the time-stamp of the posts. Hence, the first two columns of the Parent-Candidate View contain the information about the child and its connection to the chosen parent candidate.

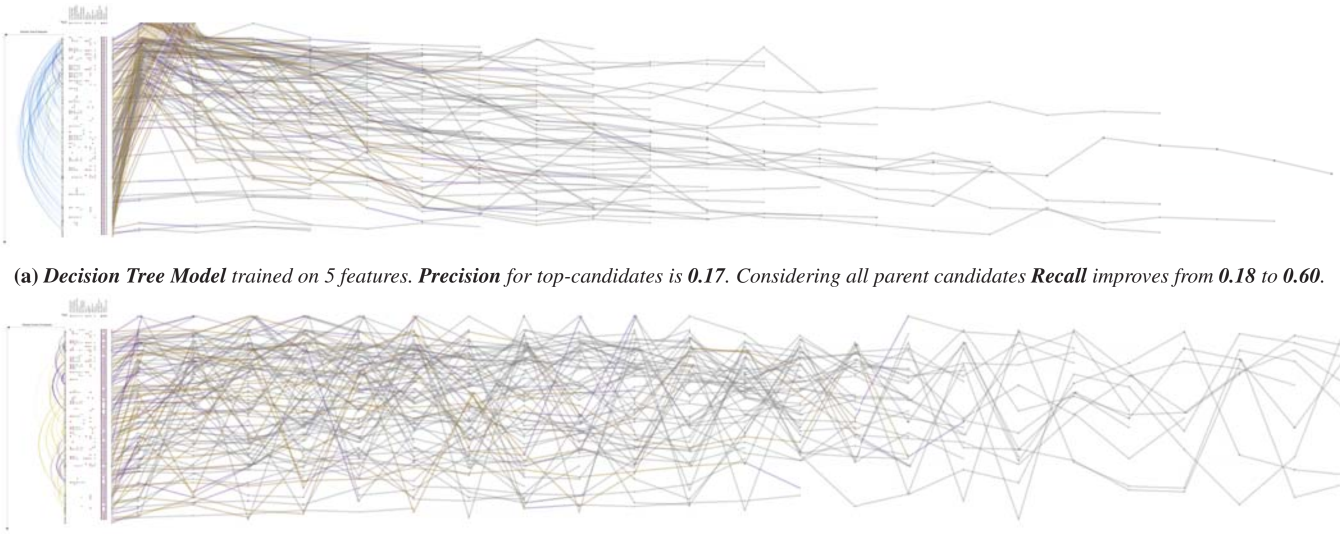
As shown in Figure 4a, this view is divided into three parts. On the left, the reply-relation from the selected model is shown for visual linking to previous views. In the middle, a feature-distribution pane displays the arrangement of features across the conversation. Lastly, on the right, the parent-candidate space is shown. In this visualization, each node has a stable vertical position. If multiple child messages have the same ranking for one parent, a node for each child will be packed in the vertical position of that parent within the same column. By default, all parent candidates for a selected child node are connected with a gray edge. If this **path is leading to a correct parent candidate** it is shown in color. The **last path segment before a true parent** is colored in purple, as is the **true parent node**. While a node or path is selected, the corresponding features for each parent, as well as their model certainty are shown. For example, Figure 4b shows that the random forest model ranked the first parent for the selected node with a certainty of 0.97. Independent of the features considered by the model, a list of relation features and their scores is available for each parent-child relation. These features highlight the sensitivity of the model to particular attributes. For this particular child, the correct parent node according to the ground-truth is the fifth-in-line (Figure 4f). Hence, the model would pick this parent only after considering four other nodes.

In addition to exploring single utterances, the parent space view is designed to give an overview of general patterns on a model-level. For example, Figure 5 shows two machine learning models in comparison. Displaying the decision spaces of the models over a complete discussion allows us to compare their biases and understand their particular sensitivities. General patterns, like edge-bundling, zigzag connections, and others, show us differences in the internal scoring of models and helps users assess their quality. It also shows the influence of certain posts on the discussion. Additionally, the length of the produced paths shows the breadth of the models' search space. In addition to using semi-transparent edges, as well as linking and brushing, a *slice&dice* technique is used to reduce the amount of displayed data in order to reduce visual clutter. Moving the sliders (on the left and top), reduces the considered data to a specific subset of messages to be explored.

7.2. Model Optimization and Transfer

In combination with the other views, the Parent-Candidate View enables users to understand how models choose parents for the reply-relations. Learning about the effect of different features on the reply-relations, as well as the sensitivity of models towards particular aspects of the data, generates an opportunity for model tuning and optimization. This is either done through adjusting parameter threshold for existing models or by refining a model through editing its features. Additionally, understanding the feature distribution in a dataset facilitates the creation of new models based on aspects that became relevant through an exploratory analysis.

Our system supports these tasks through an option to save, load, and edit models for existing and new datasets. In addition, users can combine two models based on their agreement to form a new



(a) ***Decision Tree Model*** *trained on 5 features.* ***Precision*** *for top-candidates is* ***0.17****. Considering all parent candidates* ***Recall*** *improves from* ***0.18*** *to* ***0.60****.*

(b) ***Random Forest Model*** *trained on 13 features.* ***Precision*** *for top-candidates is* ***0.25****. Considering all parent candidates* ***Recall*** *improves from* ***0.26*** *to* ***0.65****.*

Figure 5: *Comparison of two classifier models in Parent-Candidate View.* ***Model 5a*** *relates most of the children to parents from the beginning of the conversation, while showing long descending line-pattern. In contrast,* ***Model 5b*** *shows zigzag patterns with more spread out relations, considering more parent candidates.* ***Paths leading to a correct parent****, as well as the* ***agreement between the two models*** *are color-coded.*

refined one. To support a deeper analysis, we allow users to save relationships during the exploration process. These can be picked from any view, including all parent-candidates. The collection of saved relationships is accessible though selecting 'Saved Relationship' from the model drop-down menu. This provides an optimization strategy that is based on the data characteristics, i.e., creating many specialized models to capture different types of relations and combining their results. These can be applied to the whole conversation or on only selected types of messages, e.g., posts containing a question mark and longer posts (assuming we are searching for question-and-answer patterns).

Lastly, our system supports the creation and training of different models on a dataset with a ground-truth structure to be transferred to datasets with no known threaded structure. This is especially useful for our political science collaborator, as most debates they analyze do not typically have a known threaded structure. Using a manually annotated debate would allow them to train several models and transfer them to a broader dataset of discussions.

8. Scalability and Model Reliability

To quantify the performance of our approach in comparison to related techniques, we studied the scalability of our tool to long conversations and overall reliability of generated reconstruction models. First, we examined the precision, recall, and f-score values of the created classifier models on a balanced training dataset with 6926 Reddit discussions. These ranged in their length from 50 to 500 posts per file. Overall, we observed, as a result of a 10-fold cross-validation, that both decision tree and random forest models achieved higher precision and recall values when trained on 13 features as opposed to five. The highest precision value (0.79) in this artificial setting was achieved by a decision tree model, in contrast to 0.74 for the random forest. However, the recall of the random forest model was slightly higher (0.70) than of the decision tree (0.68). Both models resulted in a similar f-score (0.72 and 0.73). These results are comparable with the current state-of-the-art in machine learning.

However, while testing these models with imbalanced real-world data (40 Reddit discussions with length from 100 to 200), we observed a noticeable decline in the models' performance, as expected on imbalanced data. Both decision trees and random forests had comparably poor precision (0.16) and recall values (0.16). When relaxing the measures to include the top 10 parent candidates instead of the top one, the recall increased, but the precision dropped further. We observed that the quality of the models varied across different datasets, however, none of the models achieved a result on real-world data which was comparable to results with balanced relations. This problem is a known issue in machine learning [DPCJB15].

Examining the results of the default query-based models on the same sample of data revealed they often outperform the trained classifiers on certain aspects. For example, the precision-based query achieves a precision value of 0.87, while recall-based queries achieve recall values of up to 0.28. As with the classifiers, these results fluctuate across different datasets. However, queries can be designed as tailored models, optimized to reconstruct a specific aspect of connection. This finding manifests the importance of a visual analytics approach, as only the combination of different model generation methods allows for accurate thread reconstructions through harnessing the strengths of each method.

In contrast to the related work [SMdR07, AC], our approach showed promising results in reconstructing threads in conversations up to a length of 200 posts. To investigate the scalability of our system with relation to thread length, we conducted a small experiment, analyzing the effect on the length of a conversation on the reliability of the models. We chose to use the same sample of 40 Reddit discussions used in previous investigations, extracting from each file into a new data record containing the only the top 10 or the top 30 messages, respectively. Using the default query models, we investigated the effect of the length of a conversation on the precision and recall values of the modes. The results were apparent and confirmed findings by Wang et al. [WLK*11]. Since shorter conversations contained substantially fewer false-positive relation candidates, both the precision and recall values of models reconstructing threads in shorter discussions increased. For example, a content query applied to our data sample had an average precision of 0.36; 0.56; 0.70 and recall of 0.29; 0.48; 0.66, when applied on all posts; the top 30; the top 10, respectively.

9. Expert User Studies

To evaluate the effectiveness and usability of our system, we conducted a qualitative expert user study. Through a collaboration with a company that owns and operates a significant amount of online community forums, we gained a unique insight into their work and the challenges that they face while dealing with massive online datasources. After a description of the study arrangement, in this section, we discuss the case studies, feedback gathered, and lessons learned.

Methodology Due to the number of available views and settings in addition to the limited time of qualified experts to learn all aspects of a new system, we chose to conduct a pair analytics study [KF14]. We conducted five two-hour sessions in which a member of our team (henceforth referred to as **v**isual **a**nalytics **e**xpert, short **VAE**) worked with the one domain expert (henceforth referred to as **s**ubject **m**atter **e**xpert, short **SME**). Each session was divided into three parts. In the first 30 minutes, after a quick introduction, the VAE explained the functionality of the tool while gathering initial feedback on the overall utility and design choices through a semi-structured interview. After making sure the SME understood the basic functionality and controls of the system, in the next hour the SME and VAE embarked on an open-ended, exploratory analysis of one dataset using the tool, guided by the interests of the SME. The SME had the control over the interface with occasional input from the VAE to clarify certain interaction possibilities. During the analysis, the SMEs were encouraged to think aloud and explain the tasks they were doing. Questions from the VAE occasionally guided this study towards new analysis tasks. The last part consisted of a 30-minute feedback session, reflecting on the initial feedback and the performed analyses. This last part was guided by questions from the VAE to get a general assessment of the system, and its utility and visual design from the SME. All study sessions were audio-recorded and screen-captured for further analysis.

Dataset and Controls To ensure the validity of our study we sought a dataset that fulfills certain criteria used as controls. First, we were looking for discussions with generally familiar contents. Second, the lengths of these conversations ought to be between 100 to 200 utterances, to ensure long enough reply-chains while remaining manageable for the users. Third, in order to have a reference for evaluation, we required conversations with previously known thread structures as ground-truth data. We therefore selected two Reddit debates regarding the topics climate-change [redb] and immigration [redc], respectively. The first dataset is a discussion of a news article entitled: "*Greenhouse gases higher than any time in 800,000 years 'shows definite human effect'*" [newa], which highlights research results showing a rising level of greenhouse gases found through studying air trapped in ice cores. The second dataset is a debate on another news article entitled: "*Trump is deporting fewer immigrants than Obama, including criminals*" [newb]. Using these datasets across the five study sessions (3x climate-change, 2x immigration) allowed us to balance the study with respect to the effects of the content on the outcome. Generally, we observed that the usage of our system did not vary for different datasets.

Participants All the participants in our study work for a company that manages a substantial amount of forum data. However, the variation in their educational backgrounds, job responsibilities, as well as their self-declared technical proficiency varied a lot. Four out of five participants (in following referred to as **f**orum **m**oderators, short **FM1–FM4**) worked in forum moderation (**analysts**) with duties ranging from customer support to legal aid. Their highest educational degree also varied from secondary school across different levels of higher education. All four participants work with computers on a daily basis and are privately interested in forums (**participants**), however, none of them had a background in machine learning. In contrast, our fifth participant (in following referred to as **m**achine **l**earning **e**xpert, short **MLE**) works as a research scientist (**creator**) at the company and holds a Ph.D. in machine learning.

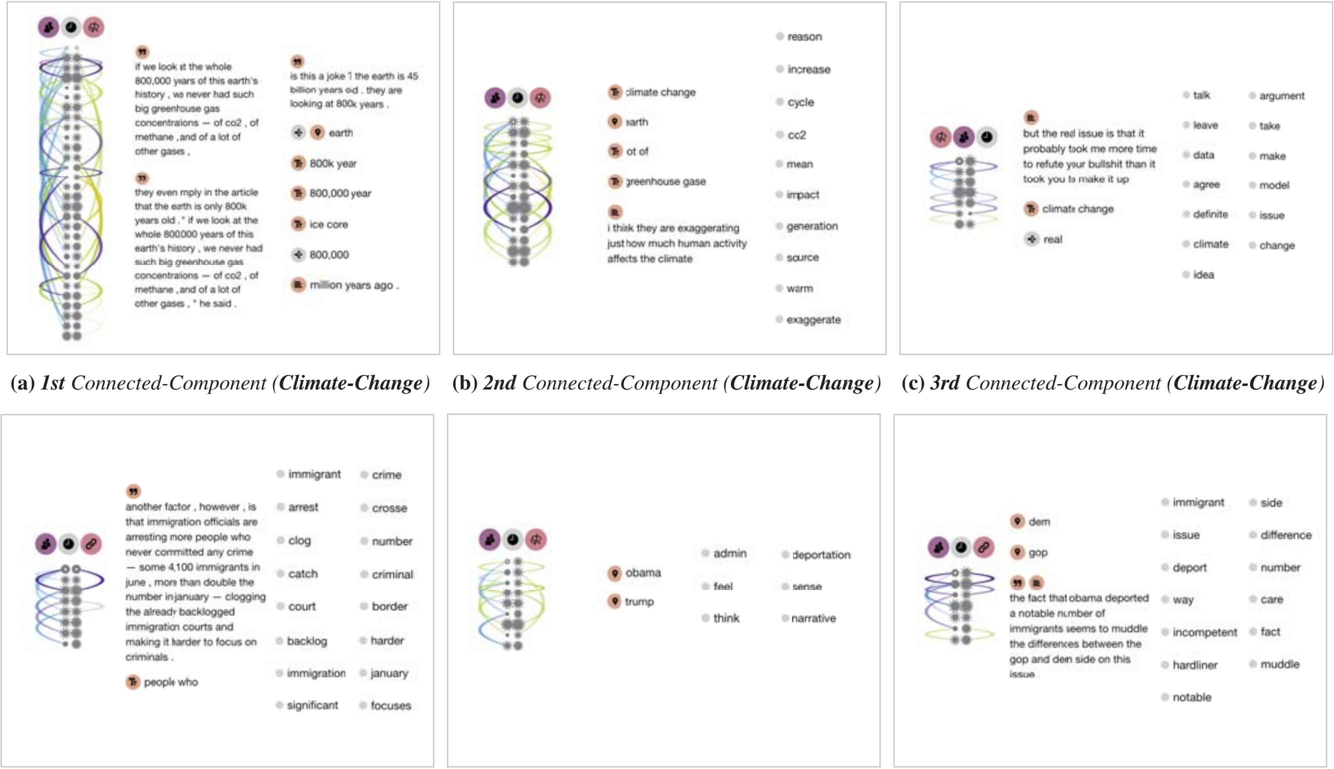
Tasks Based our problem characterization (section 3), the focus of this study was to facilitate four high-level analysis tasks, namely; (1) Data Exploration **[T1, T2]**; (2) Model Generation and Comparison **[T3, T5]**; (3) Untangling Forum Discussions **[T1, T3]**; and (4) Model Decision Understanding **[T2, T4]**. Throughout the exploration process, the SMEs were interested in different aspects of the data and tool. Nevertheless, the VAE guided the analysis towards tasks that were not covered by the SMEs through unobtrusive questions, e.g., "*How about we take a look at the different trees this model generated?*"

9.1. Case Studies

In the following, we describe a selection of case studies chosen from all five sessions. Since many scenarios were repetitive, we picked representative cases and categorized according to high-level tasks.

9.1.1. Data Exploration

Branching out from a topic of interest The first view all users see in the system is the Text-Level View. Typically, users read over the title and some posts to get a feeling of the underlying data. FM3 started his analysis with the same strategy. After finding an intriguing post accusing deniers of climate change never to be convinced regardless of the amount of available evidence, he mentioned that he "*want[s] to follow this post and see what people respond to that.*" He then selected the post to examine its relations according to the default models. Becoming excited about the topic, FM3 announced that " *[he would] create a model to look for people who agreed with each other quickly.*" The generated model was based on a query [🤝 && (📍 || Tr || ⋖) && ↝ && 👥] that favored content agreement and shorter time distances. After comparing this model to existing ones, he deemed it a good one for further exploration, using the Thematic-Forest View. While analyzing the untangled discussion, he commented that "*[...] the discussion about how deniers*


if we look at the whole 800,000 years of this earth's history , we never had such big greenhouse gas concentrations — of co2 , of methane ,and of a lot of other gases ,
they even imply in the article that the earth is only 800k years old . " if we look at the whole 800,000 years of this earth's history , we never had such big greenhouse gas concentrations — of co2 , of methane ,and of a lot of other gases , " he said .
is this a joke ? the earth is 45 billion years old . they are looking at 800k years .
earth
800k year
800,000 year
ice core
800,000
million years ago .
climate change
earth
ot of
greenhouse gase
i think they are exaggerating just how much human activity affects the climate
reason
increase
cycle
co2
mean
impact
generation
source
warm
exaggerate
but the real issue is that it probably took me more time to refute your bullshit than it took you to make it up
climate change
real
talk
leave
data
agree
definite
climate
idea
argument
take
make
model
issue
change
another factor , however , is that immigration officials are arresting more people who never committed any crime — some 4,100 immigrants in june , more than double the number in january — clogging the already backlogged immigration courts and making it harder to focus on criminals .
people who
immigrant
arrest
clog
catch
court
backlog
immigration
significant
crime
crosse
number
criminal
border
harder
january
focuses
obama
trump
admin
feel
think
deportation
sense
narrative
dem
gop
the fact that obama deported a notable number of immigrants seems to muddle the differences between the gop and dem side on this issue
immigrant
issue
deport
way
incompetent
hardliner
notable
side
difference
number
care
fact
muddle

(a) ***1st*** *Connected-Component (****Climate-Change****)* **(b)** ***2nd*** *Connected-Component (****Climate-Change****)* **(c)** ***3rd*** *Connected-Component (****Climate-Change****)*

(d) ***1st*** *Connected-Component (****Immigration****)* **(e)** ***2nd*** *Connected-Component (****Immigration****)* **(f)** ***3rd*** *Connected-Component (****Immigration****)*

Figure 6: *Thematic-Forest View showing the three first connected-components from the untangled view of Figure 1b and their top features (in the top row). They are based on a discussion about the reasons and impact of climate-change and indicate the altered arguments of the debate, i.e. (a) the validity of the scientific work, (b) the impact of greenhouse gases, (c) the frustration with the deniers of climate-change. The second row shows the first three connected-components of an untangled discussion about immigration in the US and the political strategy of Donald Trump in comparison to Barack Obama, especially concerning boarder safety and the number of arrests of undocumented immigrants.*

are wrong is interesting, you can see how much people are frustrated by deniers of climate change in the conversation." Examining more content-related reply-chains, he mentioned that "*[he] learned something about the 'earth cooling cycle'*" from the debate, while commenting as an enthusiast for forum discussions: "*This tool is interesting for me, [...] I would like to participate in this conversation because I can see the history of how the discussion evolved.*"

9.1.2. Model Generation and Comparison

Refining Default Queries All SMEs relied on the default queries for their first experiments with model generation, comparing their results to the preloaded classifier models. Through this initial analysis, they got a feel for the sensitivities of the models and their usability. After forming more concrete questions and hypothesis during this analysis phase, they would continue with generating their own models. FM4, for example, got interested in analyzing relations of named-entities in the discussion and how different entities propagate throughout reply-chains. Starting with a default content-focused query, FM4 continued exploring the relations in the Thematic-Forest. Through investigating the most prominent features in each connected component (see Figure 6), she concluded that she would add the time-distance as a feature of her next query, as this feature was present in the data but missed by the current query.

Verifying Hypotheses FM1 started her analysis with a hypothesis on the behavior of forum users, saying: "*People quoting each other are either agreeing or disagreeing.*" Consequently, she embarked on a mission to examine the aspects of agreement and disagreement around the topic of climate change. She first created a broad, query-based model [(📍 || ❞ || 🔗) && 👥] that is focused on textual or entity repetitions from different authors, commenting that she "*[wants] to create a reliable model based on real information.*" Afterwards, she created a second, more restricted, query-based model [(✣ || 👄) && ❞] in which she was searching for "*posts that are quoted [...] using lexical episodes to filter out aspects like science or politics.*" Both created models had a high amount of true relations. However, FM1 was interested in their agreement, which only contained true relations. After untangling the discussion, she closely read the connected posts (e.g., Figure 3 and Figure 6a) and concluded that "*[she] thought more people would argue against each other but through looking at the reply-chain I found that in this particular conversation people are supporting the claims more than I would have expected on such a controversial topic.*"

9.1.3. Untangling Forum Discussions

Searching for Content Patterns After reading the title of the discussion, FM2 mentioned that "*[he wants] to explore what the dif-*

ferent side-discussions are about." Expecting noticeable patterns in the content of the debate around climate-change, he started exploring the connected components separated by a content-focused query, as shown in Figure 1b. Taking a closer look at the common features of each subtree, he concluded that "*everyone is trying to prove their own theory, they are not really referring to the other theories mentioned. It's a distribution of 'random facts'.*" Seeing a post from a denier of climate-change (Figure 3), he analyzed its connections (Figure 6a) and mentioned that "*this is a person who sparked a heated debate at the beginning of the thread. Many people are outraged in their replies that this person did not understand that the article was referring to 800k years as a reference to the age of the ice cores not the age of the earth.*" During this close-reading process, he occasionally pointed to different utterances saying: "*Here you can see a thread of very argumentative posts.*" In addition, he also found many connected components sharing common URLs, commenting that "*many people refer to the same URL to prove a theory. Here is a link from Wikipedia.*" After realizing this, he declared: "*The model we created worked out very well, maybe I will use the URL feature for the next model.*"

Understanding Connections FM2 was also interested in understanding why some models relate certain posts to each other and others do not. He referred to it as "*helping [him] build more trust in the models.*" To explore this, he chose two models and used the transition between their respective Thematic-Forest Views to investigate changes in the composition of the forum according to each model. He routinely checked selected reply-chains and commented: "*I can see why these posts would connect, that makes sense. The considered features are working.*" or "*This model seems to be biased towards quotes.*" During this analysis, he made another interesting finding. He found messages posted by the *Reddit Helper-Bot* and concluded that "*[...] in some cases posts were automatically removed because participants used a lot of profanity or bad language.*"

9.1.4. Model Decision Understanding

Following the KDD Pipeline While the analysis of all forums moderators was more driven by the content of the discussions, the MLE was more interested in the model optimization task and found the Parent-Candidate View the most interesting. She described this view as showing the "soft links between datasets." Additionally, she commented that "*it visually highlights the sensitivities of the models towards certain aspects of the data.*" For example, Figure 4b indicates that the model's decision was mostly impacted by a high value for the cosine similarity of the child and first parent-candidate. Based on her prior knowledge of the used algorithms, she assumed that "*the decision tree classifier will be less robust because it tends to overfit to the training data.*" Exploring this aspect of model fitness, she continued her analysis through the lens of the KDD pipeline [FPSS96]. She first started exploring the feature distribution across the given dataset, using the feature panel of the Parent-Candidate View, as shown in Figure 4a. She was interested in the overall distribution of features, investigating the features used in a particular model (highlighted in color). She commented that "*the first step in studying the performance of a model is to understand the underlying data distribution, as this is the signal the model is receiving.*"

She continued her analysis with analyzing high-level patterns across different models, as shown in Figure 5. She observed a different pattern for decision trees and random forest and commented that "*the behavior of these two models is as [she expected]. Both models were trained on the same features. However, the decision tree is learning a very narrow pattern, which is a sign of overfitting. In contrast, the random forest model is showing a wider range of considered candidates, indicating a broader search space.*" The last step of the KDD pipeline is the evaluation and interpretation, where the MLE gave an opinion about the two models, she explained that "*[she] suspect[s] that overall the random forest model is a more reliable model for thread reconstruction as it is more robust towards overfitting.*" This confirms our independent observation in all other studies. When asked about model overfitting, she commented that "*optimizing models to achieve higher f-scores is often not practical, as the ground-truth is only one possible connection option, other alternatives might make sense for different tasks. A good connection does not always have to be about content.*" She continued, "*people often fixate on a specific aspect and treat it as the truth. However, models might show how different concepts propagate through a corpus, which helps in understanding the data.*"

Exploring the *Magic-Box* Both FM1 and FM4 were interested in understanding more about the classifier models. Without prior knowledge of machine learning, FM1 mentioned that "*[he was] intrigued to understand more about machine learning, seeing the parent candidates definitely allows me to gain more trust in some models.*" He reached this conclusion after having spent some time talking about his mistrust of automatic, black-box models as a reason for not using them. He also said: "*If I had a more practical demonstration of what machine learning can do, I would have more faith in it – this application is getting pretty high marks right now.*"

While FM1 was discussing model-trust, FM4 was interested in choosing the correct model and optimizing her existing ones. She used the Parent-Candidate View to analyze all candidates picked by a model for a particular post, as shown in Figure 4. She commented that "*reading the text of the parent candidates and seeing their feature weighting [gave her] more ideas on how to create new models.*" Despite not understanding how the classifier models work, FM4 was still interested in using and comparing them. She built herself an abstracted mental-model to deal with the classifiers, calling them "*Magic-Boxes*". Hence, her exploration of these models was purely visual. She commented that "*the random forests look much better in the Parent-Candidate View than the decision-trees.*", referring to their high-level patterns, as shown in Figure 5. After examining the patterns more closely, she concluded for the random-forest model (Figure 4a) that "*the zigzag-path shows two conflicting aspects that the model is considering—it is trying to connect the post to others* (from the top) *that are related to the same content* (entity = climate change) *but also to create connections to posts* (from the bottom) *that are using the same angry language to prove a counter-argument.*"

9.2. Feedback

Throughout the study, we collected feedback from all SMEs in the form of semi-structured interview questions. All tasks performed during the analysis sessions were motivated by the requirements mentioned in section 1. The SME expertise covered all stakeholder groups anticipated for the tools; all FMs were fulfilling the roles of analysts and discussion participants alike, while the MLE tool the role of a creator. In this section, we are reporting a summary of the feedback received during all user study sessions.

Initial Feedback During the first 30 minutes of every study session, we gathered initial feedback from the SMEs to capture their first impression about our system and the general research direction. All experts highlighted the importance of having such a system for their work and endorsed its relevance for the company. FM2 explained that "*this tool will make it a lot easier to pin down who is saying what to whom, which will be really handy for forum moderation, especially when dealing with legal issues.*" This was echoed by FM1, who said: "*I would use it for all the harassment claims that I get because these are often forums that are many pages long and take a lot of time to deal with.*" The same sentiment was shared by FM3, who complained that "*sorting through comments is always difficult, you regularly feel like there is more in there that you must have overlooked.*" He emphasized that "*having a tool like this would help [him] to find the topics [he is] actually looking for and see all their connection even in a larger forum.*"

Hence, all FMs needed to reconstruct reply-relation in forums (as such a structure is not maintained by their company-owned forums). They all reported a very time-consuming, manual process for doing this job and did not support fully-automating this task. FM2 was particularly skeptical of automatic reconstruction algorithms, saying: "*I don't have enough confidence that these approaches completely cover all aspects of the discussions I am looking for. Especially when dealing with legal issues, I have to have all points covered. Using such a system seems to be a good compromise between efficiency and thoroughness.*" He later on clarified that "*the visualization reveals all relevant aspects of the data and confirms [his] mental model, which makes [him] have more trust and confidence in using the system.*"

The MLE commented on the relevance of a gold-standard training dataset, saying: "*Looking at the agreement between the models highlights their certainty, even if the ground-truth is not agreeing with all the connections, they might still be relevant and definitely worth exploring.*" She mentioned that "*[she could] clearly see the benefit of such a tool to find patterns in many different ways, not just the expected ones.*"

Visualization Design and Usability After the analysis, we asked all users to comment on their experience. Except for minor issues (e.g., changing button labels), all participants praised the design and usability of the tool. With comments like; "*That's very beautiful!*" or "*I really like how pretty the visualizations are!*", we were pleased to see that the general aesthetics, color choices, and icons were well perceived. In particular, FM4 commented that "*[she enjoyed] using the visual query interface because it makes it seem very easy to create a model.*" FM1 appreciated having general relation features to chose for the query rather than the tag-based system she usually uses, describing that "*using tags to filter and create query assumes that the users know something about the content of the data, which is often not the case.*" All users also approved of the concept of a layered analysis through visual anchoring and appreciated the transitions between different views highlighting that it preserves the context of the analysis, while making differences visible.

General Assessment Despite the usefulness of the tool, most SMEs noticed a steep learning curve during the introduction of the system. However, all experts had no difficulty using all functionalities explained. When asked about their willingness to learn using such a system, all of them were affirmative, highlighting the benefits they gained from such an expressive tool.

Every user found a set of specific tasks they would like to perform using the tool. For example, FM3 commented that "*busier forums often get confusing to handle and when dealing with harassment issues, [he] does not want to miss a comment in a conversation*", depicting the system as "*a way for wiretapping into conversations.*" He also mentioned that "*this tool helps in revealing the influence of certain posts on a conversation. [He would] use it, therefore, to stop fights happening on forums because of misunderstandings.*" Lastly, he highlighted, that "*often people banned from a forum come back using a different account to continue talking about the same topic. [He saw] potential for the tool to reveal such cases of fraud.*"

When asked what they would do when presented with more analysis time, all users were interested in creating new models to discover other aspects of connection in the data. Surprisingly, all SMEs agreed that while optimizing a model towards a given ground-truth structure could be effective for machine learning, from their practical experience such a structure makes little sense in following a conversation flow. As FM1 explained, this is "*because people often hit the wrong reply button or do not bother at all.*" Therefore, they confirmed that automatizing the reconstruction based on precision and recall values did not fulfill their needs from a thread reconstruction system. FM2 commented that "*this given structure often reveals the order of replies but does not show general content patterns. That is why having user-defined queries is so helpful.*" The MLE highlighted that "*a post can be influenced by many previous messages, so it technically could have multiple parents,*" finding the Parent-Candidate view particularly useful for exploring this aspect.

9.3. Discussion and Lessons Learned

Addressing the requirements identified in the introduction, our tool supports a reconstruction of different reply-chains based on the user's tasks, allowing them to untangle conversations using their own mental models and semantics. The study shows that all users created different models, compared them, and tuned them based on their insights. We also observed that through this analytical process, our system could increase the model trustworthiness based on a better understanding of the generated structures.

From our analysis we see the success of supporting users across a variety of levels of expertise and depth of analysis—from getting users interested in using a tool through appropriate default settings and a starting view that is familiar, to a selection of more advanced features further in the analysis. However, even with the provisions we made for model diagnosis and understanding, model trust remains an important concern for some SMEs, especially when dealing with critical data such as legal and harassment issues.

We observed a process of learning through discovery. While we attempted to provide appropriate training in an introduction session, participants found themselves using features even when they were not fully understood. However, through the process of experimentation, the participants learned about feature extraction and machine learning. The exploratory nature of the interface, allowing for low-cost experimentation with queries and models, likely afforded this online learning.

Participants mentioned the aesthetics of the interface as an important aspect of their desire to incorporate it into their regular workflows. When building future tools to integrate visual analytics into existing workspaces, aesthetics should be carefully considered from both the ability to accurately read the visualizations, but also from the user experience point of view.

Some of the specific interests of participants in the study point to the potential for even more methods for thread reconstructions. As we entered this project, we had the traditional machine-learning view that the goal would be to recreate the ground truth. What we learned was the importance of flexibility in creating thread "*reconstructions*" that, while not representing ground-truth data, did connect messages in meaningful ways for user tasks. An example of a potential new reconstruction method arising from our study results would be to create a model which detects rapid back-and-forth discussions between conversation participants, which may be indicative of a fight.

10. Conclusion and Future Work

We presented *ThreadReconstructor*, a visual analytics approach for reconstructing threads, untangling conversations, and understanding model decisions. Our work is fueled by the need for a semi-automatic technique for revealing reply-chains in large conversation transcripts. Our system is based on a set of tailored features that lay the foundation for user-driven model generation. Models iteratively created with our tool are customized to the data and tasks of the users. In the visualization interface, all created models can be compared and optimized using four different views. We demonstrate the effectiveness of a human-in-the-loop analytics process through an expert user study with four forum moderators and one machine learning expert. Overall, the system was well perceived by all users are deemed useful for their day-to-day work.

In addition to our application-domain specific contribution, our work contributes a general visual analytics technique for tuning black-box models. Through revealing the decision space of a model, we introduce a visualization that allows the comparison of the internal workings of models, while simplifying their complexity. This technique enables machine learning experts and novices alike to tune and optimize models to their needs and data. A particularly interesting finding from our study is that the visual analytics system allowed a machine learning expert and a user with no prior knowledge of machine learning to compare two different classification algorithms in order to tune them. While each participant had their own mental model of how machine learning works (one deeply understanding the algorithms and the other referring to them as "Magic-Boxes"), both achieved similarly good results using visual analytics.

In our future work, we would like to achieve a tighter integration between machine learning and visualization. We observed that many experts used our system to perform the thread reconstruction task as an ensemble model to narrow down the search space for correct reply relations. Supporting this task through machine learning is one desirable goal for our research. Furthermore, we would like to explore other possibilities of refining the Parent-Candidate View to highlight aspects of the model decision space. For example, we would like to explicitly encode the internal model certainty in the distance between parent candidates instead of ranking them. Lastly, we are working on refining the classifier models used in our system, for example through choosing more accurate seeds for the initialization of the random forest algorithm. These could, for example, be based on known topics and structures, such as an off-topic thread, a joke thread, or a political discussion. Our work will be made publicly accessible as part of the VisArgue framework: http://visargue.inf.uni.kn/.

Acknowledgements

This research was undertaken, in part, thanks to funding from the DFG-FOR2111 Research Unit at the University of Konstanz, the Canada Research Chairs program, and the Natural Science and Engineering Research Council (NSERC). We would like to thank our collaborators at VerticalScope Inc. for their valuable partnership.

References

[AC] AUMAYR E., CHAN J.: Reconstruction of threaded conversations in online discussion forums. *Artificial Intelligence*, 26–33. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2840/3279. 3, 5, 9

[BFAD13] BALALI A., FAILI H., ASADPOUR M., DEHGHANI M.: A supervised approach for reconstructing thread structure in comments on blogs and online news agencies. *Computacion y Sistemas 17*, 2 (2013), 207–217. 3, 6

[Bre01] BREIMAN L.: Random forests. *Machine Learning 45*, 1 (2001), 5–32. doi:10.1023/A:1010933404324. 5

[CBHK02] CHAWLA N. V., BOWYER K. W., HALL L. O., KEGELMEYER W. P.: SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (2002), 321–357. doi:10.1613/jair.953. 3

[Che15] CHEN Y.: Visual opinion analysis of threaded discussions. In *IEEE Int. Conf. on Data Mining Workshop* (2015), pp. 646–651. doi:10.1109/ICDMW.2015.65. 4

[CSS06] CHEMUDUGUNTA C., SMYTH P., STEYVERS M.: Modeling general and specific aspects of documents with a probabilistic topic model. In *Proc. Int. Conf. on Neural Information Processing Systems* (2006), MIT Press, pp. 241–248. 5

[DPCJB15] DAL POZZOLO A., CAELEN O., JOHNSON R. A., BONTEMPI G.: Calibrating probability with undersampling for unbalanced classification. In *IEEE Symp. on Computational Intelligence* (2015), IEEE, pp. 159–166. 9

[DWM04] DAVE K., WATTENBERG M., MULLER M.: Ibm research report: Flash Forums and ForumReader : Navigating a new kind of large-scale online discussion. *IBM Watson Research Center 23305* (2004), 1–11. doi:10.1145/1031607.1031644. 3

[EA15] EL-ASSADY M.: *Incremental Hierarchical Topic Modeling for Multi-party Conversation Analysis*. Master's thesis, University of Konstanz, Germany, 2015. URL: https://books.google.ch/books?id=Yw8OjwEACAAJ. 5

[EAGA*16] EL-ASSADY M., GOLD V., ACEVEDO C., COLLINS C., KEIM D.: ConToVi: Multi-party conversation exploration using topic-space views. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 431–440. 4

[EAHJG*17] EL-ASSADY M., HAUTLI-JANISZ A., GOLD V., BUTT M., HOLZINGER K., KEIM D.: Interactive visual analysis of transcribed multi-party discourse. In *Proc. of the Association for Computational Linguistics, System Demonstrations* (2017), pp. 49–54. doi:10.18653/v1/P17-4009. 2

[EASG*17] El-Assady M., Sevastjanova R., Gipp B., Keim D., Collins C.: NEREx: Named-entity relationship exploration in multi-party conversations. *Computer Graphics Forum 36*, 3 (2017), 213–225. URL: 10.1111/cgf.13181, doi:10.1111/cgf.13181. 7

[EASS*18] El-Assady M., Sevastjanova R., Sperrle F., Keim D., Collins C.: Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE Trans. on Visualization and Computer Graphics 24*, 1 (Jan 2018), 382–391. doi:10.1109/TVCG.2017.2745080. 5

[FGM05] Finkel J. R., Grenager T., Manning C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. of Meeting of the Association for Computational Linguistics* (2005), pp. 363–370. 5

[FHW16] Frank E., Hall M. A., Witten I. H.: The WEKA Workbench. online appendix for "data mining: Practical machine learning tools and techniques", fourth edition. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf, 2016. 5

[FPSS96] Fayyad U., Piatetsky-Shapiro G., Smyth P.: Knowledge discovery and data mining: Towards a unifying framework. In *Proc. of Int. Conf. on Knowledge Discovery and Data Mining* (1996), KDD'96, AAAI Press, pp. 82–88. 12

[FZCQ17] Fu S., Zhao J., Cui W., Qu H.: Visual analysis of MOOC forums with iForum. *IEEE Trans. on Visualization and Computer Graphics 23*, 1 (2017), 201–210. doi:10.1109/TVCG.2016.2598444. 3

[HC14] Hoque E., Carenini G.: ConVis: A visual text analytic system for exploring blog conversations. *Computer Graphics Forum 33*, 3 (2014), 221–230. doi:10.1111/cgf.12378. 4

[HH08] Hubmann-Haidvogel A. C.: *ThreadVis for Thunderbird: A Thread Visualization Extension for the Mozilla Thunderbird Email client*. Master's thesis, Graz University of Technology, Austria, 2008. 3

[HSS10] Hansen D. L., Shneiderman B., Smith M.: Visualizing threaded conversation networks: Mining message boards and email lists for actionable insights. In *Active Media Technology* (2010), Springer Berlin Heidelberg, pp. 47–62. 4

[KDS*17] Krause J., Dasgupta A., Swartz J., Aphinyanaphongs Y., Bertini E.: A workflow for visual diagnostics of binary classifiers using instance-level explanations. *arXiv preprint: 1705.01968* (2017). 7

[Ker03] Kerr B.: Thread arcs: An email thread visualization. *Proc. IEEE Symp. on Information Visualization* (2003), 211–218. doi:10.1109/INFVIS.2003.1249028. 3, 6

[KF14] Kaastra L. T., Fisher B.: Field experiment methodology for pair analytics. In *Proc. of Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)* (2014), ACM Press, pp. 152–159. 10

[LCC13] Liu Y., Chen F., Chen Y.: Learning thread reply structure on patient forums. *Proc. of Int. Workshop on Data management & Analytics for Healthcare* (2013), 1–4. doi:10.1145/2512410.2512426. 3

[LRC*12] Lee H., Recasens M., Chang A., Surdeanu M., Jurafsky D.: Joint entity and event coreference resolution across documents. *Proc. of Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, July (2012), 489–500. URL: http://www.aclweb.org/anthology/D12-1045. 5

[LWLZ17] Liu S., Wang X., Liu M., Zhu J.: Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics 1*, 1 (2017), 48–56. 4

[LYC*09] Lin C., Yang J.-M., Cai R., Wang X.-J., Wang W.: Simultaneously modeling semantics and structure of threaded discussions: A sparse coding approach and its applications. *Proc. of Int. ACM SIGIR Conference on Research and Development in Information Retrieval* (2009), 131–138. doi:10.1145/1571941.1571966. 3, 5

[LYW*16] Liu S., Yin J., Wang X., Cui W., Cao K., Pei J.: Online visual analytics of text streams. *IEEE Trans. on Visualization and Computer Graphics 22*, 11 (2016), 2451–2466. doi:10.1109/TVCG.2015.2509990. 4

[newa] ABC News – greenhouse gases higher than any time in 800,000 years 'shows definite human effect'. http://mobile.abc.net.au/news/2017-06-01/greenhouse-gases-database-shows-co2-ch4-n2o-rising-relentlessly/8578918. Accessed: 2017-12-03. 10

[newb] Washington Post – Trump is deporting fewer immigrants than obama, including criminals. https://www.washingtonpost.com/local/immigration/trump-is-deporting-fewer-immigrants-than-obama-including-criminals/2017/08/10/d8fa72e4-7e1d-11e7-9d08-b79f191668ed_story.html. Accessed: 2017-12-03. 10

[reda] Reddit. http://reddit.com. Accessed: 2017-12-03. 2

[redb] Reddit – greenhouse gases higher than any time in 800,000 years 'shows definite human effect'. https://www.reddit.com/r/worldnews/comments/6eljby/greenhouse_gases_higher_than_any_time_in_800000/. Accessed: 2017-12-03. 10

[redc] Reddit – trump is deporting fewer immigrants than obama, including criminals. https://www.reddit.com/r/politics/comments/6sy51s/trump_is_deporting_fewer_immigrants_than_obama/. Accessed: 2017-12-03. 10

[RLR*10] Raghunathan K., Lee H., Rangarajan S., Chambers N., Surdeanu M., Jurafsky D., Manning C.: A multi-pass sieve for coreference resolution. In *Proc. EMNLP* (2010). 5

[SCS09] Seo J., Croft W. B., Smith D. a.: Online community search using thread structure. *Conf. on Information and Knowledge Management* (2009), 1907–1910. doi:10.1145/1645953.1646262. 3

[Sin01] Singhal A.: Modern information retrieval: A brief overview. *Bulletin of the IEEE CS Technical Ctte. on Data Engineering 24*, 4 (2001), 1–9. doi:10.1.1.117.7676. 4

[SK16] Sharma H., Kumar S.: A survey on decision tree algorithms of classification in data mining. *Int. J. Science and Research 5*, 4 (2016), 2094–2097. 5

[SMdR07] Schuth A., Marx M., de Rijke M.: Extracting the discussion structure in comments on news-articles. *Proc. of ACM Int. Workshop on Web Information and Data Management* (2007), 97–104. doi:10.1145/1316902.1316919. 3, 9

[TG10] Trampuš M., Grobelnik M.: Visualization of online discussion forums. *Workshop on Pattern Analysis Applications 11* (2010), 134–141. 4

[VRW13] Vehlow C., Reinhardt T., Weiskopf D.: Visualizing fuzzy overlapping communities in networks. *IEEE Trans. on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2486–2495. doi:10.1109/TVCG.2013.232. 7

[WJCR07] Wang Y.-C., Joshi M., Cohen W. W., Rosé C.: Recovering implicit thread structure in newsgroup style conversations. *Artificial Intelligence* (2007), 152–160. 3

[WLK*11] Wang L., Lui M., Kim S. N., Nivre J., Baldwin T.: Predicting thread discourse structure over technical web forums. *Proc. of Conf. on Empirical Methods in Natural Language Processing* (2011), 13–25. 3, 10

[WMZ07] Weiss G., McCarthy K., Zabar B.: Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *Dmin* (2007), 1–7. URL: http://storm.cis.fordham.edu/~gweiss/papers/dmin07-weiss.pdf. 3

[WRK11] Wanner F., Ramm T., Keim D. A.: ForAVis: Explorative user forum analysis. In *Proc. of Int. Conf. on Web Intelligence, Mining and Semantics* (New York, NY, USA, 2011), WIMS '11, ACM, pp. 14:1–14:10. URL: http://doi.acm.org/10.1145/1988688.1988705, doi:10.1145/1988688.1988705. 4

[YGLC13] Yan X., Guo J., Lan Y., Cheng X.: A biterm topic model for short texts. In *Proc. Int. Conf. on World Wide Web* (2013), pp. 1445–1456. URL: http://dl.acm.org/citation.cfm?id=2488388.2488514, doi:10.1145/2488388.2488514. 5