Feedback-Based Social Media Filtering Tool for Improved Situational Awareness

Jason Thornton, Marianne DeAngelus and Benjamin A. Miller
MIT Lincoln Laboratory
Email: jason.thornton@ll.mit.edu, deangelus@ll.mit.edu, bamiller@ll.mit.edu

***Abstract*—This paper describes a feature-rich model of data relevance, designed to aid first responder retrieval of useful information from social media sources during disasters or emergencies. The approach is meant to address the failure of traditional keyword-based methods to sufficiently suppress clutter during retrieval. The model iteratively incorporates relevance feedback to update feature space selection and classifier construction across a multimodal set of diverse content characterization techniques. This approach is advantageous because the aspects of the data (or even the modalities of the data) that signify relevance cannot always be anticipated ahead of time. Experiments with both microblog text documents and coupled imagery and text documents demonstrate the effectiveness of this model on sample retrieval tasks, in comparison to more narrowly focused models operating in *a priori* selected feature spaces. The experiments also show that even relatively low feedback levels (i.e., tens of examples) can lead to a significant performance boost during the interactive retrieval process.**

I. Introduction

Driven by the need for more comprehensive situational awareness during disasters and emergency situations, the first responder community is increasingly augmenting traditional closed information sources with information available through open sources, such as publicly available social media posts. There are two fundamental use cases in which this information is helpful to law enforcement and first responders: the detection of indicators and warnings before an event, and the assessment of impact after an event. In the second case, situational awareness can be improved by descriptions of the after-effects of the event, such as reports of infrastructure damage, physical health, or supply shortages. Useful information may take the form of text-based descriptions or posted imagery. In general, the value of social media information in these cases comes from the ease of sharing it via open platforms and the distributed contributions from many users.

The primary challenge for analysts who wish to exploit this information is effectively filtering the large volumes of content in order to separate relevant posts from clutter, or irrelevant posts with no bearing on the situation. One way to attempt this sorting is the use of keyword queries. However, keyword queries crafted by an analyst are often an imprecise expression of the information need, because of the ambiguities and inconsistent usage patterns of natural language. In addition, the relevance of the post containing the keyword(s) is sometimes determined not just by the presence or absence of those words, but by the context of their use. Finally, for some posts the relevance may not be determined by the text component but by the content of associated imagery or video.

For the reasons listed above, text and multimedia search capabilities can benefit from relevance feedback as a way to refine the search process [1], [2]. In the relevance feedback paradigm, the user inspects records returned by the retrieval system and indicates the relevance of individual documents through a feedback mechanism. In its simplest form, the feedback is an assignment of binary labels representing a "relevant" or "irrelevant" classification. The system can then use these reference documents to perform a more accurate retrieval, since the examples represent a better expression of user needs than keywords alone.

For effective analysis of multimedia data extracted from social media sources, the feature space(s) in which this analysis is performed is a critical design decision. However, the breadth of possible information extraction needs associated with emergencies makes it difficult to predict what aspects of the data will allow for accurate discrimination between content-of-interest and background clutter. Therefore, the proposed approach outlined in this work is inspired by the success of feature-rich approaches in multimedia retrieval tasks [3], [4], using such rich representations to support classifier refinement based on iterative relevance feedback. Notably, this approach leverages feature sets that are broader and more diverse than those used in previous work on relevance feedback systems [5], [6].

The rest of this paper describes the proposed algorithm for data retrieval, empirical results on benchmark datasets, and the construction of a software prototype that implements the proposed technique.

II. Proposed Data Retrieval Framework

The data retrieval framework described in this paper, and used to generate the experimental results, has several components as depicted in Fig. 1. The retrieval is initiated by a basic search filter, such as a keyword search. The output from this filter will contain some mixture of relevant and irrelevant documents. As the user provides relevance feedback,

This work is sponsored by the Assistant Secretary of Defense (Research and Engineering) under Air Force Contract FA8721-05-C-0002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government.

978-1-5090-0770-7/16/$31.00 ©2016 IEEE

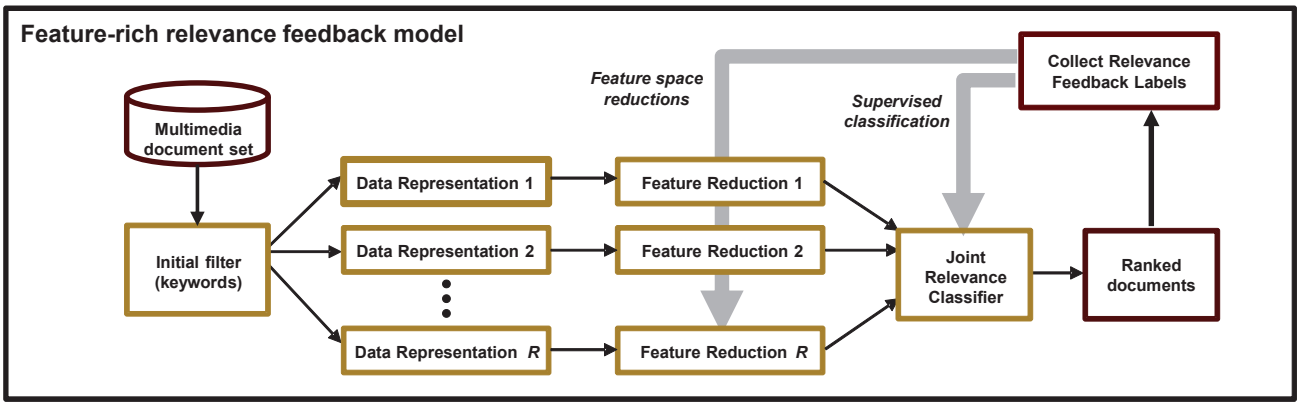

Fig. 1. Components of the feature-rich retrieval process, based on relevance feedback, described in this paper.

the framework regularly updates a secondary content filter which is designed to screen out irrelevant documents that pass through the initial filter. This is a feature-rich analysis that is customized to the search task. First, each document is converted into a wide range of feature representations by applying methods from modality-specific toolboxes. In this work, we use a text feature toolbox and an image feature toolbox with components described below. Then each feature space is subjected to a reduction phase, in which the dimensionality of the feature representation may be reduced significantly (via an unsupervised learning step) or removed entirely (via a supervised down-selection step). This yields a more compact set of representations for the final step: supervised classification. In this phase, a joint classifier across all representations is designed based on training data derived from the user feedback. The scores generated by this classifier are used to rank the search results by relevance.

In the following subsections, we give more detail about the techniques applied for feature extraction, data reduction, and classification in our implementation of the retrieval framework.

A. *Feature Toolbox*

To enable feature extraction from the text component of a document, we start by applying the Stanford CoreNLP toolbox [7] to tokenize and lemmatize the text data. After pre-processing, we extract the following set of features:

- Term Frequency Histogram: This is a simple frequency count of individual tokens (terms) within a document.
- Topic Distribution: We apply a Latent Dirichlet Allocation (LDA) topic modeler [8] to extract a corpus-specific model of 100 topics. The feature vector extracted from each document contains the relative mixture of inferred topics within that document.
- N-Gram Frequency Histogram: We count any instances of the 500 most frequently occurring word-level bi-grams and tri-grams in the corpus, representing common two-word and three-word phrases.
- Parts-of-Speech Histogram: We use the Stanford CoreNLP library to assign a part-of-speech tag to each word, then tabulate a histogram over parts-of-speech occurrences.
- Document Length: Length is determined by number of tokens in a document, which can be indicative of relevance in some microblog analysis tasks.

For image content feature extraction, we use a mixture of low-level and mid-level features which have been proposed for various image analysis tasks. This includes:

- Intensity and Color Histograms: We compute histograms over grayscale intensity, individual RGB color channels, and the CIELAB color space (which is designed to have more perceptual uniformity).
- Bag-of-Words SIFT: We compute a bag-of-words representation of local SIFT features similar to that proposed by [4], using a 400-element dictionary based on clustered SIFT keypoint vectors.
- Border/Interior Classification (BIC): BIC [9] is a compact feature that has shown promising results for content-based information retrieval tasks.
- Dense SIFT: We extract dense SIFT features from a size-normalized grayscale version of the original image.
- Histogram of Oriented Gradients (HOG): We extract HOG features [10] from a size-normalized image, which is a useful way to characterize edge content.
- Gist: The Gist feature is a well-known global descriptor of the spatial structure of a scene [11].
- Number of Faces: Along with low-level feature descriptors we include mid-level object descriptors that may be useful for content-based retrieval, such as the number of faces in a scene (detected using the model in [12]).
- Pedestrian Confidence Score: We also measure the confidence output of a pedestrian detector [13] pre-trained on the INRIA pedestrian dataset.

Collectively, the feature representations in our toolbox give us a broad set of characterizations of everything from word, term and topic frequency within text to color, edge, keypoint, scene, and object-level descriptors within imagery.

978-1-5090-0770-7/16/$31.00 ©2016 IEEE

B. Feature Reduction and Classification

The feedback-driven analysis approach outlined in Fig. 1 treats document retrieval as a binary classification problem (relevant vs. irrelevant) within a joint feature space. One of the challenges of this approach is to do effective classifier construction on high-dimensional feature sets with very limited training data, since we want the framework to be responsive to relatively small amounts of feedback. In our implementation of this framework we experimented with both unsupervised and supervised feature space reduction methods, in addition to different classification strategies. We settled on the following approach because of its robust performance across multiple data sources.

We start by applying Principal Components Analysis (PCA) to the highest-dimension image feature spaces, which contain the most statistical redundancy across individual feature values. PCA yields a transformation to a lower-dimension feature space (with a maximum basis size of 50 in our case) while eliminating much of the redundancy.

During the feedback process, the relevance model is updated via a two-stage supervised training step. First we perform a down-selection that eliminates component feature spaces based on an analysis of their discrimination value (which may vary greatly depending on the retrieval task). Then we form a new classifier jointly over the remaining feature spaces, using the cumulative feedback as labeled training data. We use a linear Support Vector Machine (SVM) classifier, which searches for the hyperplane in the joint feature space that best separates relevant from irrelevant samples. Feature space down-selection is handled by searching for the subset of component feature spaces that yields the best cross-validated SVM performance. We take a greedy approach to finding this subset, iteratively eliminating the most poorly performing individual feature spaces as long as the down-selection improves the performance of the SVM on the validation set. The joint classifier resulting from this greedy search determines the relevance model.

III. Microblog Document Retrieval Experiments

Our first set of experiments applies the proposed retrieval framework to short messages from the popular microblogging service *Twitter*. The experiments use messages taken from a database of 1.5 million geo-tagged tweets collected between April and July 2009, and focus only on text analysis (no imagery).

We define five hypothetical retrieval tasks and construct initial keyword filters to collect potentially relevant messages. The first dataset is focused on a major electrical power disruption scenario, and is created using a compound query of either of the keywords *power* or *electricity* combined with any of the keywords *no*, *out*, *lost*, or *gone*. In this case, relevant tweets are defined as those that contain first-hand descriptions of power grid outages. Irrelevant tweets, on the other hand, refer to news stories, device power outages, song lyrics, hypothetical power outages, and other unrelated contextual uses of the keywords.

The rest of the datasets relate to infectious disease monitoring and are created by aggregating all tweets that contain one of four specific keywords: *fever*, *headache*, *chills*, or *nauseous*. In these cases, we define a relevant tweet as any description of first-hand experience with the corresponding symptom. Non-relevant tweets include references to symptomatic causes that would rule out infectious disease (e.g., allergies), among other uses.

For all five datasets, we label a random subset of tweets as either relevant or irrelevant using the criteria explained above. Table I gives a few examples of labeled documents from each.

A. Feedback Simulation Experiments

We use our labeled data to simulate the user feedback process and evaluate the accuracy of the retrieval results. To run the feedback experiment on one of our labeled sets, we start by randomly ordering the labeled tweets—representing the result of the initial user query—then repeating the following steps:

- Take the top 10 records from the ordered list, and count them as either positives (relevant records) or negatives (irrelevant).
- Add these 10 records and their corresponding labels to the training set. This represents an iteration of the feedback process.
- Use the augmented training set to re-run the feature space down-selection and classifier training steps.
- Apply the new classifier to score the relevance of each remaining record in the labeled dataset.
- Sort the retrieved records in order of descending score, and repeat the process.

The experiment finishes when all labeled records have been pulled from the top of the list and counted as either positive or negative retrievals. An effective search process should encounter the majority of positive-labeled records before negative-labeled records during this sweep, indicating that relevant data is returned to the analyst first.

We note that the structure of these feedback simulation experiments is different than those of prior *ad hoc* Twitter retrieval evaluation (such as the TREC Microblog tracks), in which relevance is judged after retrieval as a way to measure performance and not continually fed back to the retrieval algorithm as relevance feedback.

We use several performance metrics to characterize effectiveness. First, we denote the cumulative number of positive labels encountered at iteration t as $p(t)$ and the cumulative number of negative labels as $n(t)$. Aside from standard precision and recall measures, this allows us to plot the tradeoff between cumulative positive recall

$$R_p(t) = \frac{p(t)}{P}, \tag{1}$$

and cumulative negative recall

$$R_n(t) = \frac{n(t)}{N}, \tag{2}$$

where P and N are equal to the total number of known positive samples and negative samples, respectively. The cumulative recall curve makes performance comparisons easy to

978-1-5090-0770-7/16/$31.00 ©2016 IEEE

TABLE I
EXAMPLES FROM LABELED TWITTER DATASETS

Query	Example Relevant	Example Irrelevant
Power Out	Power gone off again.... And again... In middle of fifa aswell!!! Arghhh	Now my car has no power steering and the breaks dont work. #cool
Chills	Not good. Fever not subsiding. Body sweating a lot. Chills for 3 hours from 4-7 for the past 2 days. Terrible headache and body lain.	Still getting chills rewatching @C_Gomez27s catch last night #myboy
Fever	Nothing like having a fever on the first day of my business law course.	Once again I have puppy fever like crazy!
Headache	Update: I still have a throbbing headache	Keeping my hair up in a bun is giving me a headache
Nauseous	I woke up with a fever and had the chills so I wore boots now it's so hot and I'm nauseous and I have a final in an hour #HotMess	This hot dog eating contest is making me nauseous

TABLE II
TWITTER QUERIES: PERFORMANCE (AUC) PER FEATURE SPACE (LDA = LATENT DIRICHLET ALLOCATION TOPICS, TF = TERM FREQUENCY, POS = PARTS OF SPEECH)

	LDA	$n$gram	TF	PoS	Length	All Features
Power Out	0.73	**0.82**	0.76	0.72	0.68	0.81
Chills	0.66	0.78	**0.79**	0.78	0.66	**0.82**
Fever	0.75	**0.89**	0.85	0.79	0.51	**0.90**
Headache	**0.76**	0.75	0.74	0.76	0.69	**0.82**
Nauseous	0.73	0.79	**0.79**	0.80	0.76	**0.84**
MAUC	0.73	**0.81**	0.79	0.77	0.66	**0.84**
MAP	0.54	**0.61**	0.59	0.58	0.52	**0.65**

interpret because it is invariant to the relative mix of positive and negative labels in the dataset.

B. Results

We use the simulated feedback experiments to investigate the value of individual feature spaces included in the feature-rich approach. For each data retrieval task, we generate the recall trade-off curve and measure the area under the curve (AUC) as a single metric of performance. This value may range from 0.5 (random performance) to 1 (perfect discrimination throughout the retrieval process). Table II lists the measured AUC values for each analysis task and for each feature space. It also highlights both the best overall performer and the best performing individual feature space for each task. The second-to-last row shows the mean AUC across all five tasks. Finally, we also include the mean average precision (MAP) scores in the last row of the table, since this is another common metric to characterize performance. The ensemble of feature spaces used by the proposed method leads to the best discrimination in aggregate (as measured by both mean AUC and mean AP) and in 4 out of the 5 individual tasks.

We also note that the term frequency (TF) approach included in Table II is a representative baseline, since the classifier will learn which keywords to emphasize or de-emphasize by weighting their contribution to the classification score. The feature-rich approach shows a boost in performance over this baseline across all tasks, indicating that the other modes of content characterization contain additional discriminative information not captured by term frequency alone.

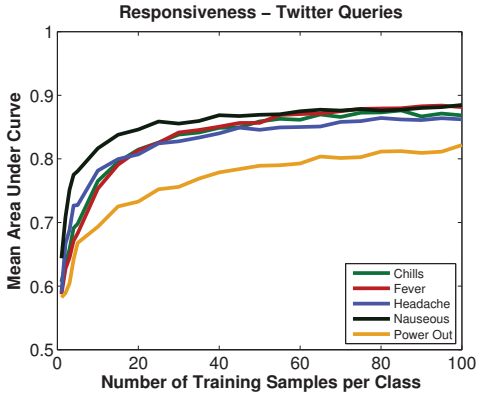

Fig. 2. Performance on Twitter retrieval tasks as a function of positive and negative feedback quantity.

C. Feedback Responsiveness

In addition to investigating the comparative performance of filtering techniques, we also wish to measure the responsiveness of such techniques as a function of user-provided feedback. Responsiveness is important for several reasons. First, a responsive filtering capability will achieve acceptable performance in operational persistent monitoring scenarios more quickly. Second, the users providing feedback are more likely to continue providing input to the system if they notice improvements in the quality of filtered results because of this input.

To measure responsiveness, we compute retrieval performance (measured as area under the recall curve) as a function of the quantity of feedback. To do so for a given search task and feedback level, we provide M positive training samples and M negative samples from the labeled data as simulated feedback. Then we hold this feedback constant as we generate the positive vs. negative recall curve for the remaining labeled data. We repeat this experiment many times, taking a new random sample of M feedback points for each trial. We average the recall curve over these trials and compute the AUC to derive a single metric. We can then plot AUC as a function of M to generate a response curve.

Fig. 2 plots the response curves corresponding to all five of the Twitter queries, for $M = 1$ to 100. We see the

978-1-5090-0770-7/16/$31.00 ©2016 IEEE

TABLE III
IMAGECLEF QUERIES: PERFORMANCE (AUC) PER FEATURE SPACE
(LDA = LATENT DIRICHLET ALLOCATION TOPICS, NG = n-GRAMS, TF = TERM FREQUENCIES, POS = PARTS OF SPEECH, INT = INTENSITY HISTOGRAM, RGB = RED-GREEN-BLUE HISTOGRAM, CL = CIELAB HISTOGRAM, HOG = HISTOGRAM OF ORIENTED GRADIENTS, DS = DENSE SIFT, BWS = BAG-OF-WORDS SIFT, F = FACES, G = GIST, BIC = BORDER/INTERIOR COLOR, PED = PEDESTRIAN SCORE, ATF = ALL TEXT FEATURES, AIF = ALL IMAGE FEATURES, ALL = ALL FEATURES)

	LDA	NG	TF	PoS	Int	RGB	CL	HOG	DS	BWS	F	G	BIC	Ped	ATF	AIF	All
Male Portrait	0.73	0.73	0.78	0.70	0.72	0.73	0.73	**0.78**	0.78	0.72	0.70	0.74	0.73	0.52	0.77	0.82	**0.85**
Chinese Char.	0.58	0.58	0.68	0.56	0.59	0.59	0.61	0.65	0.65	**0.71**	0.56	0.67	0.61	0.51	0.67	0.70	**0.74**
Map of Europe	0.71	0.70	0.74	0.59	0.60	0.61	0.61	**0.86**	0.83	0.79	0.52	0.84	0.62	0.57	0.76	0.88	**0.89**
Water Fountain	0.69	0.73	**0.79**	0.62	0.42	0.59	0.46	0.64	0.62	0.65	0.48	0.67	0.64	0.47	0.78	0.67	**0.80**
Dinosaur Skel.	0.61	0.63	0.66	0.53	0.59	0.63	0.63	0.57	0.58	0.61	0.50	**0.67**	0.59	0.55	0.65	0.68	**0.76**
Roller Coaster	**0.84**	0.82	0.84	0.70	0.60	0.56	0.62	0.62	0.64	0.76	0.52	0.71	0.61	0.52	0.83	0.75	**0.86**
Flying Bird	0.67	**0.72**	0.68	0.66	0.64	0.57	0.62	0.66	0.64	0.71	0.47	0.68	0.59	0.50	0.71	0.73	**0.79**
Skel. Drawings	0.66	0.66	**0.73**	0.69	0.63	0.62	0.67	0.69	0.69	0.70	0.52	0.70	0.72	0.42	0.71	0.77	**0.80**
Shake Hands	0.74	0.78	**0.79**	0.68	0.59	0.57	0.63	0.73	0.73	0.68	0.68	0.73	0.56	0.60	0.79	0.77	**0.80**
Yellow Flames	0.64	0.67	**0.76**	0.57	0.67	0.64	0.65	0.64	0.67	0.69	0.53	0.66	0.73	0.47	0.75	0.75	**0.81**
Satel. Desert	0.55	0.58	0.67	0.60	0.61	**0.71**	0.62	0.65	0.68	0.69	0.61	0.63	0.65	0.60	0.65	0.75	**0.77**
MAUC	0.67	0.69	**0.74**	0.63	0.60	0.62	0.62	0.68	0.68	0.70	0.55	0.70	0.64	0.52	0.73	0.75	**0.81**
MAP	0.18	0.18	**0.21**	0.15	0.15	0.16	0.16	0.20	0.19	0.19	0.13	0.20	0.16	0.12	0.21	0.22	**0.26**

most dramatic gain from $M = 0$ to $M = 20$, then some additional gain from $M = 20$ to $M = 40$, before the curves begin to flatten out. Beyond 40 samples each of positive and negative labels, we see slow but steady improvement up to $M = 100$, especially for the most challenging retrieval case (power outage). This indicates that a significant portion of the performance gain happens early in the feedback process.

IV. MULTIMEDIA DOCUMENT RETRIEVAL EXPERIMENTS

In addition to the microblog document retrieval experiments which use only text data, we also conducted experiments with multimedia documents containing coupled text and imagery. Accurate filtering in this case requires analyzing both text-derived and image-derived features to take advantage of all potential relevance indicators. Therefore, the proposed method leverages all feature extraction modes described in Section II.

Our experiments make use of a publicly available ImageCLEF 2011 [14] dataset containing coupled imagery and text data. More specifically, the dataset is a collection of Wikipedia documents used in the 2010 and 2011 ImageCLEF image retrieval challenges. The entire collection comprises 237,234 Wikipedia images, along with image captions and the original Wikipedia articles that contain the images. It includes text components in English, German and French. For the ImageCLEF 2011 Wikipedia Image Retrieval task, 50 topics were defined for ad-hoc search. For each search task, the results for all participants were pooled together and truth labels were assigned by referees to each document in the pool, indicating whether the document is relevant for the defined topic. We use the provided truth labels as stand-ins for user relevance feedback.

For our experiments, we use English-language documents that contain all three components: image, caption, and article. For the article text, we extract only the paragraphs surrounding the image reference in the original article. We also select the subset of 11 topics that each have at least 50 relevant sample images to support experimentation. The highest number of positive labels for any task is 353.

As before, we simulate the process of relevance feedback for each retrieval task defined by the dataset and compute the area under the normalized recall curve. Table III gives the complete breakdown of discrimination metrics per feature space and retrieval task. It also shows the results for all text features combined, all image features combined, and all text plus image features combined. The results indicate a wide variation in discrimination power of different feature types; some are near random performance, while other features do much better on the same task.

There are several conclusions we can draw from these results. First, the image features seem to provide more collective discrimination power than the text features, yielding higher MAUC and MAP values overall. This is not surprising, since in both sets the relevance of the documents is ultimately a function of image content and the text captions and articles are only helpful insofar as they indicate the nature of the image content. However, the text features do provide an additional performance boost over image features alone and even provide the best results in isolation for some of the retrieval tasks.

Second, there is quite a bit of variation in the value of individual feature spaces depending upon the relevance filtering task. In fact, there are seven different feature spaces that serve as the highest performing individual space for at least one of the retrieval tasks. This supports the intuition underlying the feature-rich approach: that it is difficult to anticipate the most valuable feature spaces without prior knowledge of particular data retrieval needs.

Finally we note that the responsiveness of the image retrieval tasks to feedback is similar to that observed in the text retrieval experiments. The image filtering performance also show the sharpest gains at low feedback levels ($M < 20$); however, their gradients remain higher than those of the Twitter response curves out to $M = 100$. This suggests that the image retrieval tasks require more complex content analysis and therefore see more benefit with increasing levels of feedback.

978-1-5090-0770-7/16/$31.00 ©2016 IEEE

Fig. 3. Screenshots of the prototype tool for feedback-driven data filtering. Left: User interface for text-only microblog content retrieval. Right: User interface for coupled text/imagery content retrieval.

V. Prototype Software

In addition to the performance evaluation of the proposed technique, we also constructed a prototype software tool for customizable social media content filtering based on this technique. Figure 3 shows sample screenshots of the tool.

The tool has a browser-based user interface that allows analysts to select an initial keyword query, browse the resulting records, mark individual records as relevant or irrelevant, and submit this feedback. Once submitted, the tool builds a new content filtering model and uses the model to return a re-ranked list of records, placing data with highest estimated relevance at the top. We have tested the functionality of the tool for browsing both text and imagery. In either case, the interface contains supporting visualizations (such as word clouds and image clusters) to highlight the characteristics of the analyst-identified relevant posts. Initial testing on social media data has provided some qualitative validation of the utility of the tool for representative retrieval tasks.

VI. Conclusion

Adaptivity to feedback is an important aspect of social media content filtering systems, especially when the gain in performance is worth the sustained investment of user feedback. Along with better discrimination, adaptivity allows for customization to very specific information monitoring needs, which may vary widely across analysts within the same organization. Experimentation with the proposed framework for feedback-driven social media document retrieval confirms that feature-rich models of document relevance are in general more accurate than simpler models built in individual feature spaces, and achieve significant boosts in accuracy after only a few tens of submitted feedback points. The proposed method has been incorporated into a proof-of-concept prototype tool, which is capable of real-time content filter refinement over (initially limited) social media data sources.

References

[1] J. Rocchio, "Relevance feedback in information retrieval," in *The Smart Retrieval System: Experiments in Automatic Document Processing*, G. Salton, Ed., 1971.

[2] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Syst.*, vol. 8, no. 6, pp. 536–544, April 1998.

[3] L. Tanguy, A. Urieli, B. Calderone, N. Hathout, and F. Sajous, "A multitude of linguistically-rich features for authorship attribution," in *Notebook for PAN at CLEF*, 2011.

[4] D. Pracner, N. Tomašev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "WIKImage: Correlated image and text datasets," in *Proc. of the 14th International Multiconference on Information Society (IS 2011)*, 2011, pp. 141–144. [Online]. Available: http://perun.dmi.rs/pracner/wikimage/

[5] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Image and Video Retrieval*, ser. Lecture Notes in Computer Science, 2003, vol. 2728, pp. 238–247.

[6] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using biasmap," in *Proc. Comput. Vision and Pattern Recognition*, 2001, pp. I–11–I–17.

[7] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 102–109.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[12] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[13] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *British Machine Vision Conference (BMVC)*, Aberystwyth, UK, 2010.

[14] T. Tsikrika, A. Popescu, and J. Kludas, "Overview of the Wikipedia image retrieval task at ImageCLEF 2011," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011. [Online]. Available: http://www.imageclef.org/2011/Wikipedia

978-1-5090-0770-7/16/$31.00 ©2016 IEEE