A

ATR-Vis: Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter

RAHELEH MAKKI, Dalhousie University, Canada
EDER CARVALHO, Universidade de São Paulo, Brazil
AXEL J. SOTO, Dalhousie University, Canada
STEPHEN BROOKS, Dalhousie University, Canada
MARIA CRISTINA FERREIRA DE OLIVEIRA, Universidade de São Paulo, Brazil
EVANGELOS MILIOS, Dalhousie University, Canada
ROSANE MINGHIM, Universidade de São Paulo, Brazil

The worldwide adoption of Twitter turned it into one of the most popular platforms for content analysis as it serves as a gauge of the public's feeling and opinion on a variety of topics. This is particularly true of political discussions and lawmakers actions and initiatives. Yet, one common but unrealistic assumption is that the data of interest for analysis is readily available in a comprehensive and accurate form. Data need to be retrieved, but due to the brevity and noisy nature of Twitter content, it is difficult to formulate user queries that match relevant posts that use different terminology without introducing a considerable volume of unwanted content. This problem is aggravated when the analysis must contemplate multiple and related topics of interest, for which comments are being concurrently posted. This paper presents ATR-Vis, a user-driven visual approach for the retrieval of Twitter content applicable to this scenario. The method proposes a set of active retrieval strategies to involve an analyst in such a way that a major improvement in retrieval coverage and precision is attained with minimal user effort. ATR-Vis enables non-technical users to benefit from the aforementioned active learning strategies by providing visual aids to facilitate the requested supervision. This supports the exploration of the space of potentially relevant tweets, and affords a better understanding of the retrieval results. We evaluate our approach in scenarios in which the task is to retrieve tweets related to multiple parliamentary debates within a specific time span. We collected two Twitter data sets, one associated with debates in the Canadian House of Commons during a particular week in May 2014, and another associated with debates in the Brazilian Federal Senate during a selected week in May 2015. The two use cases illustrate the effectiveness of ATR-Vis for the retrieval of relevant tweets, while quantitative results show that our approach achieves high retrieval quality with a modest amount of supervision. Finally, we evaluated our tool with three external users who perform searching in social media as part of their professional work.

CCS Concepts: • **Information systems** → **Users and interactive retrieval;** • **Human-centered computing** → **Visualization;**

Additional Key Words and Phrases: Information Retrieval, Visual Analytics, Active Learning

ACM Reference Format:
Makki et al, 2017 (In Press). ATR-Vis: Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter. *ACM Trans. Knowl. Discov. Data.* 11, 4, Article A (July 2017), 33 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

This work was carried out with the aid of grant 2013-LACREG-07 from the International Development Research Centre, Ottawa, Canada, a CALDO-FAPESP grant (Proc. 2013/50380-0), and an ELAP scholarship. Brazilian researchers are also supported by grants from CNPq (205291/2014-7 and 305696/2013-0) and FAPESP (2011/22749-8), and researchers based in Canada by grants from NSERC.
Author's addresses: R. Makki, A.J. Soto, S. Brooks and E. Milios, Faculty of Computer Science, Dalhousie University; E. Carvalho, M.C.F. de Oliveira and R. Minghim, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Twitter is one of the most popular microblogging services around the world, and as in most social communication media, politics is a frequent conversation topic. Politicians, party organizers, the press media and the general public use social media to express opinions, compete for public attention and recruit new supporters [Aharony 2012; Gruzd and Roy 2014; Xu et al. 2013]. As a result, there is a genuine interest in mining this diversely rich and endless source of public opinions. In the latest elections in the United States and Canada, for instance, Twitter was extensively used during live media events such as presidential debates [Shamma et al. 2010], and as a platform for expressing opinions about presidential candidates [Contractor et al. 2015].

In national legislatures, government laws are introduced, debated and voted by the members of the parliament or congress. Recently established open government data policies render this information readily available. At the same time Twitter has become widely adopted by the public as a platform for engaging in discussions about the various bills under debate in the parliament. While several computational tools for the textual analysis of social media data have been proposed, tools for retrieving such data effectively are still lacking [Liu et al. 2016]. This work addresses this gap. Our specific goal is to retrieve Twitter posts related to political debates held in the parliament, and associate them to the specific debate they refer to. Such association is important because it allows identifying public opinion about political decisions and bills under debate. Yet, it must be accurate, so that retrieved content is relevant, and comprehensive to ensure that diversely expressed opinions are captured. Politicians, political analysts and journalists can benefit from this association as they may want to follow specific debates/topics and need to distinguish what people are saying about their debates of interest.

A well-recognized challenge when working with tweets is their textual content being short and riddled with acronyms, slang, incorrect spelling and grammar [Rosa et al. 2011]. Adding to this, our research problem poses its own specific challenges. One factor that renders it particularly difficult is the semantic relatedness among debates. Besides all target tweets having "federal politics" as their underlying topic, different debates can have other topics in common. For instance, during the second week of May 2014 debates on the issues of "Kidnapping of Girls in Nigeria" and of "Aboriginal Affairs" were taking place in Canada. Although they are distinct, both deal with similar issues of violence against women, which makes the association task especially challenging. Most contributions reported so far on tweet classification or retrieval consider well-differentiated categories such as sports, politics, economics and technology [Lee et al. 2011; Sriram et al. 2010], which constitutes a scenario simpler than the one considered here.

An additional challenge for our problem is that the distribution of tweets over debates is severely imbalanced. Some debates may raise issues that are deemed important to a larger audience and consequently they will originate more tweets than other debates related to more specific questions. Moreover, the relevant tweets to any debate are a small fraction of the overall volume of data. Another issue is the dynamic nature of the problem, as new debates are continuously introduced whenever a legislative session is held. It would not be suitable to train a model on old data, but it would also be infeasible to label large quantities of data every time new debates start to be monitored.

Considering that no labeled data is available to train a model, our approach relies on two aspects: (1) inspired by the idea of query expansion (pseudo relevance feedback [Xu and Croft 1996]) we iteratively refine the queries by which tweets are retrieved to maximally improve recall and precision and (2) we involve the user in the retrieval process

so that her domain knowledge can be leveraged. Therefore, we introduce an interactive and exploratory tool, named ATR-Vis (Active Tweet Retrieval Visualization), to facilitate the association process using multiple active retrieval strategies. The tool visualizes the tweets that have already been associated with each debate, and also those deemed as important for the user to inspect and assign. Yet, the user is also offered other visual means to explore the data, so that s/he can associate tweets based on this exploration and beyond what it is being "suggested" by the system. While it relies on user involvement to improve the association of tweets, ATR-Vis aims at keeping this involvement to a minimum. As opposed to other active retrieval approaches, the set of strategies introduced here exploit particular features of Twitter to select the most appropriate labeling requests. The main contributions of this paper include:[1]

— The proposal of a set of active retrieval strategies that are specific to Twitter and increase the retrieval accuracy in terms of precision and recall while minimizing user effort, i.e. the number of labeling requests.
— A comparison of the proposed approach with a state-of-the-art active retrieval method and its evaluation on different datasets.
— The presentation of an interactive framework, ATR-Vis, that enables non-technical users to employ the aforementioned active learning strategies while exploring the space of potential tweets and gaining a better understanding of the results.

The paper is organized as follows. After introducing a formal definition for our problem, Section 2 describes related work, which includes active learning approaches and exploratory tools for Twitter. Our user-driven visual framework for active retrieval is described in Section 3. Section 4 describes our results, covering quantitative experiments, use cases illustrating the applicability of our proposed framework and a qualitative analysis conducted with potential end users. The section closes with a critical discussion of the results presented. Finally, concluding remarks are stated in Section 5.

1.1. Problem Definition

DEFINITION 1. *Given a set of tweets* $T = \{t_1, t_2, ..., t_n\}$ *and a set of debates* $D = \{d_1, d_2, ..., d_m\}$*, the task is to retrieve, for each debate* $d_i \in D$*, a set of tweets* $T_i \subseteq T$ *so that tweets in* T_i *are relevant to the debate* d_i*, and* $T_i \cap T_j = \varnothing$ *for* $i \neq j$*. Equivalently, we define for any tweet* $t \in T$ *a function* $f(t_k) = d_i$ *if the retrieval method associates* t_k *with* d_i *or* $f(t_k) = \varnothing$ *if* t_k *is not retrieved, and therefore considered as "non-relevant" to any of the debates.*

Consistent with a typical information retrieval setting, we assume labeled data is not available for learning. However, we extend this setting assuming an external information source may be accessed (a domain expert or oracle) that can manually label a given tweet with any or none of the debates in D. We also assume that there is a practical cost of accessing the information source, so it is important to minimize the number of instances to be presented for feedback. Note from the definition that we presume each tweet is related to at most one debate.[2]

[1]From the contributions listed, the first one has been introduced in a previous work [Makki et al. 2015], and the last two are being introduced in this paper.
[2]This can be also stated as a classification problem, where we try to assign each tweet to a debate (or assign it to an additional "non-relevant" class when it is not retrieved). However, since initially there is no labeled data and queries are inferred from the debates, the problem is more naturally posed as an information retrieval problem.

2. RELATED WORK

Given the importance of microblog content and the multiple challenges described in the previous section, several research efforts have addressed Twitter content retrieval, classification and/or analysis. Initial attempts relied on manual identification of relevant keywords that are used to filter relevant posts either as part of the Twitter API parameters [Borge-Holthoefer et al. 2011; Gaffney 2010] or in a postprocessing step [Conover et al. 2011; Romero et al. 2011]. However, due to the noisy and evolving nature of tweet terminology, it is hard to ensure a proper recall, i.e. capturing all relevant topics without biasing the results towards certain specific keywords [Filho et al. 2015]. Other approaches considered query expansion to enrich the query terms and overcome the vocabulary mismatch problem [Massoudi et al. 2011; Gurini and Gasparetti 2012]. More recent methods resorted to external knowledge sources such as Wikipedia, Freebase, and Wordnet to obtain additional related terms to expand the query [Qiang et al. 2015; Lucia and Ferrari 2014]. However, expanding queries can undermine the precision of the retrieval as more generic terms are included [Miyanishi et al. 2013]. Our query expansion approach aims at adding features while preventing a loss in retrieval precision. Similar to our work, other studies highlighted the advantages of considering the structural information surrounding a tweet such as the hashtags, URLs and replies to other posts [Luo et al. 2012a; Luo et al. 2012b].

The labeling of Twitter data is an expensive process with its usefulness and generality limited to a certain thematic context and time window. Therefore, several studies have focused on learning models with limited labeled data. Semi-supervised techniques rely on large amounts of available unlabeled data along with a small amount of labeled data. For instance, a semi-supervised Bayesian network model for Twitter topical classification was proposed in [Chen et al. 2012], and a semi-supervised SVM-rank for scoring tweets based on their credibility is described in [Gupta et al. 2014].

Similarly, active learning is a special case of semi-supervision where the method itself requests instances to be labeled from an information source. The use of active learning techniques for analyzing microblog messages is a relatively new research topic. Hu et al. [2013] showed the importance of considering social relations and user similarity among microblog posters in selecting the instances to be labeled for the task of topical classification. The AIDR (Artificial Intelligence for Disaster Response) system [Imran et al. 2014] selects tweets to be labeled through crowdsourcing to identify informative tweets for disaster management. Peetz et al. [Peetz et al. 2013] trained a Naïve Bayes classifier and selected the most uncertain samples to be labeled to improve the performance of named-entity disambiguation in Twitter. The methods adopted in the latter two works consider only the tweet's textual content to train their classifier and to select the instances for labeling requests. They do not explicitly model other specific Twitter features, e.g. reply-related information. Moreover, none of the previous approaches considered this task within an interactive exploratory setting, where a human user can participate in the labeling task.

Active retrieval would be the analog of active learning for the task of information retrieval [Jaakkola and Siegelmann 2001]. Similar to active learning, the retrieval system is allowed to request instances to be labeled on an interactive basis for the sake of improving retrieval precision or recall. Again, it is assumed that there is a cost associated with each labeling request, so the number of requests should be minimized. Two major differences are that active retrieval is expected to face a *cold start*, i.e. starting with no labeled data, and some sort of query is available representing the information need. ReQ-ReC [Li et al. 2014] can be arguably considered as one of the state-of-the art active retrieval systems. The method consists of a sequence of two cycles. The inner cycle aims to improve the precision of the retrieval by training an SVM classifier, where

uncertain instances are selected for user labeling. The outer cycle aims to improve the recall by automatically formulating a new query that is obtained from low-ranked but relevant documents in the hope of increasing the diversity of the retrieval. While ReQ-ReC does not provide any exploratory features, we adopt it as a basis for comparing the back-end of our approach.

We note that these research efforts are not exempt of controversy on the possible limitations of this type of analysis, or around the ethical and social implications on working with people's generated content who may be unaware of how their data is being used and what the limitations of these analysis are [Boyd and Crawford 2012; Bertot et al. 2010]. While these are important questions to be addressed by governments and policy makers, the provision of computational tools as the one described in this paper - specially those targeting the accurate retrieval of content - is a relevant contribution to favor a better understanding of how such data can be leveraged and what their possible implications are.

2.1. Visual Analytics for Microblog Content

A fair amount of work has been devoted to the analysis and exploration of text using visual analytics [Alencar et al. 2012]. The underlying idea behind visual analytics is to combine the benefits of data mining with the cognitive abilities and domain knowledge of a human user to perform a certain analytical task that cannot be performed automatically [Keim et al. 2010]. Interactive visualizations become the means by which the user observes and explores the data space and communicates her information needs [Thomas et al. 2006].

Twitter has been attracting considerable attention as a data platform for visual text analytics. Initially, the tools focused mainly on providing platforms for visualizing Twitter content, users and images beyond the conventional list layout [O'Connor et al. 2010; Dörk et al. 2010; Diakopoulos et al. 2010]. Most tools apply dynamic topic models and present them using ThemeRiver-inspired visualizations [Havre et al. 2000] to convey a summary of the conversations over time. A second generation of systems took the idea of extracting topics over time and applied more advanced text processing algorithms to improve content analysis. For instance, Leadline [Dou et al. 2012] aims at extracting the major events that led to changes in the topical themes and characterizing each event with information on who, what, when, and where. The analysis of evolving topics from the perspective of how they compete among each other and how they diffuse to different users were addressed in [Xu et al. 2013; Liu et al. 2014; Sun et al. 2014]. Yet other studies proposed methods for predicting revenue or stock prices from Web data. A visual analytics system to predict the box-office success of a movie was proposed by Lu et al. [2014], where the number of mentions in Twitter per day is one of the few variables of the model. Retrieval is restricted to the hashtag posted by the movie's official Twitter account. However, none of these previous works consider a systematic strategy for accurate retrieval of relevant tweets, rather they analyze all Twitter posts that match a given set of keywords.

Solutions such as SensePlace2 [MacEachren et al. 2011] and the system by Chae et al. [2012] focus on providing geo-visual analytics for understanding place, time, and theme components of evolving situations. Such solutions have been proved useful to improve situational awareness in monitoring catastrophes based on their reporting on Twitter. Scatterblogs2 [Bosch et al. 2013] is another visual analytics tool for situational awareness, but unlike the previous two it supports the expansion of an initial manual query by looking at the co-occurrence of tweets retrieved with the first query. These filtering aids are mostly based on textual content only, without taking Twitter-specific features into account.

Concerning the target goal of improving retrieval capability, the recent tool introduced by Liu et al. [2016] is probably the closest to ours. It also exploits the specific characteristics of microblog data to improve retrieval performance when looking for information in twitter posts. The authors propose an uncertainty-aware microblog retrieval model to quickly retrieve salient items, i.e. posts, users and hashtags, and provide an estimate of the uncertainty associated with the retrieval model. The retrieval relies on a previously introduced uncertainty-based mutual reinforcement graph model, in which the quality of a post content, user social influence and hashtag popularity mutually reinforce each other in order to determine the relevance of a post to a query. A composite visualization using a graph metaphor is proposed to support analysts to understand the retrieved data and interactively refine the retrieval model.

Both our work and Liu et al.'s aim at providing visual means for improving and facilitating retrieval. However, Liu et al. focus on the authorativeness and popularity of the posts, while we focus on their thematic relevance and the ability to identify as many relevant posts as possible. Although Liu et al. acknowledge the difficulties associated with evaluating recall, its estimation is highly important for our motivation. In addition, we consider the simultaneous retrieval of multiple, unbalanced and closely-related topics, which makes it challenging to achieve a high precision retrieval. We address these challenges by proposing active retrieval strategies, where a user provides feedback on request by the system—in addition to the feedback from her own exploration—to improve retrieval results.

3. ATR-VIS DESCRIPTION

The main goal of our proposed approach is, given a set of target debates, to retrieve tweets relevant to each debate, attempting to maximize precision and recall. Our framework for handling the problem has three major components. The first component is responsible for the initial unsupervised retrieval, trying to achieve the best possible response without any human intervention. Clearly, the better this component performs, the less load is put on a user in subsequent steps. Retrieved tweets provide a pseudo-relevance feedback [Xu and Croft 1996] that is used to improve the set of discriminative features.

The second component encompasses a set of strategies introduced to involve the user in those critical cases where her involvement is likely to yield an increase in the retrieval precision or recall. This component is tightly connected with the third one, i.e. the interactive visualizations. These components present selected instances for manual labeling and enable user exploration over the collection of tweets and their characterizing features. The results of user interaction are fed back into the retrieval engine, which analyzes the given information to automatically extract new discriminative features that will guide further iterations of the retrieval process. The whole process is illustrated in Figure 1, while Table I summarizes the notation used throughout this paper.

3.1. Unsupervised Tweet Retrieval

In order to formulate our queries we follow an approach similar to that introduced by Golestan Far et al. [2015], which generates the query by extracting discriminative terms from a document representative of the information need. A representative document for each debate is produced by concatenating all transcripts of a single debate, which can span multiple parliament sessions and days. From each document a set of discriminative keyterms for the debate is extracted, based on their tf-idf values. The keyterms identified for each debate define the query for retrieving the relevant tweets from T.

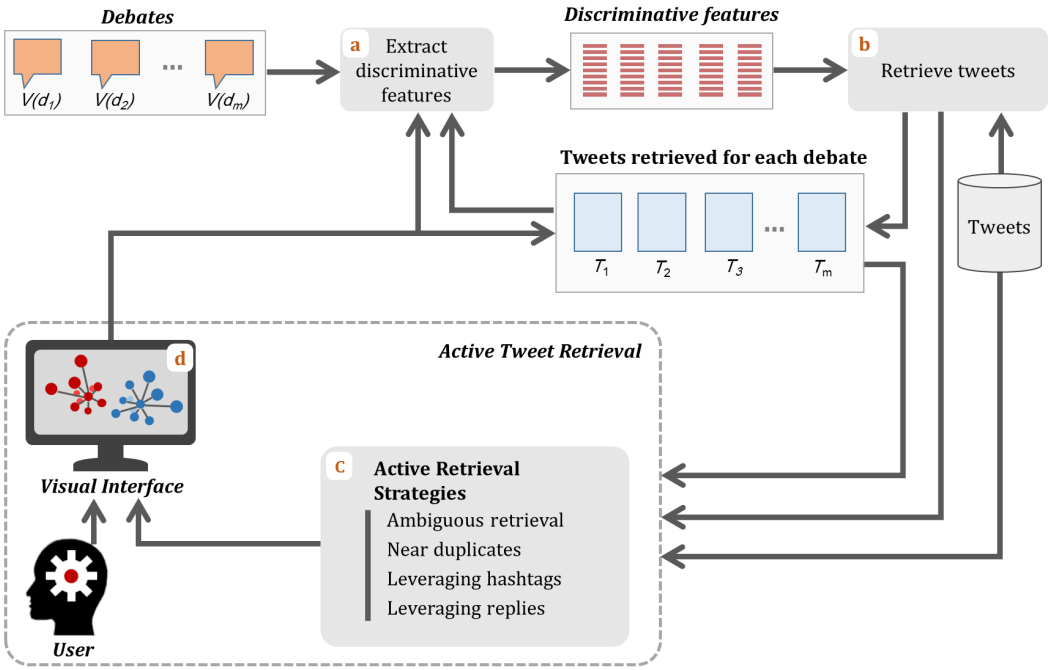

Fig. 1. The proposed framework for retrieving tweets relevant to a set of political debates. The unsupervised retrieval consists of extracting discriminative features (a) and retrieving tweets (b), and the active retrieval component selects the labeling requests (c) and updates the retrieval model based on the obtained labels (d).

Table I. Notation used in this paper.

Notation	Description
T	Set of tweets in our dataset.
D	Set of relevant debates.
$\mathbf{k}$	List of tweet features.
Ω	Matrix of weights $\omega_{i,j}$ representing the importance of feature k_i to debate d_j.
H	Set of all hashtags occurring in T.
$f(t_i)$	Returns the debate associated with tweet t_i by the retrieval method.
T_j	Set of tweets T_j associated with debate d_j by the retrieval method.
$F(T')$	Returns the frequency distribution of debates associated with a set of tweets T' by the retrieval method.
$T(h_o)$	Set of tweets having hashtag h_o.
$R(t_i)$	Set of tweets in the reply chain of tweet t_i.
$df(h_o)$	Debate frequency of hashtag h_o.
$V(T')$	Virtual document constructed by concatenating the textual content of a set of tweets.

Using just the debate transcripts as a source for the queries is not enough, however, as there is a potential mismatch between the formal vocabulary used in the debates and the informal one adopted in Twitter. Therefore, inspired by the pseudo-relevance feedback approach [Xu and Croft 1996] we use the retrieved tweets to further expand the list of discriminative keyterms. In addition to regular terms, discriminative Twitter-specific features such as hashtags, user mentions and URLs that appear in the retrieved tweets are also extracted and added to the list of features. For example, for the Canadian debate about "Bill C–23, Fair Elections Act" the features "fraud", "#unfairelectionsact", "@PierrePoilievre" and the expanded URL of "http://fw.to/oO9okOb" are added, which represent respectively a common term, a popular hashtag against the bill, the politician who introduced the bill, and a link to a news article explaining the bill.

DEFINITION 2. *Let* $\mathbf{k} = (k_1, k_2, \ldots, k_s)$ *be the list of features, and let matrix* $\Omega = (\omega_{i,j}) \in \mathbb{R}^{s \times m}$ *indicate the importance of each of the* s *features in each of the* m *distinct debates.*

Following the above definition, a matrix component $\omega_{i,j}$ contains a non-zero value if feature i is selected as a discriminative feature of debate j. The discriminative power of our different types of features (terms, URLs, user mentions and hashtags) is not necessarily the same. Thus, the non-zero values of $\omega_{i,j}$ are initially set according to the feature type. The overall intuition is that some features (e.g. hashtags) are more reliable indicators of the debate than other types of features (e.g. user mentions), so we assign them different initial weights. The setting of these weights by feature type is further described in Section 4.2. When new features are extracted at later iterations, their initial weights are decayed proportionally with the number of iterations to avoid unstable behavior.

A debate d_j is thus represented as a feature vector $\mathbf{d_j} = (\omega_{1,j}, \omega_{2,j}, \ldots, \omega_{s,j})$, while for a tweet we define its feature vector as:

$$\mathbf{t_i} = (\alpha_{1,i}, \alpha_{2,i}, \ldots, \alpha_{s,i}), \alpha_{p,i} = \begin{cases} 1, k_p \in t_i \\ 0, k_p \notin t_i \end{cases}, p \in [1, \ldots, s]. \quad (1)$$

Similarity between a debate d_j and a tweet t_i is given by the dot product of their feature vectors. When retrieving tweets we calculate the similarity score for each tweet-debate pair, and the tweet is assigned to the debate for which the similarity score is highest provided this similarity is above a certain threshold. Details about the setting of this threshold are provided in Section 4.2.

3.2. Active Tweet Retrieval

We propose four strategies for improving retrieval accuracy with user involvement. The goal is to select instances to be labeled that are most helpful in improving retrieval results while minimizing the number of labeling requests. The feedback resulting from a labeling request is important not only for that particular request, but also from what can be learned from it.

3.2.1. Ambiguous retrieval. Tweets are retrieved to the debate that has the highest similarity to its query. However, multiple debates could have very similar highest scores. This scenario suggests a good opportunity for asking the user to clarify the ambiguity. The user feedback is useful not only to determine the correct debate for these similarly-scoring tweets, but also to modify the current list of automatically extracted discriminative features.

Let $sim(t_i, d_j) \approx sim(t_i, d_r)$ with d_j and d_r having the highest scores for t_i compared to other debates in D. Upon the presentation of t_i to the user, if she assigns t_i to d_j, we can find the specific features in $\mathbf{k}$ that contributed to $sim(t_i, d_r)$ and reduce their associated weights for debate d_r, i.e. reduce $\omega_{p,r}$ for any k_p in t_i. Similarly, an increase in the weights of the features of the winning debate is also applied. This reduction (or increase) is proportional to the overall number of tweets associated with d_r for which $k_p \neq 0$.

Any specific hashtag found in t_i represents a valuable piece of information (in Section 3.2.3 we discuss how specific hashtags are identified). Therefore, any other non-retrieved tweet that includes this hashtag is also retrieved to the same class.

3.2.2. Near-Duplicates. We observed that in our Twitter dataset a large number of tweets are near-duplicates of each other (and they are not retweets). It is safe to assume that tweets that are near-duplicates should be assigned to the same debate. Therefore, we identify clusters of near-duplicate tweets, where a tweet is added to a

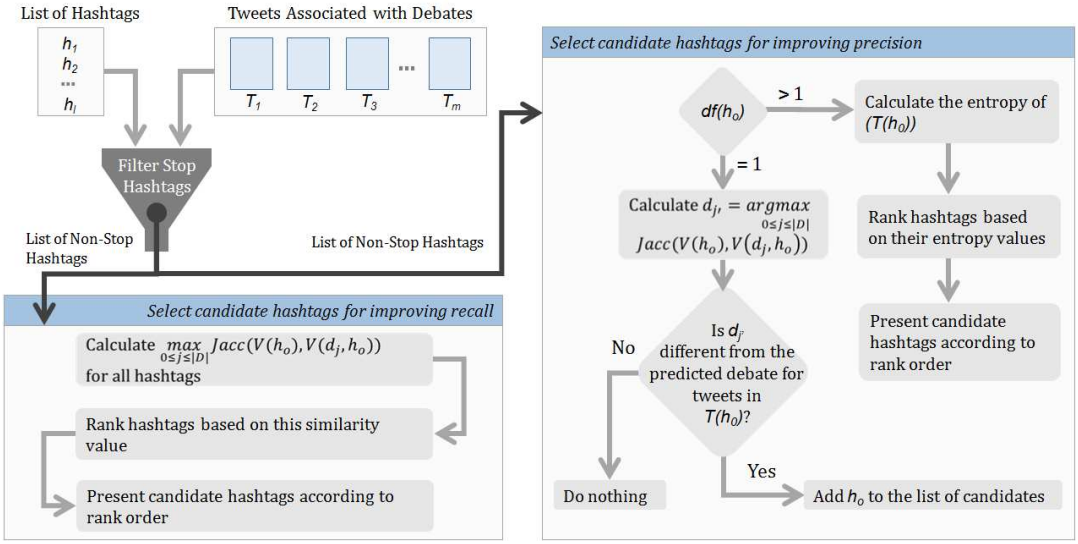

Fig. 2. Hashtag selection strategy to improve retrieval precision and recall.

cluster if it is a near-duplicate of all the tweets already in that cluster. Tweets from larger clusters have more potential as candidates for labeling requests, since labeling a single tweet from a cluster is sufficient to associate all its tweets with the same debate. Furthermore, the likelihood of the cluster belonging to a debate is an important criterion to avoid requesting the labeling of tweets in a non-relevant cluster. As a result, we rank clusters by their likelihood to belong to a debate and their cardinality, and use this rank to present to the user tweets from these near-duplicate clusters for labeling. We also take advantage of any specific hashtag identified in a labeled cluster, as described in the previous strategy.

Checking for near-duplicates naively has a quadratic time complexity, which given the typical scale of Twitter datasets becomes computationally prohibitive. Therefore, we apply Locality Sensitive Hashing [Slaney and Casey 2008], that allows near-duplicates to be found in linear time complexity. Intuitively, this method works by breaking down the text content in smaller pieces and applying a hash function to each piece. If multiple pieces of different documents are hashed to the same values, then there is a high probability that these documents are duplicates or near-duplicates. The specific approach adopted is similar to the one described by Soto et al. 2015, which consists in taking word tri-grams of the tweet content, using min-hashing to reduce the feature set, and hashing signature bands to identify near-duplicate content.

3.2.3. Leveraging Hashtags. Hashtags are possibly the most popular feature in Twitter. People use them to make their tweets easier to find and to engage in conversations with others. This third strategy relies on frequent specific hashtags to find tweets that were either retrieved to the wrong debates or failed to be retrieved. The sequence of steps for leveraging hashtag occurrences and user knowledge to improve retrieval results is illustrated in Figure 2 and described next.

Filtering Stop Hashtags. Candidate hashtags for improving retrieval accuracy are those likely to be good indicators of a specific topic in the problem domain. For instance, while "#cdnpoli" is a hashtag often associated with posts related to Canadian politics in general, it cannot be taken as a discriminative feature in the context of

our debate association problem, since it appears in tweets related to many different debates being held in the Canadian parliament. We refer to these hashtags as "*stop hashtags*"—similar to stop words in natural language processing, they do not add any useful information for the topical categorization of the data.

DEFINITION 3. *We define a virtual document as the concatenation in no specific order of the textual content of a set of tweets. We use the function $V(T')$ to indicate the virtual document obtained from the set of tweets T'.*

It is worth noting that stop hashtags are highly domain-specific. In our scenario, in order to identify a stop hashtag we must look at how many debates are likely to retrieve tweets, given that hashtag. The more debates a hashtag appears in, the more likely it should be considered as a stop hashtag. Therefore, we construct a virtual document for each debate by concatenating all its tweets as assigned by the retrieval algorithm, i.e. $V(T_j)\forall j$. Then, each hashtag h_o is ranked in ascending order based on its debate frequency $df(h_o)$, i.e. the number of virtual documents that include hashtag h_o. The result is a sorted list of hashtags, with those on the top being more likely to be specific.

Improving precision using hashtags. Our unsupervised method may retrieve content incorrectly, thus associating hashtags with the wrong debates, which in turn may lead to the retrieval of more irrelevant tweets. In order to use hashtags to improve retrieval precision we must first identify those non-stop hashtags that appear in tweets that have been incorrectly associated. Two scenarios are possible: either those tweets including a non-stop hashtag have been retrieved to more than one debate, or all of them have been associated with a unique debate. The first scenario is likely to be a conflict situation for the system, as a specific hashtag has been retrieved to multiple debates. The second scenario would indicate that either all retrieved tweets containing that hashtag are correctly associated, or they are all incorrectly associated.

DEFINITION 4. *Let $T(h_o)$ be the set of tweets that include a hashtag h_o and $F(T(h_o))$ the frequency distribution of debates associated with the tweets in $T(h_o)$ by our retrieval method. The normalized entropy for the distribution $F(T(h_o))$ is calculated as follows:*

$$\eta(F(T(h_o))) = \frac{\sum_{j=1}^{m} -F(T(h_o))(d_j)\, log(F(T(h_o))(d_j))}{log(|F(T(h_o))|)} \tag{2}$$

In the first scenario, h_o occurs in more than one debate, $df(h_o) > 1$. Assuming that h_o is not a stop hashtag, the value $df(h_o)$ is small. We first compute $F(T(h_o))$ and then its normalized entropy, $\eta(F(T(h_o)))$, using Equation 2. After computing this entropy value for all the non-stop hashtags, they are ranked in decreasing order. This normalized entropy allows to identify uniform-like distributions in $F(T(h_o))$, which could be an indicator of incorrect associations as we assumed h_o to not be a stop hashtag. Therefore, high-ranked hashtags are good candidates for involving the user to decide whether there is any issue with the retrieval of the tweets in $T(h_o)$ or not.

In the second scenario, the hashtag h_o occurs in a single debate, $df(h_o) = 1$. In this case, either all tweets in $T(h_o)$ have been correctly retrieved, or all of them have been assigned to the wrong debate. To determine which the true situation for the given hashtag h_o is, we first build a virtual document—referred to it as $V(T(h_o))$—by concatenating the textual content of all the tweets in $T(h_o)$. Then, we construct a virtual document for each debate by concatenating all the tweets previously associated with it except those in $T(h_o)$, i.e. $V(T_j - T(h_o))$. By calculating a similarity score between $V(T(h_o))$ and $V(T_j - T(h_o)), \forall j$, it is possible to identify the most similar debate to the

tweets in $T(h_o)$. If the highest-scoring debate for the tweets in $T(h_o)$ is different from the debate retrieved by our algorithm there is a good likelihood that these tweets have been mistakenly associated. In this case we involve the user to address this inconsistency and decide which debate they should be associated with.

The similarity score between two virtual documents is computed from a vector profile built for each of them containing a list of words with the highest tf-idf values. Then, we compute the Jaccard similarity using binary weights for their corresponding vector profiles, i.e. Jacc$(V(T(h_o)), V(T_j - T(h_o)))$.

Improving recall using hashtags. The unsupervised retrieval algorithm may fail to retrieve all relevant tweets mostly due to the vocabulary mismatch problem. To improve retrieval recall, we need to look for hashtags that have not been selected as discriminative features but are still "good indicators" of debates in D. A straightforward approach would be to select the most frequent non-stop hashtags occurring in the non-retrieved tweets (see Section 4.1 for a description of what is considered as our pool of non-retrieved tweets). However, these would not necessarily indicate similarity to any of the given debates. For instance, "#no2niki" is a frequent hashtag in tweets opposing a parliament member with reference to a motion on abortion. While it could be a discriminative feature to identify tweets about this topic, if "Abortion" is not among our selected debates, then this hashtag is not a good candidate for improving the recall.

To identify non-retrieved hashtags that are indicators of debates in D, we follow the same previous approach to calculate the similarity between the virtual document $V(h_o)$ of the tweets that include a given hashtag h_o and all tweets retrieved as relevant to each debate in D. Each hashtag is then associated with the debate with the highest similarity score, i.e. max Jacc$(V(h_o), V(d_j, h_o), \forall d_j \in D$. In this way, a list of hashtags ranked in decreasing order of their associated maximum similarity value to the debates is built. Given that they are likely to be relevant to some debate in D, we follow the ranking to present the candidate hashtags in this list to the user and ask for feedback on their relevance.

3.2.4. Leveraging Replies. As a social media platform, Twitter enables users to engage in conversations by replying to other users' posts. We consider this relational information between tweets as one of the selection strategies, following the hypothesis that replies to a tweet t_i are likely to be associated with the same debate as t_i.

We first trace back reply tweets to their *sources*, which are tweets that are not replies to any other tweet. We group together all the reply tweets that share the same *source*, including the *source* itself. The *reply chain* of a tweet t_i, i.e. $R(t_i)$, contains all tweets that have been grouped due to their reply relation.

Considering only the tweets in $R(t_i)$ that are already retrieved to debates in D, we calculate an entropy value using a similar approach to that described in Equation 2. If all these tweets are associated with the same debate, then the entropy value will be equal to zero, while if they are split uniformly among all debates, the entropy value will be equal to one. Reply chains with a high entropy signal some inconsistency between the conversation topic addressed by the tweets and their retrieval, and consequently it is more likely that some of these reply tweets were indeed retrieved to the wrong debates. Therefore, tweets in high entropy reply chains are good candidates for user involvement. Their cardinality is also considered in selecting the candidate reply chains. Thus, reply chains are sorted based on their entropy value multiplied by a value proportional to their cardinality, i.e. $\eta(F(R(t_i)) \times log(|R(t_i)|)$. These reply chains are presented to the user following this sorting. This strategy helps improve both precision and recall, as it allows to correct mistaken assignments and also to recover tweets that had not yet been retrieved.

It is not possible to calculate the entropy function if no tweets are retrieved in the reply chain of a source tweet t_i, i.e. $F(R(t_i)) = \varnothing$. This may be due to either a failure of the retrieval method, or these tweets are actually not related to any of the target debates. In order to determine whether these tweets are likely to be relevant, we select the largest reply chains for user inspection. In this case, any tweet labeled as relevant to one of the debates will contribute to improving retrieval recall. As in previous cases, any specific hashtag identified within a labeled tweet is leveraged as described in the first two strategies. Even if the user believes that the entire reply chain is not relevant to any debate, the method still benefits from the feedback by identifying specific hashtags and URLs that appear in these tweets and influencing the algorithm not to retrieve tweets containing these features in the future.

3.3. Interactive Visualizations

In order for the framework to be accessible and usable by non-technical persons, the retrieval process and its embedded strategies have been integrated into a visual interface that includes multiple complementary interactive visualizations. The resulting tool affords a user-driven analysis that fosters a better understanding of the retrieval strategies and enables the system to incorporate user domain knowledge in learning the assignment strategies. Therefore, the visual interface of ATR-Vis has been designed to meet the following goals:

(1) Provide an interface that incorporates and supports active retrieval strategies for an accurate and complete retrieval of tweets given a set of predefined debates. Such an interface should reduce user effort when handling labeling requests by presenting appropriate visual aids to support her task. To generate the desired impact, the tool should be understandable by non-data mining experts.
(2) Enable user-driven exploration of the retrieved and non-retrieved tweets and allow her to modify the retrieval model as a result from this exploration, beyond the handling of the labeling requests (e.g. by updating the debate-characterizing features).

ATR-Vis is a web-based application. The front end was built using D3.js [Bostock et al. 2011] and Bootstrap[3], while the backend was written in Java and uses Apache Lucene[4] for text indexing and searching. Design choices of ATR-Vis were made based on the analytical tasks and the type of data to be visualized, and by following expressiveness and effectiveness principles [Munzner 2014]. Multiple coordinated visualizations are employed due to their helpfulness for exploring intricate data with diverse attributes [North and Shneiderman 2000; Roberts 2007].

The visual interface consists of two main views: *Assignment* and *More*. The *Assignment* view enables the retrieval method to obtain feedback from labeling requests as well as multiple secondary views for aiding the user in the manual association process and the analysis of the retrieved tweets. The *More* view was designed to accommodate the two strategies that make use of the structural information of the tweets: *leveraging replies* and *hashtags*.

3.3.1. Assignment View. The Assignment View is shown in Figure 3. Each debate is uniquely associated with a color that is consistently preserved throughout the interface. The system presents the user a tweet labeling request selected either by the *Ambiguous retrieval* or the *Near-Duplicates* strategies, to be assigned to one of the predefined debates (panel a in the figure). To assist with the manual association of the

[3]http://getbootstrap.com
[4]https://lucene.apache.org

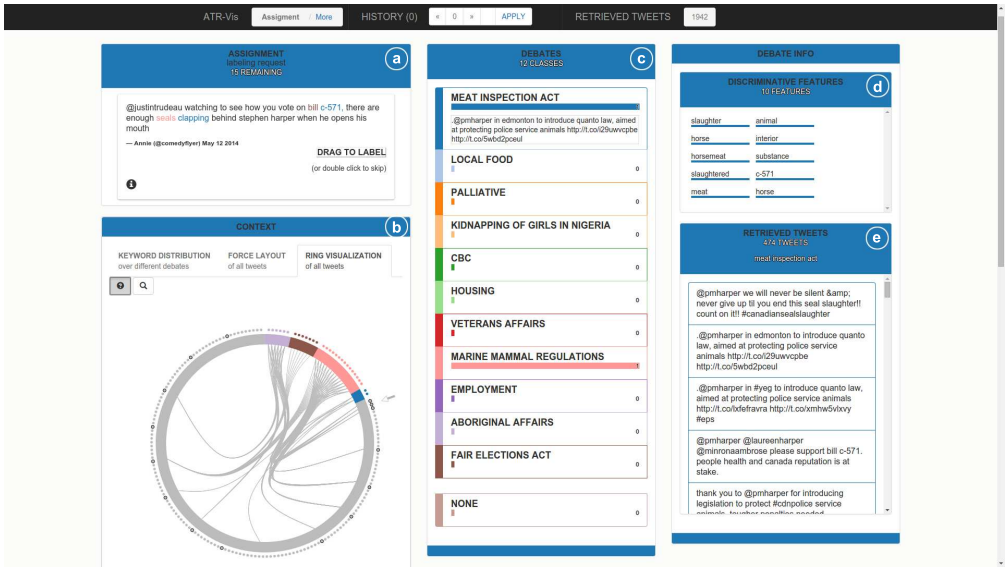

Fig. 3. Assignment View: a set of visual aids to facilitate tweet retrieval. (a) Labeling request for a Twitter post. (b) Visualization of the labeling requests in a broader context. (c) List of debates of interest. (d) Discriminative features for the selected debate. (e) Tweets retrieved by the selected debate.

current tweet, its non-stop words are shown with the color of the debate in which they occur more frequently.

Debates are presented as a vertical list (panel c), which follows the principle of grouping tweets assigned to the same debate in a same spatial region. Each debate shows its name and its similarity score with the current tweet in the labeling request. The similarity score is shown explicitly as a number, and also as a horizontal bar with length proportional to the score value, while always spanning to the panel width for the top-scoring debates. The reason for this design choice is that length is a preattentive attribute of visual perception and one of the most effective channels for encoding quantitative values [Ware 2012].

Whenever the user hovers the mouse over a debate, a sample tweet associated to that debate is shown. She can also click on a debate to access all its assigned tweets and discriminative features on a side panel. All background panels of the interface are colored according to the selected debate, except for the *Assignment* panel, which is colored with the debate that has the highest similarity score with its displayed tweet. The set of system-extracted discriminative features is shown as a sequence of terms (panel d), in which the order and length of the bar underneath are defined by the feature weight. If the user believes that a discriminative feature shown is not appropriate to this particular debate, she can modify it accordingly by dragging the feature to the proper debate.

The context panel (panel b) includes multiple visualizations to inform the labeling task. The *Keyword Distribution* tab allows observing the frequency distribution of any non-stop words occurring in the labeling request over the debates. Bar charts are a suitable choice, because they are good for analytical and accurate comparison of a value across multiple categories [Kirk 2012]. Two other visualizations place the labeling request in the global context of the currently retrieved and non-retrieved tweets. Given the potentially large number of tweets, they show only a sample, which includes tweets from the labeling requests and a stratified sample of those retrieved by each

debate. Since we aim at showing relationships between the tweets, both visualizations rely on a graph metaphor, where the nodes represent tweets. An edge connects a pair of nodes if their similarity score is above a user-selectable threshold for one common debate. The two visualizations differ in how nodes are laid out: by the forces exerted by the connections in the *Force Layout*, or arranged in a circle in the *Ring Visualization*.

The Force Layout is one of the most common solutions to show networked data [Munzner 2014]. As a result of the force-directed algorithms simulating the effect of spring-like physical forces acting on the edges and repulsive forces on the nodes, the resulting layout tends to show spatial clusters of similar, or highly related nodes. Graphs of this type are commonly employed to represent different Twitter visualizations from followers graph [Hussain et al. 2014] to retweet relationships between users [Morstatter et al. 2013] and hashtag networks [Jussila et al. 2013]. Our Ring Visualization is similar to a chord diagram with same-width chords and edge bundling to minimize cluttering and reveal high-level edge patterns [Holten 2006]. Its radial network layout is a good choice for showing relationships between different categories [Kirk 2012]. Chord diagrams have been used for visualizing different aspects of Twitter data elsewhere [Gabrielli et al. 2014; Prasetyo et al. 2016].

Furthermore, we followed the principle of attention management in ATR-Vis [Wang Baldonado et al. 2000], ensuring that a visual cue with the results of a user selection is provided [Munzner 2014]. For instance, when double-clicking on any tweet in any of the views, we immediately attract the user's attention to the assignment panel, which shows the tweet. Both the Force Layout and the Ring Visualization are coordinated with the assignment panel as the labeling request is highlighted in the graph.

Other possible interactions in this view include: labeling the current request, flipping through the list of requests, selecting new posts for inspection, visualizing all user requests, getting the closest neighbors (most similar posts) to a post, and resampling the nodes currently visualized. There is also a bar at the top of the screen for keeping track of the sequence of actions, which also serves the purpose of visualizing the impact of each change.

3.3.2. More View. This view, presented in Figure 4, allows interaction with two of the active retrieval strategies introduced to leverage the reply chains and the hashtags co-occurrence. The left-most panel (panel f) shows the reply-based conversation of a source tweet displayed as a tree layout. Hierarchical tree networks are good to represent the structure of conversations [Kirk 2012; Cogan et al. 2012; Pascual-Cid and Kaltenbrunner 2009], where the approximate number of replies and the number of different branches can be assessed at a glance. The specific reply chain is selected according to the number of tweets involved and its potential to reveal conflicting retrievals, as explained in Section 3.2.4. The user can explore the posts in the conversation, aided by the color indicating to which debate each one was retrieved, and decide whether an individual post or a whole subsequence of replies should be labeled. This view is especially useful in situations where there is topic drift.

The second view shows hashtags deemed as relevant for the user to inspect and provide feedback regarding their specificity to any of the given debates. The information is presented as a bipartite graph where one node depicts a selected hashtag and the other nodes depict the debates (panel g). Since linewidth is also a preattentive attribute of visual perception [Ware 2012], edge width is used to indicate the similarity value that measures the relatedness of the hashtag to a particular debate as explained in Section 3.2.3. Usage distribution of the target hashtag among the retrieved tweets for each debate is also shown. This is complemented with a panel on the right (panel h) that lists all the tweets containing the focus hashtag and retrieved to a specific debate. Similarly

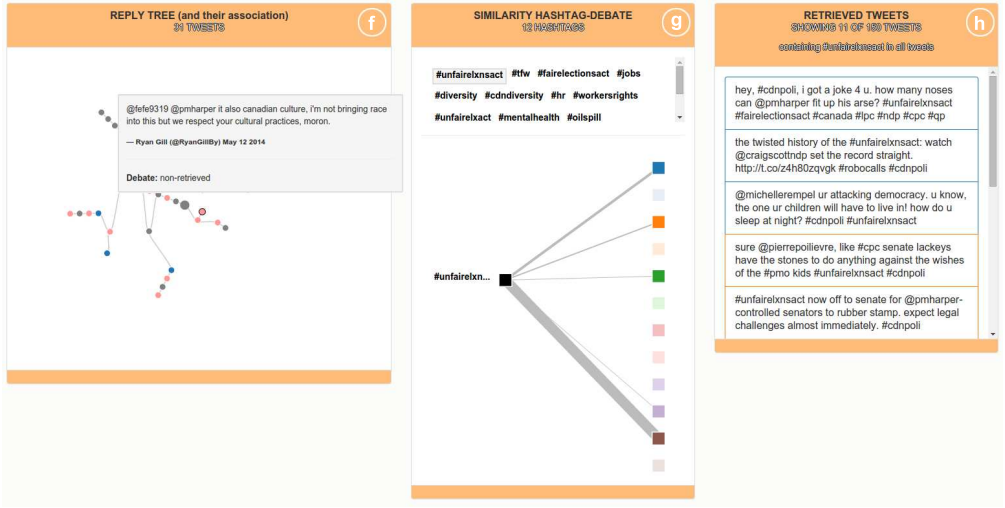

Fig. 4. More View. (f) Exploring a conversation thread on Marine Mammal Regulations. (g) Exploring how the hashtag *#unfairelxnsact* is associated with different debates. (h) Enumeration of tweets containing a hashtag.

to the previous strategy, the user can label individual posts or all posts containing that hashtag.

4. RESULTS

In this section we present results after evaluating ATR-Vis from multiple perspectives. We first present the datasets employed in these studies, followed by the parameter settings used in the strategies, and quantitative results of our retrieval strategies as compared with alternative retrieval strategies by means of a simulated user or oracle. We also describe two use cases that aim at describing typical scenarios of exploratory analysis with our framework. Finally, a user-oriented evaluation is described where three target users have interacted with the system in a pair analytics session [Arias-Hernandez et al. 2011]. Another use case illustrating the applicability of our tool in a different scenario than the parliamentary datasets considered so far is shown in Appendix A.

4.1. Datasets

Experiments were conducted on two parliamentary datasets. The first one refers to the Canadian House of Commons during the period 12-16th May 2014. We selected the 11 debates that received most attention in the parliament during that week (measured in terms of their overall length of discussion), since these were more likely to generate an expressive number of opinions in social media. The second dataset refers to 5 mainstream debates being held in the Brazilian Federal Senate from 25th to 29th May 2015. The title of these debates are presented in Tables III and V. Transcripts of the selected debates were extracted from the respective parliament websites.[5, 6]

We used Twitter's streaming API to collect tweets during the weeks of interest. Since it returns a minor fraction of the total volume of tweets at any given moment (roughly 1%), we must, to the maximum extent, restrict our search to Canadian (or Brazilian)

[5]http://www.parl.gc.ca/Default.aspx?Language=E
[6]http://dadosabertos.senado.gov.br/

political tweets in order to gather as many relevant tweets as possible without introducing many spurious ones. Furthermore, the collection procedure should not be biased towards any specific keywords, or the resulting datasets would depend on the choice of keywords and will likely report these keywords as being relevant. Bearing this in mind and in agreement with recommendations by Miranda Filho et al. [Filho et al. 2015], we used as initial information the parliament members' Twitter accounts. Twitter user names for the Canadian Parliament members were obtained from the Politwitter website,[7] while for the Brazilian case 80 Twitter user names, out of 81 senators, were identified manually. We collected tweets that were either posted by a parliament member, replied to a post by any of them, or that included one of their Twitter user names in its text. This resulted in datasets containing 16,297 and 9,625 original tweets (no retweets) for the Canadian and the Brazilian data, respectively. We chose to ignore retweets since they would be automatically retrieved to the same debate as their original tweets, distorting the retrieval results.

Although our algorithm assumes that labeled data is not available—with the exception of the active retrieval strategies, where a user provides a few labels—we still need a subset of labeled tweets for evaluation purposes. Yet, sampling this subset is far from trivial. The most unbiased strategy would be to randomly select the instances to be labeled. Let us refer to a sample obtained with this approach as the sampling subset A. Therefore, for the Canadian data we manually labeled 1,000 randomly sampled tweets. However, we observed that most of the retrieved tweets (719 out of 1,000) are not related to any of the target debates. Adopting such a strategy would require us to label several thousands of tweets in order to gather sufficient tweets for each target debate, which would evidently be infeasible.

Our alternative sampling strategy was then to randomly select a percentage of the tweets that are retrieved for each debate by our proposed algorithm. This subset suffices for us to estimate the precision of the proposed retrieval method, but in order to evaluate the recall we must also randomly sample from the non-retrieved pool of tweets. So, from this latter group we sampled as many tweets as the number sampled for the most popular debate. Overall, we manually labeled 2,634 additional tweets for the Canadian data (comprising a sampling subset B), in addition to the 1,000 randomly sampled tweets (subset A). In the case of the Brazilian data, for a good compromise between benefit and labeling effort we only applied the second strategy (sampling B), obtaining 1,064 labeled tweets. Our code and the resulting datasets, including the labeled subsets, are publicly available.[8]

4.2. Parameter Setting

To avoid retrieving tweets that are only loosely related to a debate, we assign tweets to the debate with the highest similarity score if that score is above a given threshold, which in our experiments we set to 1.0. As discussed in Section 3.1, the values in matrix Ω indicate the relevance of the extracted features to the different debates, and hence their contribution to the similarity scores of tweet-debate pairs. Based on an initial appreciation of feature importance, these weights have been initialized to 1.5 for hashtags, 1.0 for keyterms and URLs, and 0.5 for user mentions. Weights are adapted when employing the *ambiguous retrieval* strategy, as described in Section 3.2.1.

The initial number of features is $|D| \times \kappa$ keyterms, $|D| \times \mu$ user mentions and URLs, and $|D| \times \tau$ hashtags, where κ, μ and τ were set to 5, 2 and 1, respectively. The active retrieval component requires setting a single parameter, namely the minimum debate frequency for filtering stop hashtags. As this value should be proportional to the num-

[7] http://politwitter.ca/page/canadian-politics-twitters/mp/house
[8] http://web.cs.dal.ca/~soto/debatesTweets.html

Table II. Accuracy, macro-precision, macro-recall, R-precision and MAP for the unsupervised retrieval method, a random active retrieval strategy, ReQ-ReC, and the ATR-Vis' selection strategies using the Canadian dataset. *The number of labeling requests reported for ReQ-ReC is an average over all debates.

Retrieval method	Accuracy	Macro-Pr	Macro-Re	R-precision	MAP	#Requests
Unsupervised (1^{st} iteration)	0.61	0.74	0.55	0.67	0.70	0
Unsupervised	0.80	0.75	0.68	0.70	0.71	0
Random active retrieval	0.81	0.76	0.70	0.71	0.73	100
ReQ-ReC	0.29	0.26	0.70	0.66	0.64	116*
Ambiguous retrieval (1)	0.83	0.80	0.75	0.75	0.75	15
(1) + Near-Duplicates (2)	0.84	0.81	0.76	0.76	0.76	24
(1) + (2) + Hashtags (3)	0.89	0.82	0.81	0.79	0.80	60
(1) + (2) + (3) + Replies	0.92	0.83	0.86	0.82	0.84	100

ber of debates, in our experiments we filtered out hashtags that appear in over a fourth of the debates, i.e. $df(h_o) > (|D|/4)$. The sensitivity to parameter settings is discussed in Section 4.6. We adopted the same parameter settings on both datasets.

4.3. Retrieval results

We consider multiple evaluation metrics in reporting our results, namely *accuracy*, given by the ratio of correctly retrieved tweets to the total number of retrieved tweets, and *macro-precision* and *macro-recall*, due to the strong imbalance in the distribution of tweets in the debates. Furthermore, we consider two additional metrics that take into account the ranking or scores of the retrieved instances. One is *Mean Average Precision* (MAP), which indicates the average precision of the results across different levels of recall. The other metric is *R-precision*, which measures the precision of the top R retrieved documents, where R is the number of known relevant documents [Manning et al. 2008]. We also report *precision*, *recall*, *R-precision* and *MAP* for each debate.

As per our initial assumption, we found out during labeling that most tweets in our datasets are single-labeled. For those few multi-labeled tweets we consider their retrieval to be correct if they are associated with any of their multiple debate labels. Note that although we can only report the results for the labeled tweets, we perform the retrieval steps and selection strategies considering all tweets in T, all of them having the same probability of being selected for labeling requests regardless of their labeled status.

A summary of the performance of the unsupervised retrieval method and the various active retrieval strategies on the Canadian dataset is presented in Table II, which also includes the number of labeling requests needed by each retrieval approach. For the unsupervised method, we report two separate results. One refers to the method after its first iteration, i.e. considering only the keyterms extracted from debates as discriminative features, while the other refers to applying the full unsupervised method, i.e. considering relevance feedback from the tweets to expand the list of features. We can see from Table II that the pseudo-relevance allows retrieving many more relevant tweets while retaining a comparable precision.

To examine the impact of the proposed selection strategies on improving retrieval accuracy, we first applied each strategy separately and then calculated the number of tweets that were correctly retrieved as a result of its application. When using the hashtag-based strategy, for each non-stop hashtag h_o we simulate the oracle by randomly selecting three labeled tweets from $T(h_o)$. If the three tweets are labeled with the same debate, then all tweets including this hashtag will be considered as belonging to that debate. A similar approach was adopted when simulating the user in the reply-based strategy. The results for each strategy, on 100 labeling requests (divided into blocks of 10 requests), are illustrated in Figure 5. The line representing the ambigu-

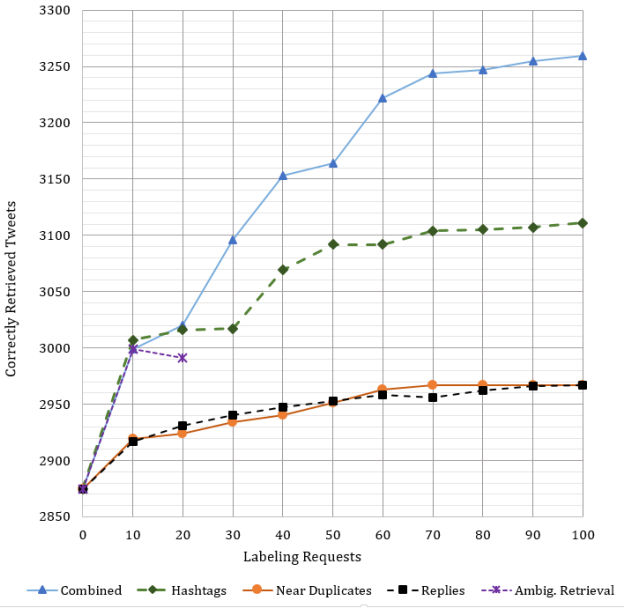

Fig. 5. Canadian dataset: for each selection strategy, the number of correctly retrieved tweets. Blue triangles show the results of applying the four strategies in sequence.

ous retrieval strategy is shorter because only 20 tweets from our test set have nearly equal scores to their most similar debates. We also report the results of applying all the selection strategies in sequence (in the same order as described and reported in Table II). The results highlight that combining these four strategies for retrieval outperforms the independent application of any of them.

Evaluation metrics of the retrieval methods relative to each Canadian debate are presented in Table III. We can see that for most debates the proposed active retrieval method effectively improves both precision and recall, while for some debates it only improves one of these measures while the other drops, or improves much less so. For instance, there is an increase in recall and in precision, respectively, for the debates "Fair Elections Act" and "Marine Mammal Regulations". Moreover, the extent of the improvement varies: it is expected that on debates for which the unsupervised method already performs extremely well, as it is the case of "Meat Inspection Act", the improvement introduced by the active retrieval strategies is not as significant as in those for which the unsupervised retrieval is not as effective, as in the case of "Marine Mammal Regulations".

To further evaluate the effectiveness of the selection strategies, we compare the retrieval results with a random-based selection strategy, which serves as a baseline for our approach. 100 instances are randomly selected and their labels requested from the user. Similarly to our active retrieval strategies, labeled tweets are used to expand the list of discriminative keyterms, which in turn results in more tweets being retrieved. Likewise, we identify specific hashtags that occur in these user-labeled tweets and add them as discriminative features to their corresponding debates. The evaluation measures for this experiment are also included in Table II, which shows averages over 10 runs with randomly selected labeling requests.

We also compare our approach with ReQ-ReC [Li et al. 2014], a state-of-the-art active retrieval method. ReQ-ReC executes two iterative loops, where the outer-loop is responsible for improving recall by forming new queries and retrieving additional rel-

Table III. Results obtained with the unsupervised retrieval method and the ATR method, for each debate in the Canadian dataset. Debate abbreviations stand for: Fair Elections Act (FEA), Meat Inspection Act (MIA), Employment (EMP), Aboriginal Affairs (AAF), Veteran Affairs (VAF), Kidnapping of Girls in Nigeria (KGN), Canada Broadcasting Corporation (CBC), Marine Mammals Regulations (MMR), Housing (HOU), Local Food (LFO), Palliative (PAL).

		FEA	MIA	EMP	AAF	VAF	KGN	CBC	MMR	HOU	LFO	PAL
	#Labeled tweets	670	536	365	394	125	122	101	73	22	8	9
Precision	Unsupervised	**0.98**	**0.98**	**0.98**	**0.9**	**0.99**	**0.92**	**0.79**	0.67	0.89	0.09	0.18
	ReQ-ReC	0.47	0.54	0.35	0.31	0.12	0.14	0.24	0.11	0.09	0.25	0.24
	ATR-Vis	0.97	**0.98**	0.89	**0.9**	0.97	0.89	0.78	**0.88**	**0.91**	**0.35**	**0.67**
Recall	Unsupervised	0.65	0.96	0.63	0.8	0.56	0.81	**0.97**	0.51	0.36	0.37	0.67
	ReQ-ReC	0.64	0.84	0.8	0.68	0.96	0.85	0.55	**0.8**	0.9	0.22	0.45
	ATR-Vis	**0.94**	**0.99**	**0.92**	**0.81**	**0.61**	**0.92**	**0.97**	0.78	**0.95**	**0.75**	**0.89**
R-Prec	Unsupervised	0.87	0.96	0.87	**0.81**	0.61	0.83	0.85	0.15	0.73	0.37	0.67
	ReQ-ReC	0.59	0.82	0.72	0.62	**0.91**	0.85	0.55	0.66	0.87	0.22	0.45
	ATR-Vis	**0.94**	**0.98**	**0.89**	**0.81**	0.64	**0.89**	**0.86**	**0.79**	**0.91**	**0.5**	**0.78**
MAP	Unsupervised	0.91	0.97	0.87	**0.81**	0.77	0.88	**0.88**	0.12	0.78	0.19	0.67
	ReQ-ReC	0.49	0.79	0.73	0.59	**0.9**	0.83	0.48	0.71	0.87	0.22	0.45
	ATR-Vis	**0.95**	**0.99**	**0.91**	0.78	0.76	**0.91**	0.85	**0.77**	**0.95**	**0.59**	**0.85**

evant tweets, while the inner-loop's job is to maximize the precision of the retrieved tweets. The labeling requests are selected considering the uncertainty of an SVM classifier, which is trained as part of the method's inner-loop. At each inner-loop iteration, 10 tweets with the minimum distance to the decision boundary, i.e. 5 from each side, are selected as the labeling requests. Once these requests are labeled by the user, the classifier is retrained and applied to the corpus of retrieved tweets. The inner loop ends when the classifier performance converges. Then, a new query based on the retrieved tweets is formed in the outer loop and used to recover additional relevant tweets.

ReQ-ReC's authors compared variations of their method regarding expanding the query based on retrieved instances. On our Canadian dataset and on the datasets presented in [Li et al. 2014], the "Active" method, which uses Rocchio's method for query expansion [Manning et al. 2008], outperforms any other variation for query expansion. Since ReQ-ReC works under the assumption of having one single query at a time, it has been applied separately to each debate. The results of applying the slightly modified version of the original ReQ-ReC (Active method) to the Canadian political dataset are shown in Tables II and III. The number of labeling requests (#Requests) for ReQ-ReC in Table II indicates the average number over all debates.

Results show that our method, which uses specific Twitter features, outperformed ReQ-ReC on all the evaluation metrics. Moreover, we observe that Macro-Precision for ReQ-ReC is much lower than R-Precision and MAP. This may happen since the system does a good job of finding relevant tweets in the first outer iterations and ranking them at the top of the retrieved tweets. However, most of the lower-ranked tweets retrieved may not be highly relevant to the debate, and since they are considered to formulate new queries, it leads to a deterioration of the overall retrieval precision.

We partly replicated the previous experiments considering the Brazilian parliamentary dataset. A comparison of the retrieval results before and after the application of our retrieval strategies is presented in an aggregated manner and segregated by debates in Tables IV and V, respectively. The results indicate that the pseudo-relevance feedback of the unsupervised approach increase the retrieval recall at the expense of a drop in the precision. This is somewhat expected as adding new features to the debates may introduce spurious tweets. However, the active learning strategies seem to identify the incorrectly retrieved instances as the retrieval precision surpasses the initial values while also succeeding in finding new relevant tweets that previously failed to be retrieved.

Table IV. Accuracy, macro-precision, macro-recall, R-precision and MAP for the unsupervised retrieval method, and the ATR-Vis' selection strategies using the Brazilian dataset.

Retrieval method	Accuracy	Macro-Pr	Macro-Re	R-precision	MAP	#Requests
Unsupervised (1^{st} iteration)	0.74	0.74	0.73	0.70	0.73	0
Unsupervised	0.71	0.72	0.77	0.71	0.66	0
Ambiguous retrieval (1)	0.73	0.73	0.82	0.73	0.67	3
(1) + Near-Duplicates (2)	0.78	0.79	0.82	0.74	0.69	24
(1) + (2) + Hashtags (3)	0.80	0.82	0.88	0.78	0.8	60
(1) + (2) + (3) + Replies	0.77	0.79	0.9	0.78	0.75	90

The results broken down by debates shed some light on the reasons for the retrieval performance. Despite being handled as separate bills at the Brazilian Senate, debates on the topics of social security changes and workers' rights reforms are tightly related. Therefore, it is only after some user feedback that false discriminative keyterms are identified. Better retrieval rates are attained on all debates after simulating user feedback, as depicted in Table V.

Table V. Results obtained with the unsupervised retrieval method and the ATR-Vis' selection strategies, for each debate in the Brazilian dataset. Debate abbreviations stand for: Social security (PRE), Workers' rights (TRA), Political reform (REF), Fiscal adjustments (AJU) and Brazilian Development Bank (BNDES).

		PRE	TRA	REF	AJU	BNDES
	#Labeled tweets	89	116	175	121	129
Precision	Unsupervised	0.42	0.78	0.66	**0.9**	**0.82**
	ATR-Vis	**0.67**	**0.87**	**0.67**	**0.9**	0.81
Recall	Unsupervised	0.67	0.69	0.92	**0.83**	0.75
	ATR-Vis	**0.95**	**0.85**	**0.95**	0.81	**0.93**
R-Prec	Unsupervised	0.47	0.73	0.66	0.83	0.83
	ATR-Vis	**0.56**	**0.82**	**0.73**	**0.84**	**0.94**
MAP	Unsupervised	0.29	0.58	0.75	**0.81**	0.84
	ATR-Vis	**0.56**	**0.74**	**0.82**	0.74	**0.92**

4.4. Use Cases

In this subsection we discuss and compare the application of ATR-Vis to two use cases, the Canadian parliament debates and Brazilian federal senate debates. In order to showcase the suitability of the tool to other domains, another use case featuring major international news during the period 15-27th July 2016 is also presented in Appendix A.

4.4.1. Canadian parliamentary dataset. We now describe how an analyst could use the ATR-Vis framework to retrieve relevant tweets about the Canadian parliamentary debates. A short video showcasing the tool and its interaction functions is provided as part of the Supplementary Material to this paper.

The user is first presented the *Assignment* view, shown in Figure 3. She may start exploring the list of tweets retrieved for each debate and the list of discriminative features, in case some of the retrieved tweets are unexpectedly associated with a mistaken debate. Assuming no major issue is identified in the retrieval, the user may focus on the labeling request presented in the assignment panel (panel a), as addressing the request is likely to improve the retrieval accuracy considerably. In this case the tweet requested is "*@justintrudeau watching to see how you vote on bill c-571, there are enough seals clapping behind stephen harper when he opens his mouth*", and we can

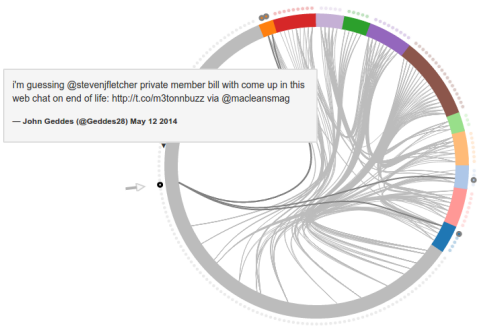

Fig. 6. Ring visualization: identifying weak connections between unretrieved tweets and under-represented debates.

anticipate why the method has found evidence it could belong to the debate on Marine Mammal Regulations, which addresses the regulation of seal hunting in the arctic, and to the debate on "Meat Inspection Act", debated as bill c-571 (notice both debates are shown with a long horizontal bar, unlike the others in the debates list). Its manual labeling to the correct class ("Meat Inspection Act") will contribute to strengthening the evidence that keyterm "c-571" is a highly discriminative feature for this particular debate.

The Ring Visualization, shown in Figure 6, allows the identification of connections among tweets retrieved by different debates, which is an indication of potentially conflictive retrieval. Let us assume that the user is interested in finding more tweets associated with the debate on Palliative Care (shown in orange), which seems to be under-represented in the current retrieval. After exploring the connections between tweets related to Palliative and unretrieved tweets (shown in gray), she finds out that some tweets mention the apparent mentor of the Palliative Care bill (@stevenjfletcher), albeit without using any of its typical keyterms. This interaction could lead to its incorporation into the features characterizing this particular debate, and therefore contribute to increasing the retrieval recall.

The *More* view enables taking advantage of the tweets' structural characteristics to improve the retrieval process (Figure 4). On the left hand side interesting reply chains according to the criterion described in Section 3.2.4 are presented for user inspection using a space-filling tree layout. The one shown refers to a conversation around the topic of seal hunting regulation. Twelve of these tweets have been correctly retrieved to the debate "Marine Mammal Regulations" as indicated by their associated color. However, three tweets were incorrectly retrieved to the debate "Meat Inspection Act", whose inspection and correction represents an opportunity to improve retrieval precision for both classes. The user can notice that other slightly related topics are mentioned in some of the tweets, such as Canadian cultural aspects related to seal hunting. These tweets were not retrieved because they use a different vocabulary than the typical topics mentioned in the debate. Incorporating them increases the recall and enriches the vocabulary associated with this debate. Also some "branches" of this conversation turn into an exchange of aggressive posts not at all discussing debates or political differences. The user can anticipate the nodes where this situation is likely to happen by inspecting branches with several non-retrieved (gray) nodes and refrain from labeling this branch.

In the middle panel, specific hastags can be leveraged by supervising or correcting the most likely retrieval of the tweets that include them. The first nine hashtags shown in the figure are specific to the debates "Fair elections act" and "Employment", with the

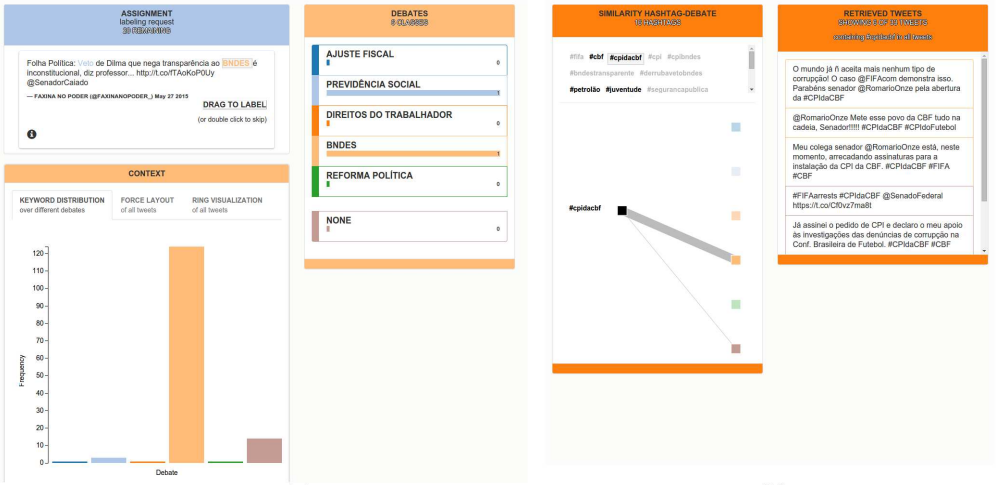

Fig. 7. Beginning of the user involvement considering the Brazilian Parliamentary dataset. (a) Ambiguous retrieval. The word "*veto*" is mistakenly associated to the debate "Social security" (*Previdência social*), and hence in conflict with the correct debate BNDES (b) Leveraging hashtags. The hashtags "#cpidacbf" is mistakenly associated to "BNDES" due to a transitivity with the hashtag "#CPI" which is also frequent in posts related with the debate "BNDES".

exception of "#jobs" which also appear in tweets related to other debates. The user can further inspect the tweets making use of each hashtag (on the right panel) and the bipartite graph depicting hashtag-debate connections for a selected hashtag. She may decide to label the hashtag as a discriminative keyword of a particular debate, with the additional option of automatically retrieving all the tweets including them to that debate. At any time the user can submit the modifications and see the impact of her interactions in the overall retrieval process.

If our user were a political journalist, there are several interesting findings that she could discover from the interaction with the tool. For example, the debate "Fair elections act" attracted by far the most attention in social media. By skimming over some retrieved tweets in this class, it can be noted that major concerns were raised by Twitter users regarding the possibility of using the bill in an anti-democratic spirit by the government in power. In addition, by inspecting the discriminative features for this debate, a list of the URLs, which includes several news articles, can be identified. This can help find the most influential articles on Twitter.

4.4.2. Brazilian federal senate dataset. As opposed to the previous use case, in this case we assume that the user starts interacting with the system from scratch. Therefore, at the beginning several features learnt from the debate transcripts are not completely discriminative. For instance, in the debate about "Social security" (*Previdência social*) the term *veto* is used extensively, while curiously other senators did not use it at all while discussing other debates that week. As a result, some of the initial labeling requests prompt the user to provide feedback when tweets talk about a veto for other bills (Figure 7-a). The user can indicate the correct debate for these cases, which leads to a decrease in the weight of the term *veto* for the "Social security" debate, or directly removing keyword *veto* from its discriminative features.

As indicated by the previous experiments, leveraging hashtags is the most effort-efficient way of providing feedback as it is likely to affect a considerable number of retrieved tweets. Due to alleged corruption cases in Brazil, senators and the general

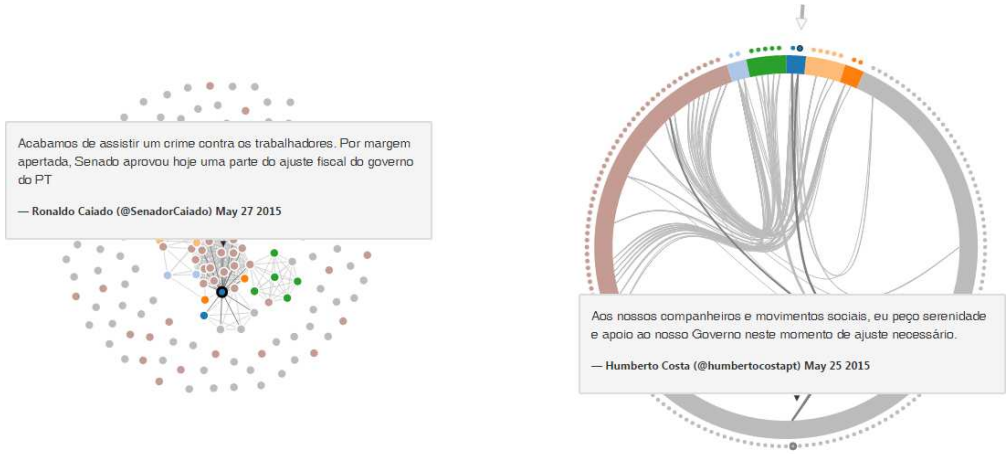

Fig. 8. After a batch of user feedback using the Brazilian Parliamentary dataset. (a) The tweet translates as "*We just witnessed a crime against workers. By a tight difference, the Senate approved today part of the fiscal adjustment proposed by the PT government*" and could be connected to the debates "Fiscal adjustment" and "Workers' rights" (b) The highlighted tweet, which translates as "*To our fellows and social institutions, I ask for your patience and support to our government during this time of necessary adjustments*", signals the association to the "Fiscal adjustment" debate due to the presence of the word "adjustment".

public requested internal investigations through a CPI (Portuguese acronym for "Parliamentary Investigation Committee"). Therefore, the system recognized frequently used hashtags like #cpibndes or #cpidacbf (Figure 7-b). The first one is strictly related to the BNDES debate, whereas the second one is about investigating the Brazilian Confederation of Football. After submitting the appropriate feedback (jointly with the feedback on other hashtags), the user can observe a 70% increase in the number of tweets retrieved.

After a first batch of user feedback, features indeed become more specific, and the following labeling requests presented to the user are ambiguous even for humans. These difficult cases can also be identified from the graph visualizations. In the force-based graph, bridges and clusters of nodes with mixed colors are likely to indicate posts exhibiting some ambiguity regarding the features learned for each debate, as illustrated in Figure 8-a with the debates "Fiscal adjustment" (*Ajuste fiscal*) and "Workers' rights" (*Direitos do trabalhador*). Similarly, by interacting with different link thresholds in the Ring Visualization, it is possible to identify additional non-retrieved tweets that are related to a specific debate. Figure 8-b illustrates a case of a tweet weakly connected to the debate "Fiscal adjustment" because it includes the word "adjustment". There are also replies to these senators' posts, which potentially lead to the identification of additional relevant tweets.

4.5. ATR-Vis Pair Analytics Evaluation

We performed an evaluation of the system with three domain expert users by means of a pair analytics process [Arias-Hernandez et al. 2011], which is carried out with one Subject Matter Expert (SME) and one Visual Analytics Expert (VAE). Our first SME is a journalist and expert in online news and multimedia. As part of her daily work, she searches in Twitter to find interesting topics for stories and to identify active influential users to follow, study and interview with. She is also a university professor teaching journalism in the context of modern digital platforms. The second SME is a

university professor in Sociology and Criminology, whose recent research interests include the portrayal of crime and racism in social media. Our last SME is a student taking a multidisciplinary degree in Criminology and Computer Science, and a research collaborator of the second SME. The latter two users are interested in understanding the discourse related to crimes in social media. In their opinion, analyzing social media is very important for social science researchers as people express their opinions more freely and more honestly than they would in answering questionnaires. We refer to the first, second, and third SMEs as SME1, SME2 and SME3, respectively. The VAE was one of the authors of this paper.

Inspired by the work of Lam et al. [2012], in this evaluation we followed the guidelines of two of their seven evaluation scenarios. The first scenario is "Visual Data Analysis and Reasoning" (VDAR), which discusses the evaluation of tools to support analytical tasks. Since these are typically complex and context sensitive, these evaluations are usually case studies with realistic tasks and domain experts. Questions for this scenario address ways in which the tool can help users to find the information they are seeking, form hypothesis and make decisions. The second scenario is "User Experience" (UE), which aims at evaluating people's opinions and their personal experiences about a tool, to what extent it was successful in assisting them to complete the tasks in their minds, and their suggestions for improvement. Questions for this scenario address the user appreciation of the system, whether they would consider using ATR-Vis in their work/research, and their suggestions for improvements.

Before the evaluation session, we asked the SMEs to quickly familiarize themselves with the debates of our Canadian parliament dataset. They were given links to news articles and Wikipedia pages of the corresponding bills. We started the evaluation session by asking background questions, such as whether they need to search/analyze Twitter data in their profession, what information they look for and how they obtain this information; e.g. whether they use any external tools and whether they feel their information needs are satisfied. Then, we overviewed our dataset and explained the motivation of our system. We followed this by showcasing ATR-Vis along with its main interactive features. Then, the SMEs, assisted by the VAE, conducted the retrieval of tweets by interacting with different features of ATR-Vis. SMEs were encouraged to provide feedback and to review the effects of their interactions on the assignment of tweets, hashtags, and discriminative features to the debates. At the end of the session, they answered questions about ways ATR-Vis can be used to meet their information needs, comments for improving the system, and their general evaluation of the system.

4.5.1. Before pair analytics with ATR-Vis: SMEs' background. As part of her daily work, in order to find relevant information or stories about a topic, SME1 tends to search for an active Twitter account or a hashtag related to that topic. She uses Twitter advanced search and Hootsuite[9] to find such accounts, and Storify[10] and Banjo[11] to generate stories and identify people to interview. However, SME1 stated that the tools she is currently using miss relevant information: "*we tend to fall back to those things that are easiest often because we are rushed, we follow an account, or we follow a hashtag, but we do miss a whole lot. Because there are relevant tweets that do not have those keywords and that has always been a problem*".

SME2 has extensive experience in performing content analysis or discourse analysis on traditional media such as news articles. However, she has not used any content analysis tools, but manually studied the news articles, as in her opinion, content and

[9]https://hootsuite.com/
[10]https://storify.com/
[11]http://ban.jo/

discourse analysis tools are not rich enough to always capture the meaning of the article. She added that ensuring that important information is not missed is a very difficult task and an issue for social science research. SME2 plans to perform content and discourse analysis on Twitter for her research, and she knows that she cannot rely only on some keywords for finding relevant information.

SME3 has made use of Twitter advanced search and stated that its results are not very accurate: "*It gives you tweets that does not have anything with what you are looking for*". She added that in her opinion this search engine also misses relevant tweets. She supports her opinion with an example about how hashtags can result in receiving irrelevant information. During the last winter Olympics, the hashtag "#WeAreWinter" was used in Canada in tweets supporting Canadian teams or reporting news about these games. However, some people used this hashtag in posts not related with the Olympics. For instance, somebody might just say "I just went to the supermarket #WeAreWinter".

4.5.2. During pair analytics with ATR-Vis: main interactions. SME1 found that "*the tool is pretty straightforward*" to use and that showing the discriminative features are useful especially because the user can control their assignment to different debates. She also mentioned that the Similarity Hashtag-Debate panel in the More view is useful for labeling all tweets containing a specific hashtag with one single assignment, which in turn may reduce the number of labeling requests as well. In addition, in her opinion, the Force Layout View is very helpful in determining the clusters of tweets. Therefore, the user can perform a deeper analysis on these clusters and see whether a story exists or not. For instance, SME1 commented that there is a dense cluster for the "Fair Elections Act" and she would be interested in analyzing this cluster to determine whether "*one political party or one political group is really responsible for a lot of this conversation*". She also mentioned that the Force Layout can be used for determining debates with a broader range of topics from the composition of the clusters. She exemplified his remark with the observation that debate "Aboriginal Affairs" seems to have a broader range of topics, which are in common with other debates, as compared to "Fair Elections Act".

SME2 liked all the features and mentioned that the Assignment View is "*very useful and user friendly*". She found particularly useful that she can assign discriminative features to different debates and see the effect of this assignment on how tweets are retrieved. She added that: "*people make a lot of bizarre references to things that have nothing to do with something else*" and therefore it is important that ATR-Vis gives the ability to examine tweets retrieved by the automatic method and change the debates/topics of tweets. She mentioned that such a feature is very useful when the computer or the user makes a mistake: "*It alleviates human error that is riled in social sciences*". Also, by looking at different branches of a conversation and the colors of its nodes in the Reply Tree, she noted how fast people can change their minds about a topic.

In SME3's opinion all features of ATR-Vis are useful: "*I think it is all really useful and what part becomes the most useful depends on the individual topic*". For instance for some topics, the Reply Tree will be extremely useful, but with other topics, it might not be as useful as visualizations in the context view. Then, she continued "*I do not think that there is anything on here that is a waste of space and it is all useful*". She also found the ability to make corrections even to her own errors, with simple interactions such as drag and drop, very helpful.

4.5.3. After pair analytics session: questionnaire. We posed three main questions to SMEs after finishing using the tool: "What advantages and possible other uses you find for ATR-Vis?", "Would you consider using this system for your own work/research?" and

"What limitations did you find and/or what suggestions can you give us to improve the tool?"

What advantages and possible other uses you find for ATR-Vis?

SME1 mentioned that ATR-Vis can be useful for other tasks that go beyond the goal of just searching for tweets. The first task is to find active Twitter accounts in her topics of interest: "*I can see this being useful at various points along the process for a journalist, one is looking for people...When you are assigned to a story and you are doing background information, so one way would be to find people. Because if you find people who are actively engaged on Twitter, you can track them down, you can call them up, you can do interviews*". The second task is to learn about emerging topics and events. The SME currently utilizes Google News alerts to receive information about her topics of interest, but she mentioned that it searches for News stories only, not tweets. The last task is to look for patterns and trends and identifying related debates/topics from clusters of tweets that look interesting for story ideas; e.g. "Marine Mammal Regulations" may be related to "Employment" considering the Inuit communities. She also exemplified this point saying: "*just looking at the connection between Nigerian girls and missing and murdered indigenous women, ... people are kind of putting the two of them together; that could be a news story*". SME1 also highlighted the serendipity allowed by ATR-Vis: "*Sometimes we don't know what it is we are looking for and sometimes it's like you have a hypothesis, so I think I know what my story is about, but I can ignore the evidence or I can ignore what is in front of me and I have to rethink my story and my focus and then reassess. So, a tool like this is great at every step of the game of the story*".

SME2 pointed out that showing the similarities between hashtags and debates in the Similarity Hashtag-Debate panel is useful for the content/discourse analysis as only considering the appearance of hashtags in tweets cannot always capture their meaning: "*in criminology people use a lot of hashtags specially with race, issues, fear and crime in general*". SME2 added that ATR-Vis can be used in identifying the connections that people make: "*in my line of research that is what makes it important and it is something that I could never do on my own and that's what makes a program like this so important is to actually look at the verbal connections that people make on their own*".

SME3 commented on the possibilities for ATR-Vis to gather accurate public discourse: "*I definitely stand by thinking that it's going to be the best way to get really good public discourse on issues in any social science*". She added that performing traditional research in social sciences, through face-to-face interviews or questionnaires, has many difficulties such as finding people who are willing to be interviewed in today's fast-paced world and also avoiding social desirability bias. In her opinion, ATR-Vis can help social science researchers gather more accurate opinions faster and easier than traditional methods in social sciences, which is very important in their line of research. She concluded saying: "*ATR-Vis is far more accurate than using Twitter Search*".

Will you consider using this system for your own work/research?

SME1 stated: "*I would definitely try this again*" and "*I would even have this as one of the tools in our students' toolbox to use when they are working on their stories*". SME2 commented "*This has actually exceeded all of my expectations because it just makes the possibility of my research big*". She added that the research possibilities are endless and the fact that there is a system that can make it happen is interesting. The SME mentioned that, although she is relatively new to the study of social media, she finds ATR-Vis very useful: "*This is something that I would use for every single piece of research, something that students can do master theses on*". SME3's response was: "*Yes, it is definitely very user friendly and well designed*".

What limitations did you find and/or what suggestions can you give us to improve the tool?
Both SME1 and SME2 commented that they miss the capability of adding new debates on the fly to the list of debates. For instance, SME1 mentioned that she found interesting tweets and discussions about the debate on abortion among the tweets and wanted to add this topic/debate to the list in order to retrieve its relevant tweets.

SME1 added that being able to put different tags/notes on tweets, hashtags and features would be very useful, especially when multiple users are working together or the user is interacting with the system in multiple sessions. For instance, the user may put some tags showing how certain she is about the label of tweets and hashtags or even about the authority of an account. In this case, she may want to further investigate cases with low certainty or ask her colleagues' opinions about them. Finally, integrating ATR-Vis with Facebook posts and Instagram is another addition to the system commented by SME1. She also mentioned that being able to track back stories and see when they started and what was the trigger, i.e. having a timeline, would be useful.

4.6. Discussion

In this subsection we discuss various aspects of the experiments described above.

4.6.1. Numerical experiments. The experimental results obtained on two distinct datasets showed the advantages of applying our selection strategies. In general, the poorer the retrieval for a given class is, the more the debate benefits from the active retrieval strategies for improving its retrieval results. The selection strategies based on the ambiguous retrieval and hashtags are the most effective ones as shown in Tables II and IV. However, the strategy of simulating the user introduces certain limitations. For the Canadian dataset we leveraged 12 hashtags, which generate 36 requests as we assume that 3 tweets are needed to inspect the hashtag. More realistically, in view of her domain knowledge a user may not need to inspect a hashtag to understand how it is used. Likewise, the reply strategy is more suitable for visual inspection rather than for a massive retrieval after some random posts in the reply chain are inspected.

4.6.2. Sensitivity due to data sampling. We also analyze how sensitive the experimental results are to the specific sample considered by investigating how they are affected when more labeled tweets are added. Therefore, we first considered only subset B, which contains 60% of the retrieved tweets for different debates, as our test set and then we evaluated the effects of adding to it the 1,000 randomly selected tweets of subset A. Precision remains stable for most debates as it is observed in Figure 9. For debates with relatively few tweets such as "Local Food" and "Palliative and End of Life", which are more sensitive to small changes in their retrieved tweets, there is a reduction in their precision. This also explains the low precision of the retrieval methods (Table III), which results from retrieving a few wrong tweets for these debates. For recall, the effect of adding subset A into subset B is even less significant. The comparison of the overall accuracy ($B = 0.98, B \cup A = 0.99$), macro-precision ($B = 0.82, B \cup A = 0.83$) and macro-recall ($B = 0.84, B \cup A = 0.85$) also indicates no major effect of extending the test set with randomly sampled tweets.

4.6.3. Model selection. The results reported in Tables II and III show the accuracy of our methods for the parameter settings introduced in Section 4.2. Although the performance of the method may be sensitive to different parameter configurations, we hypothesize that the proposed active retrieval strategies can compensate for negative effects incurred due to parameters not set optimally. The accuracy of our unsupervised and ATR methods in terms of the number of keyterms used in the initial step, κ, is presented in Figure 10. The solid blue line shows the accuracy for the unsupervised

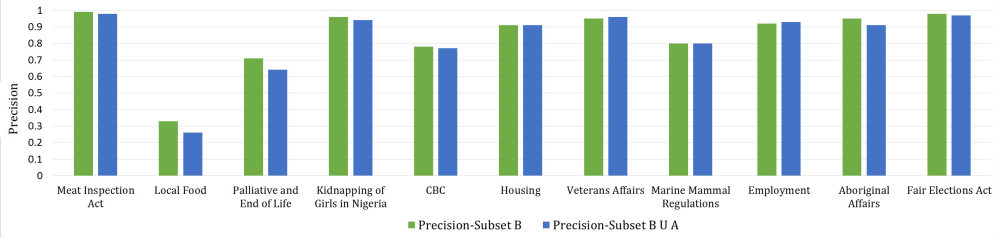

Fig. 9. Precision for each debate before and after adding tweets of subset A to subset B.

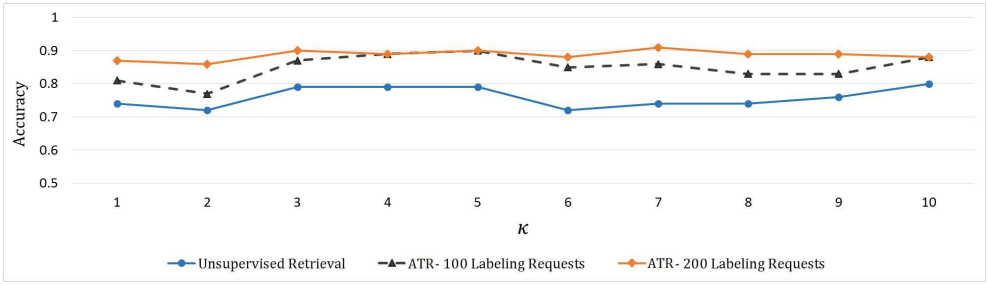

Fig. 10. Accuracy before and after applying ATR strategies with different values for the number of keyterms. A value of $\kappa = 5$ is used in our experiments.

retrieval model, where one observes that performance is affected by the choice of this parameter. The black-dotted line represents the result after the ATR strategies, and we can see that these strategies improve the results of the unsupervised retrieval fairly independent of the parameter choice. The solid orange line shows the results of ATR after doubling the number of labeling requests. The result supports our hypothesis that additional requests to the user can compensate a non-optimal parameter configuration, as in all cases a similar upper bound in accuracy is obtained regardless of the number of keyterms used.

While developing the active retrieval strategies we tested several other approaches. This involved different keyterm extractions for debates and tweets, different text similarities, different vector-space text representations (e.g. character and word-bigrams with binary, integer and real-valued weights) and different approaches to identify stop hashtags (e.g. degree centrality on the co-occurrence graph). We selected the design options that worked best for our tasks, while we favored the simpler model or approach when no clear difference was observed.

4.6.4. Use cases. The debates in the Brazilian dataset are more semantically related to each other than the Canadian one, and hence their retrieval performance is lower. Furthermore, the week of May 2015 was shaken by the news of investigations on corruption and arrests of high-level managers in the federal international football association (FIFA). Given that football is a very popular and sensitive topic to many Brazilians, this resulted in a large number of noisy posts being merged with the discussions of the ongoing debates at the Brazilian Senate. Yet, this allowed us to evaluate ATR-Vis in the context of uncontrolled external events.

Regarding the use cases, some interesting differences were observed between the two datasets. The typical reply pattern in the Brazilian dataset is the public replying to senators without interacting among each other, while in the Canadian dataset people tend to engage more in conversations. In the Brazilian dataset we also noted that

tweets involving the senator Romário de Souza Faria—who used to be a highly popular football player—seem to generate a substantially higher traction among Brazilians compared to any other senator. In the case of the Canadian Parliament, we note that the leaders of the major political parties are the ones that trigger most of the tweets followed by those senators involved in the proposal of the bills of interest.

4.6.5. Retrieval error analysis. The proposed automatic retrieval method is not flawless, which indeed prompted our motivation for incorporating the user into the retrieval loop. This is particularly the case at the beginning where some initial features may not be good indicators for discriminating the debate of interest, so user intervention is necessary. The correct identification and filtering of hashtags by the system plays a major role in the retrieval of tweets. False negatives in the identification typically leads to a large number of tweets being retrieved to the wrong debates, or failing to retrieve those. False positives are less harmful but imply in unnecessary labeling effort from the user.

Inadvertent user mistakes when providing feedback are also possible. In some scenarios, some of the visual components of the tool can be useful to detect such inconsistency. For instance, if a user mistakenly assigns hashtag "#c23", which is the bill number for "Fair Elections act", to the debate "Marine Mammal Regulations", numerous connections between these two debates in the Ring Visualization provide an indication that there has been an error in the retrieval/supervision, especially if those connections are unexpected for the user. Subsequent inspection of the tweets and their discriminative keywords can fix the incorrect supervision.

However, for some cases neither the system nor the user could identify possible inconsistencies, and hence incorrect retrieval can pass unnoticed. Most of these errors result from difficulties in understanding the semantics of tweets, as the method mostly relies on vector similarities. Sarcastic tweets are one common example. Another example is spam tweets that use misleading URLs or hashtags just for the sake of being visible among trending topics.

5. CONCLUSIONS

This paper presented ATR-Vis, a user-driven visual framework for active retrieval targeted specifically for Twitter data. This framework addresses an existing challenge in the analysis of social media data, which is to assure that the information relevant to an analytical task is retrieved attempting to maximize both recall and precision. The proposed framework has been applied and evaluated in a task scenario of retrieving Twitter posts related to a set of target debates occurring in a parliament house over a certain period.

The experimental results demonstrate that the framework can successfully integrate a user into the retrieval task so as to improve both retrieval precision and recall. User involvement is kept to a minimum by carefully selecting and submitting Twitter entities, i.e. posts or hashtags, for user supervision based on their estimated potential to improve the retrieval outcome. Our strategies for selecting these entities were favorably compared against other approaches including a state-of-the-art system for interactive retrieval.

The interactive interface enables a user to explore, inspect and modify the retrieval process, so that user interactions actually modify how the system works. It gives non-technical users—who might be political analysts or journalists—the tools for obtaining a reliable Twitter collection responding to their information needs before carrying out any data analyses. Furthermore, this user-driven approach yields a higher versatility to adapt the framework to different domains without any additional model refinement, which would typically require a data mining expert. We showcased possible flows of

analysis using ATR-Vis in the context of two datasets corresponding to the retrieval of tweets related to parliamentary debates being held in Canada and Brazil. No language tuning was required in handling content in English or in Portuguese, as all methods implemented in the framework are language independent.

In order to have a stronger understanding of ATR-Vis, its different interactive visual features, and its efficiency in assisting users in finding relevant information, we performed an evaluation with three domain experts. All three experts provided positive feedback for ATR-Vis, and acknowledged the need for this type of tools for the accurate retrieval of tweets. They also shared interesting insights on potential improvements and further developments. Although beyond the scope of this article, we have also applied ATR-Vis for the retrieval of non parliamentary debates, where we considered a set of top news stories that received great media attention as our topics of interest to retrieve tweets about. While Appendix A discusses some use cases showing how ATR-Vis can support the retrieval of relevant tweets to selected news stories, our assumption is that in other domains additional features related to poster's time and location, URL's text content and user profiles need to be further investigated.

There are possible avenues for extending this work. We conceived ATR-Vis as a necessary step for retrieval of Twitter data, where retrieved posts are likely to be pipelined to other tools for content analysis. In addition, assigning impact scores to each labeling request based on its probability of improving the overall accuracy of the retrieval can help users in assessing the labeling effort. Finally, another interesting aspect is to evaluate this framework in the context of stream data—as opposed to the static dataset evaluations considered in this paper. This will involve finding strategies to see how the system faces the cold start problem at the beginning and how it should "forget" data when memory is exceeded.

ACKNOWLEDGMENTS

This work was carried out with the aid of grant 2013-LACREG-07 from the International Development Research Centre, Ottawa, Canada, a CALDO-FAPESP grant (Proc. 2013/50380-0), and an ELAP scholarship. Brazilian researchers are also supported by grants from CNPq (205291/2014-7) and FAPESP (2011/22749-8), and researchers based in Canada by grants from NSERC. The authors thank "Cheng Li" for sharing the code of ReQ-ReC.

REFERENCES

Noa Aharony. 2012. Twitter use by three political leaders: an exploratory analysis. *Online Information Review* 36, 4 (2012), 587–603.

Aretha B Alencar, Maria Cristina F de Oliveira, and Fernando V Paulovich. 2012. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 6 (2012), 476–492.

Richard Arias-Hernandez, Linda T Kaastra, Tera M Green, and Brian Fisher. 2011. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.

John Carlo Bertot, Paul T Jaeger, Sean Munson, and Tom Glaisyer. 2010. Social media technology and government transparency. *Computer* 11 (2010), 53–59.

Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, and others. 2011. Structural and dynamical patterns on online social networks: the Spanish May 15th movement as a case study. *PloS one* 6, 8 (2011), e23883.

Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Puttmann, Steffen Koch, Robert Kruger, Michael Worner, and Thomas Ertl. 2013. ScatterBlogs2: Real-Time Monitoring of Microblog Messages through User-Guided Filtering. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2022–2031.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology, 2012 IEEE Conference on*. IEEE, 143–152.

Yan Chen, Zhoujun Li, Liqiang Nie, Xia Hu, Xiangyu Wang, Tat-seng Chua, and Xiaoming Zhang. 2012. A Semi-Supervised Bayesian Network Model for Microblog Topic Classification.. In *Proceedings of the 24th International Conference on Computational Linguistics*. Citeseer, 561–576.

Peter Cogan, Matthew Andrews, Milan Bradonjic, W Sean Kennedy, Alessandra Sala, and Gabriel Tucci. 2012. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*. ACM, 25–31.

Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* 133 (2011), 89–96.

Danish Contractor, Bhupesh Chawda, Sameep Mehta, L Venkata Subramaniam, and Tanveer A Faruquie. 2015. Tracking political elections on social media: applications and experience. In *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2320–2326.

Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology, 2010 IEEE Symposium on*. IEEE, 115–122.

Marian Dörk, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. 2010. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on* 16, 6 (2010), 1129–1138.

Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. 2012. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology, 2012 IEEE Conference on*. IEEE, 93–102.

Renato Miranda Filho, Jussara M. Almeida, and Gisele L. Pappa. 2015. Twitter Population Sample Bias and its Impact on Predictive Outcomes: A Case Study on Elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. ACM, 1254–1261.

Lorenzo Gabrielli, Salvatore Rinzivillo, Francesco Ronzano, and Daniel Villatoro. 2014. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in Sensor Networks*. Springer, 26–35.

Devin Gaffney. 2010. #iranElection: Quantifying online activism. In *Proceedings of WebSci10: Extending the Frontiers of Society On-Line*.

Mona Golestan Far, Scott Sanne, Mohamed Reda Bouadjenek, Gabriela Ferraro, and David Hawking. 2015. On Term Selection Techniques for Patent Prior Art Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 803–806.

Anatoliy Gruzd and Jeffrey Roy. 2014. Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet* 6, 1 (2014), 28–45.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

Davide F Gurini and Fabio Gasparetti. 2012. *TREC Microblog 2012 Track: Real-Time Algorithm for Microblog Ranking Systems*. Technical Report. DTIC Document.

Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*. IEEE, 115–123.

Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics* 12, 5 (2006), 741–748.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. ActNeT: Active Learning for Networked Texts in Microblogging. In *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM13)*. 306–314.

Ajaz Hussain, Khalid Latif, Aimal Tariq Rextin, Amir Hayat, and Masoon Alam. 2014. Scalable visualization of semantic nets using power-law graphs. *Applied Mathematics & Information Sciences* 8, 1 (2014), 355–367.

Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 159–162.

Tommi Jaakkola and Hava T. Siegelmann. 2001. Active Information Retrieval. In *Advances in Neural Information Processing Systems*. 777–784.

Jari Jussila, Jukka Huhtamäki, Hannu Kärkkäinen, and Kaisa Still. 2013. Information visualization of Twitter data for co-organizing conferences. In *Proceedings of International Conference on Making Sense of Converging Media*. ACM, 139–145.

Daniel A Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. 2010. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann.

Andy Kirk. 2012. *Data Visualization: a successful design process*. Packt Publishing Ltd.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 251–258.

Cheng Li, Yue Wang, Paul Resnick, and Qiaozhu Mei. 2014. Req-rec: High recall retrieval with query pooling and interactive classification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 163–172.

Mengchen Liu, Shixia Liu, Xizhou Zhu, Qinying Liao, Furu Wei, and Shimei Pan. 2016. An Uncertainty-Aware Approach for Exploratory Microblog Retrieval. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 250–259.

Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, and Baining Guo. 2014. Topic Panorama: a Full Picture of Relevant Topics. In *Proceedings of the IEEE Symposium on Visual Analytics and Science Technology (IEEE VAST 2014)*. IEEE Computer Society, 183–192.

Yafeng Lu, Robert Krüger, Dennis Thom, Feng Wang, Steffen Koch, Thomas Ertl, and Ross Maciejewski. 2014. Integrating predictive analytics and social media. In *Visual Analytics Science and Technology, 2014 IEEE Conference on*. IEEE, 193–202.

William Lucia and Elena Ferrari. 2014. Egocentric: Ego networks for knowledge-based short text classification. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1079–1088.

Zhunchen Luo, Miles Osborne, and Ting Wang. 2012a. Opinion Retrieval in Twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. 507–510.

Zhunchen Luo, Miles Osbornez, Saša Petrovic, and Ting Wang. 2012b. Improving Twitter Retrieval by Exploiting Structural Information. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. 648–654.

Alan M MacEachren, Anuj Jaiswal, Anthony C Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. 2011. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology, 2011 IEEE Conference on*. IEEE, 181–190.

Raheleh Makki, A.J. Soto, S. Brooks, and E. Milios. 2015. Active Information Retrieval for Linking Twitter Posts with Political Debates. In *IEEE International Conference on Machine Learning and Applications*. IEEE, 1–10.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, and others. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge.

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in information retrieval*. Springer, 362–367.

Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. 2013. Improving pseudo-relevance feedback via tweet selection. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 439–448.

Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. 2013. Understanding twitter data with tweetxplorer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1482–1485.

Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.

Chris North and Ben Shneiderman. 2000. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 128–135.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 384–385.

Victor Pascual-Cid and Andreas Kaltenbrunner. 2009. Exploring asynchronous online discussions through hierarchical visualisation. In *2009 13th International Conference Information Visualisation*. IEEE, 191–196.

M-H Peetz, Damiano Spina, Julio Gonzalo, M Rijke, and others. 2013. Towards an active learning system for company name disambiguation in microblog streams. In *CEUR Workshop Proceedings*. CEUR.

Philips Kokoh Prasetyo, Palakorn Achananuparp, and Ee-Peng Lim. 2016. On analyzing geotagged tweets for location-based patterns. In *Proceedings of the 17th International Conference on Distributed Computing and Networking*. ACM, 45–50.

Runwei Qiang, Feifan Fan, Chao Lv, and Jianwu Yang. 2015. Knowledge-based query expansion in real-time microblog search. *arXiv preprint arXiv:1503.03961* (2015).

Jonathan C Roberts. 2007. State of the art: Coordinated & multiple views in exploratory visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on*. IEEE, 61–71.

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 695–704.

Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. Topical clustering of tweets. *Proceedings of the ACM SIGIR: Social Web Search and Mining* (2011).

David Shamma, Lyndon Kennedy, and Elizabeth Churchill. 2010. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events. *CSCW Horizons* (2010), 589–593.

Malcolm Slaney and Michael Casey. 2008. Locality-sensitive hashing for finding nearest neighbors. *Signal Processing Magazine, IEEE* 25, 2 (2008), 128–131.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 841–842.

Guodao Sun, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan J. H. Zhu, and Ronghua Liang. 2014. EvoRiver: Visual Analysis of Topic Competition on Social Media. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1753–1762.

James J Thomas, Kristin Cook, and others. 2006. A visual analytics agenda. *Computer Graphics and Applications, IEEE* 26, 1 (2006), 10–13.

Michelle Q Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. 2000. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 110–119.

Colin Ware. 2012. *Information visualization: perception for design*. Elsevier.

Jinxi Xu and W Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 4–11.

Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, J.J.H. Zhu, and Huamin Qu. 2013. Visual Analysis of Topic Competition on Social Media. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 2012–2021.

Online Appendix to: ATR-Vis: Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter

RAHELEH MAKKI, Dalhousie University, Canada
EDER CARVALHO, Universidade de São Paulo, Brazil
AXEL J. SOTO, Dalhousie University, Canada
STEPHEN BROOKS, Dalhousie University, Canada
MARIA CRISTINA FERREIRA DE OLIVEIRA, Universidade de São Paulo, Brazil
EVANGELOS MILIOS, Dalhousie University, Canada
ROSANE MINGHIM, Universidade de São Paulo, Brazil

A. USE CASE: RETRIEVING TWEETS ASSOCIATED TO RECENT NEWS STORIES

ATR-Vis has been also assessed with other non-parliamentary datasets, where news stories that received great attention from the media during the period 15-27 July 2016 have been used, namely: "Terrorist attack in Nice", "Brexit", "Colombia's government and FARC", "Dallas shooting", "Israeli-Palestinian conflict", "Killing of Afro-Americans", "Orlando nightclub shooting", "Refugee crisis", "Rio 2016 Olympics", "Turkey attempted coup", "US Presidential campaign". We considered news articles from CNN and Fox News related to each story to extract keywords and set our initial queries. The tweets collected during this period accounted for 9,277,751 after retweets and non-English posts were discarded. We have made this dataset also publicly available.[12]

We now describe how a user could interact with ATR-Vis to retrieve relevant tweets to these stories and even learn more about them. Due to several terrorist attacks and events of a violent nature that happened during the period we collected our dataset, the selected news stories have a high degree of similarity to each other, which makes the automatic retrieval of tweets a difficult task. The user can perceive this similarity between the stories through different visual components of ATR-Vis. The user may start with the Assignment View and the first suggested labeling request. As shown in Figure 11 the labeling request is "BoingBoing: RT AkyolinEnglish: 17 Turkish police officers killed in Ankara - by the junta-would-be. Horrible. It seems the coup wont be sub...". While "Turkey attempted coup" is the correct association for this tweet, the system also finds "Dallas shooting" as a potential candidate, due to the presences of terms: "police" "officers" and "killed". After the user assigns the tweet to the correct event, the retrieval system also learns from this interaction to dampen the ambiguous terms for "Dallas shooting" as they may also appear in other stories.

Then the user may want to focus on the context panel, where she can explore the connections between tweets based on their similarities to the stories. Figure 12 shows that "Dallas shooting" (in orange) and "Killing of Afro-Americans" (in pink) are strongly related stories. For instance, the inspected tweet: "Watch: Baton Rouge: Timeline of Shooting: A look at the attack that left three police officers dead. https://t.co/UjqfSO8RXh" contains elements from both stories as there is a mention to killed officers, and to Baton Rouge, the city where an Afro-American was killed. ATR-Vis is able to find the connections between these stories, and also find intermediate

[12] http://web.cs.dal.ca/~soto/debatesTweets.html

 1556-4681/2017/07-ARTA $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

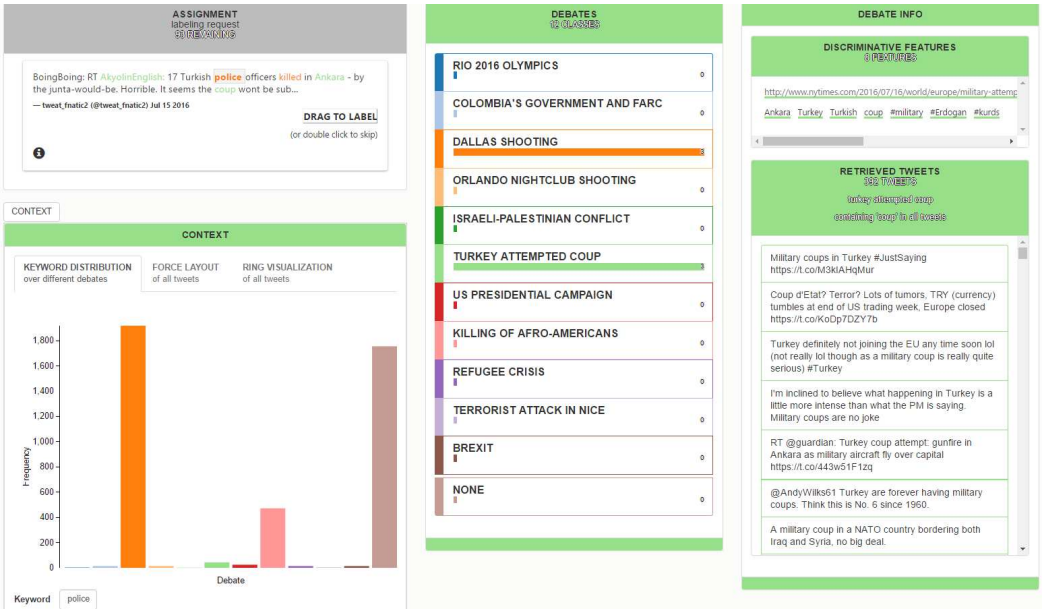

Fig. 11. The Assignment View showing the labeling request, discriminative features, and keyword distributions for tweets on emerging news stories.

Fig. 12. The Force Layout shows the tight connection between the story of "Dallas shooting" (in orange) and "Killing of Afro-Americans" (in pink).

stories (appearing as a hub between two clusters) that may not belong to any of the stories of interest.

Let us assume the user is interested in finding more tweets relevant to the story "US Presidential campaign" (shown in red), which seems to be under-represented. Exploring the connections between tweets related to this story and non-retrieved tweets (shown in gray), allows the user to find some relevant tweets, which in turn refines its discriminative features and improves its retrieval accuracy (recall and precision). For instance, Figure 13 shows the tweet "LA Times suggests MILITARY COUP when Trump wins Presidency https://t.co/oiRGbunXj3", which also has connections to tweets

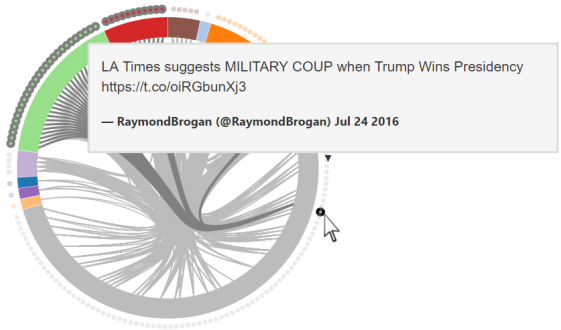

Fig. 13. ATR-Vis can be used to improve the recall of a story. This image shows the connection of a non-retrieved with two of our stories, which the user can inspect to decide on its correct retrieval.

associated with the story “Turkey attempted coup”. Since this tweet contains the terms “military” and “coup”, ATR-Vis could not make a confident decision about its assignment. Associating this tweet with “US Presidential campaign”, results in adding the expanded URL of the tweet to the list of discriminative features for this story, which will be in turn used to retrieve more relevant tweets.

The last interaction is with the Similarity Hashtag-Debate. The identification of hashtags that are good indicators of our selected news stories and assigning them to the proper story can be an important contribution to improve recall and precision of the retrieved tweets. After reviewing different hashtags and the tweets containing them, the user can easily assign hashtags like “#CharlesKinsey”, “#TurkeyCoupAttempt” and “#NiceFrance” to their corresponding story. In addition, there are other less obvious hashtags, like “#BlueLivesMatter”, which after some inspection can be understood as a response to the hashtag “#BlackLivesmatter”, in support for the officers killed in the “Dallas shooting”.