Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

Topic Tomographies (TopTom): a visual approach to distill information from media streams

B. Gobbo[1], D. Balsamo[3,2], M. Mauri[1], P. Bajardi[2], A. Panisson[2], P. Ciuccarelli[1]

[1] Politecnico di Milano, Italy
[2] ISI Foundation, Torino, Italy
[3] University of Turin, Italy

Abstract

In this paper we present TopTom, a digital platform whose goal is to provide analytical and visual solutions for the exploration of a dynamic corpus of user-generated messages and media articles, with the aim of i) distilling the information from thousands of documents in a low-dimensional space of explainable topics, ii) cluster them in a hierarchical fashion while allowing to drill down to details and stories as constituents of the topics, iii) spotting trends and anomalies. TopTom implements a batch processing pipeline able to run both in near-real time with time stamped data from streaming sources and on historical data with a temporal dimension in a cold start mode. The resulting output unfolds along three main axis: time, volume and semantic similarity (i.e. topic hierarchical aggregation). To allow the browsing of data in a multiscale fashion and the identification of anomalous behaviors, three visual metaphors were adopted from biological and medical fields to design visualizations, i.e. the flowing of particles in a coherent stream, tomographic cross sectioning and contrast-like analysis of biological tissues. The platform interface is composed by three main visualizations with coherent and smooth navigation interactions: calendar view, flow view, and temporal cut view. The integration of these three visual models with the multiscale analytic pipeline proposes a novel system for the identification and exploration of topics from unstructured texts. We evaluated the system using a collection of documents about the emerging opioid epidemics in the United States.

CCS Concepts

• ***Human-centered computing*** → ***Visualization;*** • ***Information systems*** → *Document topic models; Expert search;*

1. Introduction

As digital platforms become more and more pervasive in individuals' everyday-life, mining information from media streams is a popular strategy to detect large scale social events, make sense of public debates, customers sentiment and rumour spreading [DWS*12, LG10, Haw09, FR11]. A large body of work has been developed by researchers in different fields to identify patterns and model human behaviours on social media [LBK09, FCYJMT*15, LGRC12] while novel business opportunities were raised from the development of commercial platforms directed to tackle some specific tasks in the context of media monitoring [CMZ10, MHP*12].

Devising representations of textual data that reduces information overload and reveal inter- or intra-document structure extracting latent patterns (i.e. underlying topics) is a powerful yet challenging approach [BNJ03]. Topics have often fuzzy boundaries, and according to the way we analyze them they might be grouped in categories of neighbouring concepts at different granularities. With the lack of a clear-cut semantic definition of topics, it is hard to identify anomalous behaviours since micro-topics are ephemeral and disappear quickly while macro-topics tend to average the debate. Moreover, while most of the existing literature presents solutions for single platforms, often social debates span across different kind of web platforms, each one with its specific logic [Rog17].

The goal of the system presented hereafter is to provide experts with a tool able to monitor multiple data sources and to browse the resulting debate space from macro to micro aggregation, with automatic suggestions on where anomalous behaviours can be found, letting the user to browse smoothly the entire temporal dimension. *TopTom* therefore implements a visual tool aiming at i) distilling the information from thousands of documents in a low-dimensional space of explainable topics, ii) drilling down or zooming out details of stories and facts around the topics, iii) spotting trends and anomalies in terms of information volume around such topics. Hence the contributions of the presented research are twofold: on one hand, to devise an analysis pipeline able to work with heterogeneous textual data sources (micro-blogging, discussion boards, news) on heterogeneous time-scales (hours, days, weeks), on the other hand, to identify a set of visual models able to represent results independently from the source analyzed.

Designing a visual system able to represent the entire set of

features resulting from the algorithmic pipeline is challenging on many sides. First, the richness of the semantic and the temporal aggregations at different levels of granularity makes hard to adopt a single visual model to represent all of them. It is therefore required a multiple-views approach [MPCC13], providing different perspectives on the data through which users can rebuild the overall structure of the topics. Second, the user should be able to drill down from the most aggregated level (macro-topics) to the single documents. Third, there are multiple items (keywords, topics, documents) strongly tied together, and there is the need of showing their mutual relationships. Finally, every level of aggregation might witness different kind of trends and anomalies. In this spirit, in this work, we focus on devising a coherent data visualization interface to explore large and heterogeneous temporal corpora, leveraging information extracted through state of the art topic modeling methodologies that include temporal and hierarchical dimensions and anomaly detection procedures. TopTom has been developed in a flexible and modular fashion, so that different topic modeling and anomaly detection approaches can be plugged in the back-end while preserving the multiple navigational interactions and the full visualization capabilities of the interface. In this perspective, part of the topic modeling pipeline was built on top of existing and referenced methodologies, and the resulting analyses are translated to data structures through an application interface that can be easily consumed by the visualization interface.

Taking as example the American opioid epidemic, we show how TopTom can be used to get an overall picture of the public debate mining texts from newspaper, then have a glimpse of the social reactions from a micro-blogging platform such as Twitter, and finally observe the debate among people dealing with detox and first hand use on discussion boards such as Reddit. The capability of TopTom in dealing both visually and computationally with multi-temporal scale phenomena enables cross-platform analysis, providing the users multiple complementary perspectives on the same subject of interest. Through temporal data coming from the selected sources we can see how debates and discussions evolve over time, identifying anomalous behaviours at different levels of aggregation: a sudden change in the used language, an increase in the volume of texts related to a specific topic, or the extinction of a topic in favour of another one. These kind of collective behaviours are often hard to be spotted on high-volume data streams, since they happen at different levels of aggregation.

2. Related Work

In this section, we review and summarize well-established and common approaches to extract information from social media data by mean of visual tools.

2.1. Topic modeling and social media monitoring

Topic modeling has emerged as one of the most effective methods for classifying, clustering, and retrieving textual data, and has been the object of extensive investigation in the literature. Many topic analysis frameworks are extensions of well known algorithms, considered as state-of-the-art for basic (non temporal and non hierarchical) topic modeling. *Latent Dirichlet Allocation* (LDA) [BNJ03] is the reference for probabilistic topic modeling. *Nonnegative matrix factorization* (NMF) [LS99] is the counterpart of LDA for the matrix factorization community. Both approaches have been considered as starting points to handle temporal and hierarchical topic extraction.

The extraction of signals that expose complex correlations between topics and temporal behaviours has been the object of investigation in more recent years. Blei et. al extended LDA for *Dynamic Topic Models* (DTM) [BL06], and other graphical models have been proposed to incorporate the temporal dimension in LDA [WM06, Kaw11, WAB12, HCLB10, NDLU07, ACL*12]. Saha and Sindhwani [SS12] propose an algorithm based on NMF that captures and tracks topics over time at the daily temporal scale.

Among the topic modeling approaches aforementioned, the number of extracted topics can be either nonparametric and estimated by the model, or might be informed as a parameter of the model. The a priori knowledge of the number of topics is almost never given, and there are not many approaches that can be used to validate the estimated number of topics. To avoid producing a fixed number of extracted topics, hierarchical representations of topics can be used. Among the works that use nonparametric models, it was shown that the *nested Chinese restaurant process* [GJTB04] can be used as a nonparametric prior for a hierarchical extension to LDA. HierarchicalTopic [DYW*13] was proposed to deal with a large number of topics by constructing a topic hierarchy based on a list of topics, and to facilitate their representation.

The methodology for the temporal topic extraction proposed in this paper as described in Section 3.2 is an extension of [PGQC14]. It combines the extraction of temporal topics using nonnegative matrix factorization and a hierarchical representation of the extracted topics.

2.2. Visualization and interaction

Due to the richness and the complexity of data resulting from topic analysis, in literature there is a wide use of data visualization to make information easy to explore and understand. A common approach is to develop a main visualization for the overall perspective, then providing details adding information through interaction. Usually the combination of elements on multiple panels on the interface might create an information overload [ZCW*14, DYW*13, CLT*11, HHKE16]. *TopTom* improves the interaction flow keeping only elements essential to the users and focusing on views connections. Topic visualization usually deals with two broad categories, static and dynamic, being the latter one the main focus of this paper.

In most of the solutions the main visualization adopts a time-based visual models, using the horizontal axis to represent time and the vertical one to represent volume of documents. *Stacked graphs*, *area charts* and *stream graphs* [BW08] are widely used to visually represent the change in volume of the same object [WLS*10, LYK*12] When dealing with topics aggregating over time *alluvial diagrams* [RB10], are used to provide an idea of flows tearing apart and reconnecting [CLT*11, SWL*14, LYW*16].

To enable users to drill down from the global overview to the keywords, *tag clouds* are well-known and accepted visual elements for representing topics composition [BSH*10, DWS*12,

SWL*14], with some variations on the way they are displayed. Tiara [WLS*10] proposes an overview using a *streamgraph*, and creates visual text summaries by putting texts directly inside graphic elements. TextFlow [CLT*11] uses juxtapose visualizations: a *tag cloud* and a *timeline* placed on the bottom side of the interface encode words and sentences correlation. By now, *force layout algorithms* are also used in the literature to highlight topic similarities [LWC*14] but they have never been used as dynamic elements in topic modeling visualizations.

HierarchicalTopics [DYW*13] uses a *dendrogram visualization* next to a *streamgraph* to show the hierarchical structure to the users, and leverages on the interaction to allow them to explore sub-topics contained in a main one. Cui et al. [CLWW14] solves the issue visually hiding the hierarchical structure and proposing to the user some algorithmically identified pre-set cuts.

Anomalous behaviors of topics are usually represented as an additional layer to the overview visualization, with two main graphic approaches. The first one uses an additional layer of glyphs, as in TextFlow [CLT*11] A second approach is to color-code single visual marks [ZCW*14] or shapes defined by multiple points over time using gradients [PGQC14].

As described in section 4, the system adopts an approach based on multiple alternative views, with two main visualizations extending the *streamgraphs* and *forces-based tag cloud* visual models, that smoothly connect each other avoiding information overload and providing only useful information for the users. Moreover TopTom allows users to immediately detect and browse anomalous topics from different sources, temporal granularities and vocabularies.

3. Methods

The system is composed of a data ingestion phase (i.e. connectors to streaming/rest API or repositories of target data sources), a preprocessing pipeline aimed at annotating and normalizing the raw texts in structured JSONs, a data lake for data storage and an unsupervised machine learning pipeline to identify emerging topics running in timed batches. The resulting output is a summary file containing information about temporal topics described in hierarchical fashion, keywords and documents representative of such topics, and information about potential outlying trends that is finally fed into the visualization interface specifically designed according to the system requirements described in Section 4 (Figure 1).

TopTom platform lies on the following technological stack: the analytics pipeline has been developed in Python2.7, the data lake is based on MongoDB (and connectors for ElasticSearch has been developed as well), the APIs are Flask-based and the visualization interface is developed on D3.js and React.js [†].

3.1. Ingestion and pre-processing

The present pipeline has been tested for three major sources of information, namely Twitter (a micro-blogging platform), Reddit (a social news aggregation and discussion website) and GDELT (a

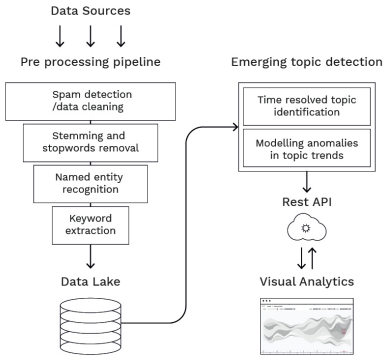

Figure 1: ***Overview of TopTom architecture****. Once that the system is configured, it is able to run the entire pipeline autonomously providing near-real time results to the visual interface.*

project that monitors worldwide web news in over 100 languages), but it can be extended to different media sources. Given the different nature of content and users' behavior on such platforms, for the case study presented in section 5, different data gathering procedures have been implemented. We relied on Twitter streaming API based on a list of keywords, scheduled REST API for Reddit posts on selected subreddits and scheduled massive crawling of news containing themes of interest from GDELT. To limit the well-known issue of Twitter social-bots spamming and potentially polluting the corpus under investigation [FVD*16], a basic semi-supervised anti-spam filter has been implemented. The system trains a naive Bayes classifier on manually annotated messages with the addition of a cascade set of rules to tag users as bots if they i) have high frequency activity, ii) re-shared spam messages, iii) follow other bots, iv) have a synchronized tweeting activity with other users, thus suggesting a non-human behavior [VFD*17]. Such anti-spam filter is periodically re-applied to the whole twitter corpus while a black-list of bots is progressively updated.
At this stage, every document (i.e. tweet, Reddit submission, news article) is parsed, stemmed and cleaned from stop-words while persons/locations/organizations are identified and annotated as "entities". Entities are also enriched by important contextual keywords identified via *TextRank* algorithm [MT04]. The results of the preprocessing pipeline are finally stored on the data lake (MongoDB-based).

3.2. Analytics Pipeline

The methodology used in this work, an extension of [PGQC14], is able to extract a hierarchical representation of temporal structures from a text dataset $\mathbb{D}$ with N documents, each document associated to a timestamp. Each one of these structures is represented as a *temporal topic*.

The *temporal span* indicates the interval between the minimum and maximum document timestamps selected for the analysis. Once a *temporal scale* is set for the analysis, the dataset $\mathbb{D}$ is divided in T *time chunks* with the duration of the time scale (e.g. a dataset with temporal span of 24 hours is processed using a time scale of one hour, hence it is divided into 24 time chunks of one

† https://github.com/densitydesign/toptom-frontend

hour). The current prototype of TopTom is designed to process data up to a temporal span of two months with a temporal scale of two days, but larger temporal spans with different temporal scales can be included in the platform through configuration.

We extract from $\mathbb{D}$ a vocabulary with V terms. Each document is then represented as a term count vector $\mathbf{x} \in \mathbb{R}^V$. These vectors form a document-term count matrix $\mathbf{X} \in \mathbb{R}^{N \times V}$, where each position x_{ij} represents the number of times the term j appears in document i. For GDELT data, the entire pipeline is computed twice: the first using all the words of the documents as terms of the vocabulary, the second using a vocabulary built only on the entities extracted from the text as mentioned above, yielding two complementary perspectives.

3.2.1. Topic extraction

The dataset $\mathbb{D}$ is divided in T time chunks. Each time chunk t contains a subset $(\mathbb{D}^{(t)}$ with $N^{(t)}$ documents, and is represented as a term count matrix $\mathbf{X}^{(t)}$. Each $\mathbf{X}^{(t)}$ is modeled as a set of K topics, where K might be different for each time chunk. As discussed in Section 2, among the many methods for topic modeling in the literature, in this work we chose a method based on *nonnegative matrix factorization* to decompose $\mathbf{X}^{(t)}$ in two matrices: a matrix of left vectors $\mathbf{W}^{(t)} \in \mathbb{R}^{N^{(t)} \times K}$ and a matrix of right vectors $\mathbf{H}^{(t)} \in \mathbb{R}^{K \times V}$, where K is the number of topics used in the decomposition, $N^{(t)}$ is the number of documents in time chunk t and V is the number of terms in the vocabulary (Figure 2a). Nonnegative factorization minimizes the following error function,

$$\min_{W,H} \|\mathbf{X}^{(t)} - \mathbf{W}^{(t)}\mathbf{H}^{(t)}\|_F^2,$$

where $\|\|_F^2$ is the Frobenius norm, subject to the constraint that the values in $\mathbf{W}^{(t)}$ and $\mathbf{H}^{(t)}$ must be nonnegative. The nonnegative factorization is achieved using the *projected gradient method* with sparseness constraints, as described in [Lin07, Hoy04].

The topic-term matrix $\mathbf{H}^{(t)}$ stores the term vectors of the extracted topics at time chunk t. The document-topic matrix $\mathbf{W}^{(t)}$ stores the strength with which each document is associated to each topic. Thus, a single topic k can be represented as two vectors $\mathbf{w}$ and $\mathbf{h}$, where $\mathbf{w}$ is a weight vector, column of $\mathbf{W}^{(t)}$ and $\mathbf{h}$ is a term vector, row of $\mathbf{H}^{(t)}$. The output of this phase is a set of topics for each time chunk t, each topic represented as its corresponding weight vector $\mathbf{w}$ and term vector $\mathbf{h}$.

3.2.2. Agglomerative clustering

To track the evolution of topics over time, we merge the topics extracted at each time chunk based on their lexical similarity, forming clusters of topics. In this phase, each term vector $\mathbf{h}$ of topics extracted in the previous phase are treated as data points in an *agglomerative clustering algorithm*. We used an approach similar to UPGMA [SM58], that at each step aggregates the two most similar clusters into a higher-level one, with the following differences: (1) distances between children are defined as the cosine distance between their vectors and (2) cluster distance is defined in terms of "complete" linkage: that is, the distance between two clusters c_1 and c_2 is defined as the maximum of all pair-wise distances between the children of c_1 and those of c_2.

The agglomerative clustering produces a hierarchical structure represented by a dendrogram (Figure 2b) that can be cut at a given similarity threshold to yield a set of clusters at a chosen level of detail. That is, by varying the similarity threshold of the cut we can go from a coarse-grained topical structure with few high-level clusters to a fine-grained topical structure with many small clusters composed by single topics. The output of this phase is a hierarchical representation of the topics that at each level combines a pair of nodes, each node representing a single topic (a leaf) or an aggregation of topics (a cluster). In the next, to distinguish between leaves and clusters, we are going to represent the leaves with their corresponding time chunk as k_i^t and the clusters just as k_i.

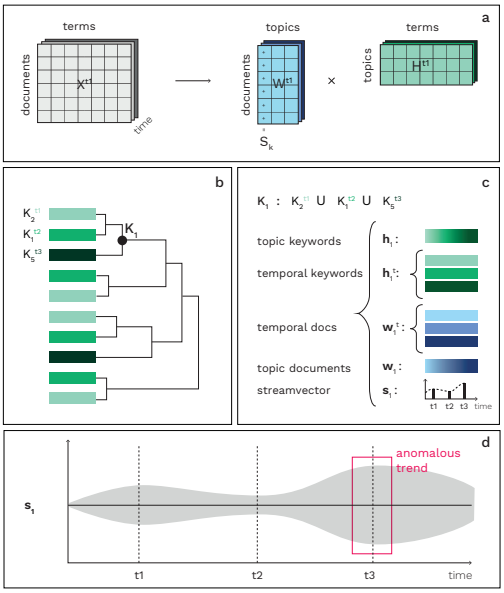

Figure 2: ***Analytics pipeline****. a) For each temporal chunk the document-term matrix $X^{(t)}$ is decomposed in a document-topic matrix $W^{(t)}$ and a topic-term matrix $H^{(t)}$ using NMF. b) The agglomerative clustering produces a hierarchical representation of the topics. c) Each node in the hierarchy is represented as a temporal topic. d) An anomalous trend is detected in the stream vector s of a temporal topic.*

3.2.3. Hierarchical representation of temporal topics

At this step, we build the temporal representation associated to each node of the hierarchy, i.e., the *temporal topics*. We start by building the temporal topic representation for the leaves, and then we proceed to the agglomeration for building the temporal representation for the clusters of the hierarchy. For the two types of nodes in the dendrogram (leaves and clusters), we proceed as follows:
For a leaf k_i^t, we keep the term vector $\mathbf{h} \in \mathbb{R}^V$ without any change, while the respective weight vector $\mathbf{w} \in \mathbb{R}^{N^{(t)}}$ is extended to a weight vector of dimension N where all weights that are not associated to the $N^{(t)}$ documents are set to 0.

For a cluster k_i, we compute its respective term vector $\mathbf{h} \in \mathbb{R}^V$ as a weighted average of the term vectors of its children. Also, we compute its respective weight vector $\mathbf{w} \in \mathbb{R}^N$ as the sum of the weight vectors of its children.

At this point, each node in the hierarchy is represented as a couple of vectors $\mathbf{w}$ and $\mathbf{h}$ respectively in the space of documents (**topic documents**) and in the space of terms (**topic keywords**). This static representation is finally completed by its temporal representation with the following information (Figure 2c):

- **stream vector**, a vector $\mathbf{s} \in \mathbb{R}^T$ where each element s_t is the sum of the weights from $\mathbf{w}$ associated to the $N^{(t)}$ documents at each time chunk t (i.e. the volume of the temporal topic at time t).
- **temporal keywords**, a term vector $\mathbf{h}^t$ for each time chunk: for each t the term vectors $\mathbf{h}$ of all the children of the topic represented in t are averaged. If the topic is not represented in t, a null vector is used. Since leaves are extracted from a single time chunk, there's only one non-null $\mathbf{h}^t$ for a leave k_i^t.
- **temporal documents**, a weight vector $\mathbf{w}^t$ for each time chunk: for each t, the weights of the $N^{(t)}$ documents in that time chunk are extracted from $\mathbf{w}$.

It is worth highlighting the role of the temporal *span* and the role of the temporal *scale* in the analysis. The former indicates the whole temporal domain of the documents included in the analysis, and it is interconnected with the overall length of events in the corpus, the latter indicates the time resolution of the analysis and it is related to the speed at which changes happen in the topics. When a larger temporal span is considered, a bigger temporal scale is chosen, and vice versa, ranging from a maximum of two months span with two days scale to a minimum of 24 hours span with hourly resolution. The larger the temporal span the more the analysis is capable of grasping long-term topics and general trends, at cost of losing local details due to larger time chunks. Conversely, the shorter the temporal scale the more the results are capable of representing fast changes and short-term events, at the cost of neglecting a bigger picture. Moreover, each source has its own peculiar time scale and interaction time, requiring different typical time scales and spans: Twitter is the most versatile source, suitable to investigate short-term dynamics [LGRC12] with a minimum of 24 hours span and 1 hour resolution, GDELT and news in general have a quick decrease in thread volume [LBK09] so that we rely on a minimum resolution of 1 week span and 12 hours resolution due to the almost-daily nature of news, whilst Reddit covers longer periods, with one day resolution over a month span, as discussed in Section 6. In choosing scales and temporal spans, for the sake of readability of the visual interface, we always used a number of time chunks ranging between 20 and 40.

Since topic modeling with nonnegative matrix factorization is able to detect patterns over aggregated data, this approach is suitable for batch or near-real time processing, where the perceived delay before updated information arrive to the visual interface is in the order of the temporal scale (e.g. hours or minutes). The pipeline has to first accumulate the documents that arrive during the last time interval, and then the topics for the new interval can be extracted. After that, the new topics are connected to the existing topic stream by updating the cluster hierarchy, and then the visual interface is updated with the new results accordingly. The perceived delay can be made arbitrarily small by choosing smaller temporal scales, combined with online nonnegative matrix factorization approaches [GTLY12].

3.2.4. Anomaly detection

The final step in the analysis is the detection of anomalies in the temporal structure of topics. In order to avoid signalling anomalies on little meaningful topics (e.g., high level aggregations that may merge unrelated topics, lower aggregations that may separate topics that otherwise could be regarded as the same continuous topic over time) we evaluate the presence of anomalies only on a subset of topics $\tilde{k}$, considered as *consistent aggregations* with respect to the structure of the hierarchy.

The algorithm that extracts such subset of topics is implemented as follows. We consider the hierarchy of clusters ordered in a dendrogram, where the leave nodes k^t are the topics to be aggregated, r is the root node, the nodes k in between represent aggregations and the distance between each node $z(k_i)$ is the cophenetic distance between its two children nodes. The idea behind this algorithm, inspired by [RL14], is to "cut" each leaf branch before its most dissimilar aggregation (i.e., before the topic is merged with the most unrelated among its closest topics) and to cluster all the nodes ending at the same aggregation. Thus, taken the ordered list $\mathbf{p}^i$ of *parent nodes* of a leaf (i.e. all the nodes on the branch connecting k_i^t to the root r), we evaluate the shift in cophenetic distance $\delta z(p_j^i) = z(p_{j+1}^i) - z(p_j^i)$ between each node and the following, labelling the node k_j as a *consistent aggregation* if $\delta z(p_j^i)$ is maximal along the whole path. We repeat the procedure to assign a consistent aggregation to each leaf and finally consider all the leaves with the same consistent aggregation as part of the same cluster $\tilde{k}_i$. This method yields a partition of clusters that can be considered as the result of cutting a dendrogram at different heights in different areas.

We cover two types of anomalies: **volume increase** and **volume decrease**, reflecting a trend (increase or decrease) in the volume of documents associated to a topic. Both types of anomalies are extracted from the stream vector $\mathbf{s}$ of each temporal topic in the set of consistent topics $\tilde{k}$ (Figure 2d): a large increase from values of $\mathbf{s}$ at time t to $t+1$ represents an increase in the number of documents associated to that topic. Such a volume increase is considered an anomaly if it exceeds the mean volume increase for $\mathbf{s}$ and it is classified in different levels based on standard deviation from its mean value, e.g., a five fold increase in the value of $\mathbf{s}$ has higher level than a two fold increase. The result of this phase is a list of anomalies, each anomaly with the following information:

- the type of the anomaly (volume increase, volume decrease);
- the level of the anomaly (e.g., high, low, medium);
- the temporal topic in which the anomaly was detected;
- the time chunk in which the anomaly was observed.

3.2.5. Processed information

The algorithmic pipeline distills the information contained in the corpus, returning a highly informative set of data structures summarized in JSONs and provided via API to the visual analytics framework. Each *temporal topic* is represented with its attributes, the *stream vector*, the *anomalies*, the *topic keywords* and *topic documents* as pairs of terms (or documents) and their associated weights

in **h** (or **w**), and the *temporal keywords* and *temporal documents* as lists with T elements, where each element is a list of terms (or documents) with the same structure as *topic keywords* and *topic documents*. The weights of keywords and documents are used to retain only the most relevant ones for both topic documents, temporal documents, topic keywords and temporal keywords. Additional information about the hierarchical structure of temporal topics and daily aggregated anomalies are further summarized and provided via API.

4. Visual Analytics

4.1. Design Requirements

In order to define the main actions that should be available through the interface, we implemented a co-design process that involved a group of end users. They were experts in the field of public health and epidemiology, familiar with similar sets of data since in the past they have already worked on granular data about the opioid epidemics in USA, familiar with basic visualization techniques but unfamiliar with interactive browsers including a set of different visual models. During this phase, eight requirements for the interface have been defined according to users' characteristics, knowledge and skills [WGK14].

- **R1** - **The big picture**: a near-real time "sensor" that shows at first the evolution of the discussions on social media and news about a group of topics within a selected time-frame. A first (over)view should depict the selected temporal data in terms of (visual) importance: (i) volume of contents for each topic; (ii) type of source media; (iii) anomalous behaviors of identified topics.
- **R2** - **Heterogeneous corpora browsing**: an interface that, independently from the data source, enables the users to distinguish topics and explore related documents, allowing them to choose the corpus according to their current needs.
- **R3** - **Topics along time exploration**: An interaction pattern that allows the selection of the time span and its granularity should be designed in order to temporally locate topics and identify both short-term and long term patterns in online conversation.
- **R4** - **Hierarchical structure exploration**: A representation of the hidden hierarchical structure behind the topic clustering should be incorporated in the visual exploration strategy, allowing the user to identify how topics appear at different levels of semantic similarity.
- **R5** - **Anomalies location detection along the hierarchical aggregation**: A visual variable can be used to highlight anomalous situations both on a single hierarchical layer and along the hierarchical structure for as much as the users want to immediately spot throughout all the views where anomalies are located.
- **R6** - **Direct handling of contents**: acting on the visual representations of data has been recognized as a potential driver for engagement. Interaction patterns that enable a direct manipulation of graphical objects in the views should be designed.
- **R7** - **Interactive drill down to documents**: a visual system that enables the users to access multiple dimensions of the data unwrapping the hierarchical structure of topics.
- **R8** - **Familiar visual model**: according to the preliminary co-design process, the users are familiar with specific and well established visual models; the risk of creating a barrier for adoption with the use of unconventional visual models has also been discussed. The first steps of the user experience should be designed using the most common and solid visual models in topic modeling literature and cases.

While in the literature review we found visual solutions for some of the requirements, no platform was able to provide all of them within a unified and coherent user experience and interface. Our design process therefore started from the declination of consistent metaphors into a visual grammar able to encode the multiple dimension of the data.

4.2. Visual Metaphors and Encoding

The need to produce an organic system of interaction has led us to design a system based on three main visual metaphors: the flowing of particles, the section of three-dimensional objects as in tomographies, and the contrast analysis of biological tissues.

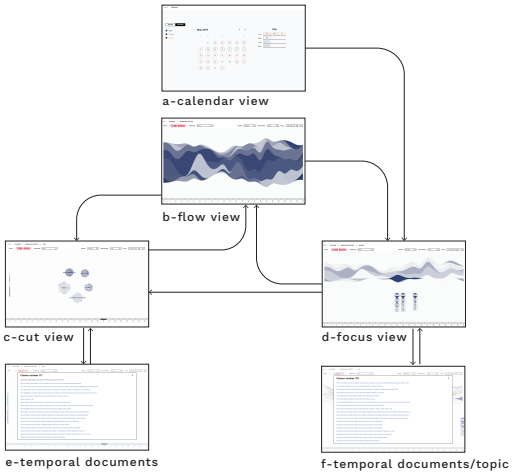

Figure 3: ***TopTom interaction flow*** *A visual summary of TopTom interaction flow. From the list of anomalies on the calendar view (a) users access the focus view (d) with the highlighted anomaly. From the focus view it is possible to go back to the flow view (d) or look inside a single time chunk by browsing the cut-view (c). From both (c) and (d) users have access to the lists of topic documents.*

Conceptually, the data resulting from analysis described in Section 3 is seen as three-dimensional object that develops over three axes: time, volume, and hierarchical aggregation (Figure 5). In any point of this ideal 3D object an anomalous behavior may occur.

We summarized in (Figure 3) the visual architecture and the interactions between the different views implemented in TopTom.

4.2.1. Flow view

We started from the biological metaphor identified by Susan Havre's [HHN99], adopting a *streamgraph* to represent the first two dimensions: time and volume.

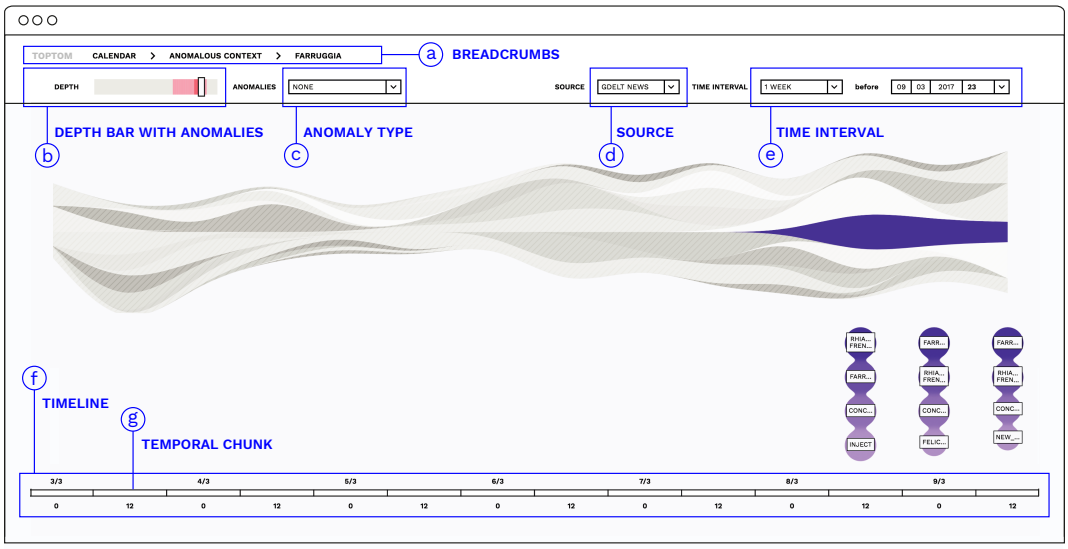

Figure 4: ***Focus view and interface overview***. *A screen-shot of the application opened on the focus view (see section 4.2.2), with interface components highlighted. A breadcrumbs system allows users to browse different views (a). As a litmus paper, the depth bar (b) indicates where anomalies are located along the hierarchical structure. Users can change anomaly type (c), source (d) and time interval (e) using the drop-down menus on the top. Each time chunk (g) on the interactive timeline (f) allows to explore the temporal cut view (see section 4.2.3).*

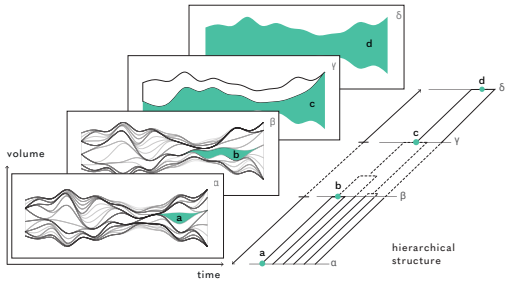

Figure 5: ***Tomographic sections***. *An abstract representation of the hierarchical structure. The foreground sketch represents the most detailed aspect of the streamgraph. Changing the depth level, the user can ideally move on the z axis and see how topics aggregate.*

Such model simplifies the user's task of following individual topics over time having an immediate overview of the whole volume of debate (R1, R3). Moreover, the users were familiar with this kind of time-based visual models (R7).

Each *stream vector* provided by the APIs (see section 3.2.3) is represented as a colored stream, and its size represents the volume for each time chunk; colors are assigned to create a high contrast among adjacent areas. This visualization is able to represent the temporal unfolding of topics for a given depth in the hierarchical structure. To disaggregate the overall discussion at different levels we adopted as metaphor the Computed Tomography, as a way to explore the three-dimensional space by slicing over the different level of aggregation (Figure 5). In order to avoid an information overload and to end up with an overly complicated user interface, the whole hierarchical structure as sketched in (Figure 5) is not directly available to the user, whereas the interface allows for dragging the depth slider (Figure 4b) so that users can choose the level of aggregation according to the presence of anomalies (R4,R5). The lowest depth shows a single aggregated topic, while the highest depth shows the most fragmented view of the topics.

4.2.2. Focus view

Selecting a single stream in the *flow view* (Figure 4) users can isolate the corresponding topic, and read the first four most representative keywords related to each time chunk (*temporal keywords*, see section 3). To leave space for keywords the *streamgraph* collapses and moves on the top side of the interface.

- **Selected area**: The overall alignment of the streamgraph is centred on the selected stream, allowing a better reading of changes in volume. Furthermore, the other streams are grayed out, making the selected area more readable .
- **Connected circles**: for each time chunk in which the selected topic exists, a column of circles represents the four most relevant keywords. The size of the circles indicates the relevance of each keyword.

Clicking again on the selected stream shows the list of related documents from *topic documents*. In the same way, clicking on the connected circles shows the list of documents related to that specific time chunk. This drill-down view is created using data contained in *temporal documents* (see section 3.2.3).

The *focus view* can be accessed as well from the *calendar view*: if an anomaly is selected in the calendar, the *focus view* is opened and the topic associated to the anomalous behavior is highlighted.

4.2.3. Temporal cut view

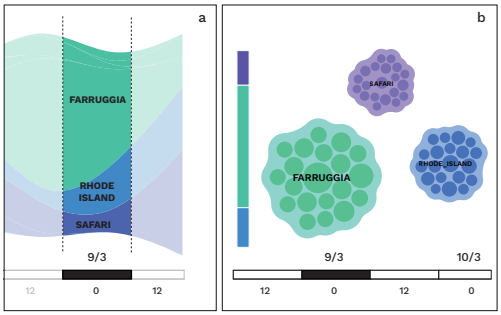

Figure 6: ***Temporal cut view***. *When users select a time chunk (Figure 4f) the temporal cut view shows the details of topics composition (b).*

The Computed Tomography metaphor is extended also on the time axis, allowing users to cut conceptually the streamgraph for a given time chunk and look at the keywords composing each topic. Each time chunk on the timeline (Figure 6) is a button (R6) which allows the user to look at the topics and keywords composing the stream in that precise time chunk. This view is informed by the data defined in *temporal keywords* (see section 3.2.3).

The generated view appears as a slice of the *streamgraph*, where each topic is depicted as a hull containing circles representing the related keywords in that specific time chunk.

- **Outer contour**: the outer contours represent the topics in the selected time chunk. Contours are computed as *metaballs*, areas following the shape of contained circles, to stress the organic metaphor and visually communicate a continuity among the items especially when users move from one time chunk to another.
- **Inner circles**: each hull contains 20 smaller circles that represent the most relevant keywords for that topic. Rolling over the circles, a label with the keywords appears.
- **Stacked bars**: a visual link to the *stream view*.

The user can switch from one time chunk to another observing how the topics change in keywords composition and volume.

4.2.4. Anomalies as contrast agent

The third adopted visual metaphor to show topics' behaviors also come from medical field: the anomalies in the tool are emphasized (R5) by a different color tone that highlights location and intensity acting as a red contrast liquid in a Computed Tomography. The chromatic scale varies from red to pink to show the anomaly intensity, different shades of gray are used to represent the non-anomalous elements. This filter can be activated across all the views in the interface (Figure 7) helping the user to follow anomalous behaviors in the different visual models and on different hierarchical levels.

4.2.5. Depth slider

As a litmus filter, the slider (Figure 8) bar shows the location of anomalies along the cluster hierarchy where the colors encode the

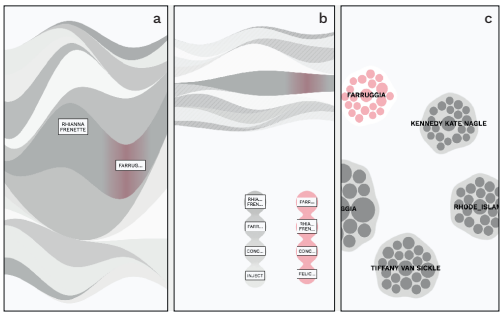

Figure 7: ***Anomalies***. *When users select a specific type of anomaly from the dropdown menu (Figure 4c), it will be highlighted along the flow view (a) the focus view (b) and the temporal cut view (c).*

strength of anomalies, helping the users to choose potentially interesting levels of topic hierarchy that call for closer investigation. Moving the slider from left to right the visualization shows different level of topic aggregation both in the focus view and in the temporal cut view.

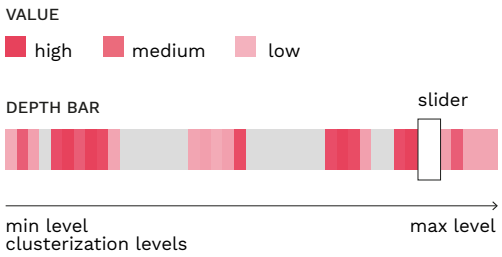

Figure 8: ***Depth slider.*** *Users can navigate the hierarchical structure of topics on the different views through the depth slider.*

4.2.6. Calendar view

The presented views are useful to explore a given temporal span from multiple perspectives, but users don't have an overview of the anomalous behaviors across all the computed temporal spans, which is one of the identified needs (R3, R5). To provide an overview on the overall analysis we designed a calendar view as a summary of anomalous behavior in topics across different data sources. From this view the user can choose the type of analysis (entities or keywords) and activate the different sources (Twitter, Reddit and Gdelt) as overlay layers. Each days that contains at least one anomaly is surrounded by a circular glyph and each circle is divided in 24 slices that represent hours in a day (Figure 9). Selecting a day and a specific topic the user is directly addressed to the *focus view* highlighting the anomaly for that topic. (R2)

5. Results

The analytic core of *TopTom* system is designed to work in a fully unsupervised fashion to help analysts in the discovery of stories and emerging topics in large and heterogeneous corpora. In the following section we report a qualitative assessment of the prototype to a

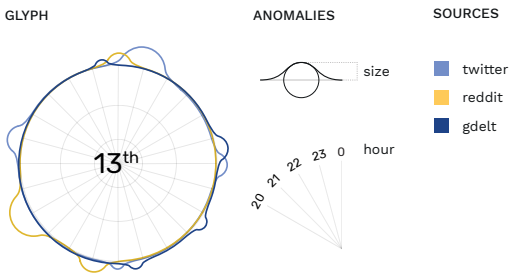

Figure 9: ***Glyph for the calendar view.*** *In the calendar view, each day that contains at least one anomaly is surrounded by a circular glyph.*

specific use case, showing how it can be used to increase situational awareness and identify salient facts. In particular, we will show by examples how the algorithmic pipeline and the visual analytics tool is robust and useful to explore heterogeneous corpora.

5.1. Case study and user scenarios

Starting from the '80s, opioid-based pain killers gained popularity in the United States and physicians have increasingly prescribed them for chronic conditions besides cancer treatment or post-surgery recovery [LMS*17]. Such highly addictive treatments led to a subsequent soar of illegal drugs users (mainly heroin and fentanyl) [MBK*14] triggering an exponential growth of overdose deaths that reaches epidemic proportions in the last years [DSC*15, Rud16]. While epidemiological data and official statistics provide crucial information to policy makers and researchers, there is a huge effort in monitoring the phenomenon through a bottom-up approach. To gain insights from first-hand users to promptly understand emerging trends, risk factors and behavioral shifts, researchers and doctors enrolled cohorts and volunteers to perform detailed interviews and follow ups. To support such endeavour, digital epidemiology techniques [SBB*12, Haw09] to scale standard methodologies to monitor health conditions at the population level in near-real time at low expenses might be extremely valuable. In this section we present the application of TopTom to explore, understand and track trends and major events around the opioid crisis in US (**`http://labs.densitydesign.org/toptom`**).

5.1.1. One month of distilled news

The system has been fed with news collected via GDELT project, crawling all the articles related to US stories between January and March 2017 and belonging to "Drugs and Narcotics" or "Alcohol and Substance Abuse" categories or the url contained the word "opioid". Moreover, an additional filter was implemented to retain articles closely related to the topic of interest, i.e. if their url contained at least one of the following words: opioid, epidemic, heroin, painkiller, prescription, naloxone, abuse, overdose, addict, recover. A total of 31429 articles have been identified but only 60% of them where available for crawling at the reported urls and were indeed collected and analyzed. In the following section an overview of the major events occurred in March 2017 with a daily temporal resolution is presented.

From the calendar view, an anomaly tagged as "oxycontin", a popular opioid pain killer that can cause severe addiction, emerged on March 2nd. As described in the previous sections, clicking on the anomaly we open the *flow view* highlighting the temporal unfolding of the topic at the week level, having direct access to the titles and documents related to the topic. We immediately recognized an overall pattern related to lawmakers in different states (California, North Carolina,...) proposing more restrictions on prescriptions and taxes to address opioid abuse crisis to impact the widely (over)used OxyContin and other opioids. Zooming out the anomalous topic, climbing the hierarchy and aggregating similar topics together, we widened the view to the overall trend about stricter legislations, as reported by the titles of articles related to the most representative documents of the topic (*"New enforcement increases penalties for dealers of deadly opioids", "Wisconsin lawmakers take up bills to combat opioid addiction", "Missouri Senate passes bill for prescription drug tracking"*). Observing a fairly aggregated view of the entire month we could track bills and novel restrictions to contain the opioid epidemic in different US States (e.g. *"NYC officials announce new plan to fight opioid epidemic", "Maryland lawmakers outline package on opioid overdoses"*) and by drilling down into the topic and focusing on specific days we were able to distinguish between policies implementation and results (e.g. Republican plan for Medicaid that might have a detrimental effect by hurting opioid abusers), and official advices (e.g on March 8th officials from the Boston area warning of dangerous batch of heroin circulating). Getting back to the calendar view, we focused on March 9th and explored the anomaly identified as "Farruggia". Clicking on the topic, the whole story unfolded clearly: it was about a woman, Felicia Farruggia, taking heroin during labour with the help of a friend of her. The most important documents were about the event itself whereas the last ones were, more in general, related to the topic of children exposed to drugs. While the "Farruggia" topic was still very well defined in the following time window, an increasing amount of articles related to children and drugs were added to the topic. It is worth noting that the topic was still very robust even when the entire month was considered. Focusing on the single day (March 8th) three major clusters emerged. Two of them were still related to coherent events (the largest topic about Farruggia, the second one about the actress "Eliza Dushku" revealing her former drug addiction). On the other hand, a spurious topic emerged as a collection of off-topic news (e.g *"Cheese is so addictive, one doctor calls it dairy crack"*) Following the topics evolving over time by looking at the *temporal cut view* of the following day, the Farruggia-topic and Dushku-topic were still very well defined, with more article news describing the facts and a new topic emerged, collecting articles showing an increasing evidence that preoperative opioid use may lead to complications, readmissions and high costs post surgery and about a funding effort to expand access to naloxone, an overdose antidote, promoting state assistance for the development and maintenance of prescription drug databases. On March 10th the media attention about the stories emerged in the previous two days was finally over.

5.1.2. One month of distilled reddit posts

Reddit submissions from thematic sections (i.e. subreddits) dedicated to opioid-related discussions were crawled. More than 66k

submissions spanning 5 months have been collected; hereafter a monthly time span is investigated at daily temporal resolution. On May 26th, we observed an anomalous trend about "tramadol", with documents focusing on that specific synthetic opioid mainly from a harm-reduction/withdrawal point of view (e.g. *"Day 2 with no tramadol", "how to stop nausea from tramadol?",*). From the topic documents, we immediately recognized some of the common side effects of tramadol (i.e. nausea), an advice-seeking behavior from users that were going to take a combination of drugs and that tramadol might be used to quit stronger addictive drugs (i.e. oxycodone). Exploring well sustained temporal topics looking at a more aggregated view, we observed that the reddit community is also very supportive towards people who are trying to quit the severe opioid addiction. Also, we found that a stable and quite important topic along the entire month was about "day", "feel", "clean" and "good": the documents belonging to this topic were all about users logging their daily experience of withdrawal and receiving huge encouragement from other users. Moving forward, a topic about "bag" and "stamp" clearly emerged: the small bags of glossy paper (glassine) that are used as envelops to sell heroin and illicit drugs, are often branded through artistic stamps so that drug users can actually recognize the different doses in terms of quality and effects, so a fairly large corpus of posts was about reviews, discussions, or HAT (Has anyone tried?) threads on specific heroin doses. Finally, we recognized two topics mainly related about specific heroin types (i.e. "black tar" and "china white") and discussions about route of administration of painkillers (i.e. "snort", "smoke", "patch"). The ability to distill such detailed first hand usage habits is extremely valuable to have a deeper understanding about the opioid crisis and to map harmful behaviors in order to inform policy makers with field data.

5.1.3. One day of distilled tweets

Quantitative analysis have already shown that Twitter is massively flooded by headline news [KLPM10], thus limiting the original content that can be mined from the analysis of its stream while enabling researchers and policy makers to measure the emotional impact of news on the population. For the presented case-study, a wide range of keywords regarding commercial names and slang names of specific drugs names were used as filter to collect almost 1 million tweets over 13 non-consecutive days between the end of April 2017 and mid May 2017. From the calendar view, we observed an anomaly rising on April 23th about "Cherokee". Focusing on a daily time span with hourly resolution, we observed that the topic appeared as a weak signal early in the day (between 8 and 9 AM) and was about *Cherokee Nation* suing CVS Health, Walgreens, and other drug companies and retailers, alleging the companies didn't do enough to stop prescription painkillers from flooding the tribal community and creating a crisis of opioid addiction. Thirteen hours later we witnessed an anomalous trend highlighting that the discussion sparked on Twitter with a large volume of messages. Exploring the same day at the same hierarchical level, a thin but well sustained topic whose main temporal keywords were "detox" and "percocet"(a popular painkiller often misused for recreational activities) crossed several hours. Although the two terms had perfectly sense together, after a brief exploration of the topic documents it was clear that the underlying twitter debate was not that straight forward. On one hand there was a lot of buzz about "detox" in general terms and not necessarily related to the opioid crisis (alcohol, junk food ...), then there were posts focusing about the drug first hand abuse (e.g. *"I need a Percocet", "Percocet got me high"*). Interestingly enough, the most important documents of such topic were about news related to the musician Prince (*"Prince was in rehab for his Percocet addiction", "Prince reportedly had a long-standing addiction to Percocet"*) and represented the bridge between "percocet" and "detox" topics. Leveraging the tomographic approach and zooming into the hierarchical structure of topics, the "detox" topic and the "percocet" one split into separate components, with the latter carrying specific information about first hand use and habits reported by users.

6. Evaluation and Discussion

Following a qualitative interview and debriefing with the experts listed in Section 4.1, we collected some valuable feedbacks for potential improvements and the overall application usability. The solution was regarded as particularly helpful, because it implements an advanced algorithmic pipeline with a temporally resolved hierarchical topic extraction, and an advanced visual tool so that the analyst can both follow the topic evolution over time and drill down towards point-wise facts. When asked about the most effective views, the streamgraph was considered the core of the system. Most of the user's activity was about exploring different temporal granularities (i.e. monthly, weekly, ...) so that the underlying corpus on which the topics are extracted represents different temporal horizons.

Tomographic exploration and automatic detection of anomalies were recognized to be helpful to keep an eye on unexpected trends and to select the most interesting and useful information. On this side, the suggestion was that future work should be devoted to the implementation of more sophisticated algorithms able to identify weaker signals and anomalies (e.g. triggering intermittent or periodic dynamics of topics).

The current prototype is fully actionable, but we acknowledge some limitations in terms of the underlying algorithmic pipeline, user interactions and visualization, and to platform-specific issues. In *TopTom* we face the common issues related to unsupervised learning analysis and in particular to the estimate of the more suitable number of topics to extract given a certain corpus. This can not be fixed once for all, but an adaptive mechanism informed by the semantic differences of the underlying documents might help in tuning such process, although extremely computationally expensive. Second, an additional layer of supervised pre-processing to categorize documents in human-defined categories to explore topics within narrower domains would probably enhance the overall performances. Similarly, great benefits would be expected from well trained spam filters and additional classifiers able to retain most informative documents dropping off-topics pieces or too-short-to-be-informative social media posts thus reducing the signal-to-noise ratio. Unfortunately, at the present stage the system does not allow a feedback loop from the analysts to the training phase of the algorithms from the visual interface.

The current visual interface is the first result built upon the described algorithmic pipeline. Through the visualization design process it was possible to identify new user needs, in particular the

ability of having an overview of the most relevant anomalies independently from time aggregation and topic aggregation that led us to design the *calendar view*, able to show in which days anomalies arose and future work will be focused on providing a global access to all the different levels of temporal aggregation. A second issue that emerged is that users would like to understand which topics have a stronger identity (being consistent across the hierarchy) when changing the depth and in relation with anomalies that can be solved by adding the possibility to compare the results of analysis of different hierarchical layers. Finally, the co-occurence of keywords (or entities) might be handled as a dynamical graph by defining a suitable framework for their interplay with the hierarchical topics.

From a more conceptual point of view, additional effort will be devoted to explore and validate different approaches to handle the temporal dimension in discussion-based media like Reddit. Although human activity on social media seems to follow similar statistical patterns [FCYJMT*15], the nature of the platforms (individual micro-blogging vs thematic subsections) tend to spark different discussion dynamics. While news and twitter posts have a clear generation timestamp and tend to have a limited attention over time [LGRC12], thematic discussions might cover longer and sustained debates. In such perspective, the concept of time for Reddit and forum posts should be handled differently from the bare timestamp of creation. The solution adopted in this paper is to flatten Reddit posts about the same discussion in a single document with the timestamp related to the first submission, however it might be important to devise a well-principled and robust approach to this issue.

7. Conclusion

TopTom has demonstrated a great potential in exploring and distilling information from heterogeneous textual data, extracting a temporally-resolved hierarchical topic structure and highlighting anomalous trends in topic volume. The solution demonstrated a great balance between a tailored visual analytics interface with effective visual metaphors organized in multiple coordinated views and an adaptable back-end where algorithmic improvements can be tested and easily deployed via APIs. Such highly flexible approach is adaptable to users needs and makes it a helpful tool both for academic researchers and analysts in the private sector.
The natural steps for further investigations follow several directions. From the algorithmic side, we will consider to test more sophisticated topic modeling approaches merging the temporal and the semantic dimensions and to explore more elaborated anomaly detection techniques to grasp weaker signals. From the user-interface and system architecture perspective, we will implement functionalities able to collect the analysts' feedback in a structured way to feed supervised machine learning algorithms. Moreover we will consider to supplement *TopTom* with additional views to integrate other content features. In general, *TopTom* will help researchers to systematically and easily investigate social media platforms identifying emerging trends about specific behaviors and targeted populations.

Acknowledgments

DB acknowledges support from the Lagrange Project and CRT Foundation (http://www.isi.it/en/lagrange-project/project). MM, BG and PC would like to thank Riccaldo Scalco for the support in the development of the front-end visualization.

References

[ACL*12] ALSAKRAN J., CHEN Y., LUO D., ZHAO Y., YANG J., DOU W., LIU S.: Real-time visualization of streaming text with force-based dynamic system. *IEEE Computer Graphics Application 32*, 1 (Jan 2012), 34–45. doi:10.1109/MCG.2011.91. 2

[BL06] BLEI D. M., LAFFERTY J. D.: Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 113–120. 2

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022. 1, 2

[BSH*10] BERNSTEIN M. S., SUH B., HONG L., CHEN J., KAIRAM S., CHI E. H.: Eddi: Interactive Topic-based Browsing of Social Status Streams. *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10* (2010), 303. doi:10.1145/1866029.1866077. 2

[BW08] BYRON L., WATTENBERG M.: Stacked graphs - Geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1245–1252. doi:10.1109/TVCG.2008.166. 2

[CLT*11] CUI W., LIU S., TAN LI SHI C., SONG Y., GAO Z., QU H., TONG X.: TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (dec 2011), 2412–2421. 2, 3

[CLWW14] CUI W., LIU S., WU Z., WEI H.: How Hierarchical Topics Evolve in Large Text Corpora. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (dec 2014), 2281–2290. doi:10.1109/TVCG.2014.2346433. 3

[CMZ10] CULNAN M. J., MCHUGH P. J., ZUBILLAGA J. I.: How large us companies can use twitter and other social media to gain business value. *MIS Quarterly Executive 9*, 4 (2010). 1

[DSC*15] DART R. C., SURRATT H. L., CICERO T. J., PARRINO M. W., SEVERTSON S. G., BUCHER-BARTELSON B., GREEN J. L.: Trends in opioid analgesic abuse and mortality in the united states. *New England Journal of Medicine 372*, 3 (2015), 241–248. 9

[DWS*12] DOU W., WANG X., SKAU D., RIBARSKY W., ZHOU M. X.: Leadline: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 93–102. 1, 2

[DYW*13] DOU W., YU L., WANG X., MA Z., RIBARSKY W.: HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2002–2011. doi:10.1109/TVCG.2013.162. 2, 3

[FCYJMT*15] FERRAZ COSTA A., YAMAGUCHI Y., JUCI MACHADO TRAINA A., TRAINA JR C., FALOUTSOS C.: Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 269–278. 1, 11

[FR11] FISCHER E., REUBER A. R.: Social interaction via new social media:(how) can interactions on twitter affect effectual thinking and behavior? *Journal of business venturing 26*, 1 (2011), 1–18. 1

[FVD*16] FERRARA E., VAROL O., DAVIS C., MENCZER F., FLAMMINI A.: The rise of social bots. *Commun. ACM 59*, 7 (June 2016), 96–104. doi:10.1145/2818717. 3

[GJTB04] GRIFFITHS T. L., JORDAN M. I., TENENBAUM J. B., BLEI D. M.: Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems* (2004), pp. 17–24. 2

[GTLY12] GUAN N., TAO D., LUO Z., YUAN B.: Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems 23*, 7 (2012), 1087–1099. 5

[Haw09] HAWN C.: Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs 28*, 2 (2009), 361–368. 1, 9

[HCLB10] HE Q., CHANG K., LIM E.-P., BANERJEE A.: Keep it simple with time: A reexamination of probabilistic topic detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 10 (2010), 1795–1808. 2

[HHKE16] HEIMERL F., HAN Q., KOCH S., ERTL T.: Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization & Computer Graphics*, 1 (2016), 190–199. 2

[HHN99] HAVRE S., HETZLER E., NOWELL L.: ThemeRiver: In Search of Trends, Patterns, and Relationships. *InfoVis 99* (1999), 4. 6

[Hoy04] HOYER P.: Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research 5* (2004), 1457–1469. 4

[Kaw11] KAWAMAE N.: Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM international conference on Web search and data mining* (2011), ACM, pp. 317–326. 2

[KLPM10] KWAK H., LEE C., PARK H., MOON S.: What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 591–600. 10

[LBK09] LESKOVEC J., BACKSTROM L., KLEINBERG J.: Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 497–506. 1, 5

[LG10] LERMAN K., GHOSH R.: Information contagion: An empirical study of the spread of news on digg and twitter social networks. *Icwsm 10* (2010), 90–97. 1

[LGRC12] LEHMANN J., GONÇALVES B., RAMASCO J. J., CATTUTO C.: Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 251–260. 1, 5, 11

[Lin07] LIN C.: Projected gradient methods for nonnegative matrix factorization. *Neural computation 19*, 10 (2007), 2756–2779. 4

[LMS*17] LEUNG P. T., MACDONALD E. M., STANBROOK M. B., DHALLA I. A., JUURLINK D. N.: A 1980 letter on the risk of opioid addiction. *New England Journal of Medicine 376*, 22 (2017), 2194–2195. 9

[LS99] LEE D. D., SEUNG H. S.: Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (1999), 788. 2

[LWC*14] LIU S., WANG X., CHEN J., ZHU J., GUO B.: TopicPanorama : a Full Picture of Relevant Topics. *IEEE Symposium on Visual Analytics Science and Technology* (2014), 183–192. 3

[LYK*12] LUO D., YANG J., KRSTAJIC M., RIBARSKY W., KEIM D.: Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics 18*, 1 (Jan 2012), 93–105. doi:10.1109/TVCG.2010.225. 2

[LYW*16] LIU S., YIN J., WANG X., CUI W., CAO K., PEI J.: Online visual analytics of text streams. *IEEE transactions on visualization and computer graphics 22*, 11 (2016), 2451–2466. 2

[MBK*14] MARS S. G., BOURGOIS P., KARANDINOS G., MONTERO F., CICCARONE D.: "every never i ever said came true": Transitions from opioid pills to heroin injecting. *International Journal of Drug Policy 25*, 2 (2014), 257–266. 9

[MHP*12] MATTERN F., HUHN W., PERREY J., DÖRNER K., LORENZ J. T., SPILLECKE D.: Turning buzz into gold. *McKinsey & Company white paper* (2012). 1

[MPCC13] MAURI M., PINI A., CIMINIERI D., CIUCCARELLI P.: Weaving data, slicing views. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI on - CHItaly '13* (New York, New York, USA, sep 2013), ACM Press, pp. 1–8. doi:10.1145/2499149.2499159. 2

[MT04] MIHALCEA R., TARAU P.: Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (2004). 3

[NDLU07] NALLAPATI R. M., DITMORE S., LAFFERTY J. D., UNG K.: Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), ACM, pp. 520–529. 2

[PGQC14] PANISSON A., GAUVIN L., QUAGGIOTTO M., CATTUTO C.: Mining concurrent topical activity in microblog streams. *Proceedings of the 4th workshop on 'Making Sense of Microposts' (WWW'14)* (2014). 2, 3

[RB10] ROSVALL M., BERGSTROM C. T.: Mapping change in large networks. *PLoS ONE 5*, 1 (2010). doi:10.1371/journal.pone.0008694. 2

[RL14] RODRIGUEZ A., LAIO A.: Clustering by fast search and find of density peaks. *Science 344*, 6191 (2014), 1492–1496. 5

[Rog17] ROGERS R.: Digital Methods for Cross-platform Analysis. In *The SAGE Handbook of Social Media*, Burgess J. J. E., Marwick A. E., Poell T., (Eds.). SAGE Publications Ltd, London, 2017, ch. 5, pp. 91–108. doi:10.4135/9781473984066.n6. 1

[Rud16] RUDD R. A.: Increases in drug and opioid-involved overdose deaths - united states, 2010–2015. *MMWR. Morbidity and mortality weekly report 65* (2016). 9

[SBB*12] SALATHE M., BENGTSSON L., BODNAR T. J., BREWER D. D., BROWNSTEIN J. S., BUCKEE C., CAMPBELL E. M., CATTUTO C., KHANDELWAL S., MABRY P. L., ET AL.: Digital epidemiology. *PLoS computational biology 8*, 7 (2012), e1002616. 9

[SM58] SOKAL R. R., MICHENER C. D.: A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin 28* (1958), 1409–1438. 4

[SS12] SAHA A., SINDHWANI V.: Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), ACM, pp. 693–702. 2

[SWL*14] SUN G., WU Y., LIU S., PENG T. Q., ZHU J. J., LIANG R.: EvoRiver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1753–1762. doi:10.1109/TVCG.2014.2346919. 2

[VFD*17] VAROL O., FERRARA E., DAVIS C. A., MENCZER F., FLAMMINI A.: Online human-bot interactions: Detection, estimation, and characterization. *CoRR abs/1703.03107* (2017). arXiv:1703.03107. 3

[WAB12] WANG Y., AGICHTEIN E., BENZI M.: Tm-lda: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), ACM, pp. 123–131. 2

[WGK14] WARD M., GRINSTEIN G., KEIM D.: *Interactive Visualizations. Foudations, techniques and applications*, second ed. CRC Press, 2014. 6

[WLS*10] WEI F., LIU S., SONG Y., PAN S., ZHOU X. M., QIAN W., SHI L., TAN L., ZHANG Q.: TIARA: a visual exploratory text analytic system. *Kdd '10* (2010), 153–162. doi:10.1145/1835804.1835827. 2, 3

[WM06] WANG X., MCCALLUM A.: Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th*

ACM SIGKDD international conference on Knowledge discovery and data mining (2006), ACM, pp. 424–433. 2

[ZCW*14] ZHAO J., CAO N., WEN Z., SONG Y., LIN Y. R., COLLINS C.: #fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1773–1782. doi:10.1109/TVCG.2014.2346922. 2, 3