CrowdE: Filtering Tweets For Direct Customer Engagements

Jilin Chen, Allen Cypher, Clemens Drews and Jeffrey Nichols
IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA
{jilinc, acypher, cdrews, jwnichols}@us.ibm.com

Abstract

Many consumer brands have customer relationship agents that directly engage opinionated consumers on social streams, such as Twitter. To help agents find opinionated consumers, social stream monitoring tools provide keyword-based filters, which are often too coarse-grained to be effective. In this work, we introduce CrowdE, a Twitter-based filtering system that helps agents find opinionated customers through brand-specific intelligent filters. To minimize per-brand effort in creating these brand-specific filters, the system used a common crowd-enabled process that creates the filters through machine learning over crowd-labeled tweets. We validated the quality of the crowd labels and the performance of the filter algorithms built from the labels. A user evaluation further showed that CrowdE's intelligent filters improved task performance and were generally preferred by users in comparison to keyword-based filters in current social stream monitoring tools.

Introduction

Social stream platforms such as Twitter are attracting many consumers to express opinions about brands, and consumer brands are therefore increasingly motivated to directly engage consumers on these platforms (Jansen et al. 2009). Many brands (e.g. Delta Airlines, BestBuy) have since adopted social stream monitoring tools (e.g. Hootsuite, Seesmic, Radian6) to initiate customer engagements from dedicated Twitter accounts (e.g. @deltaassist, @bestbuy).

To achieve maximal effect, customer relationship agents from the brands constantly monitor live social streams, identify as many opinionated customers as possible, and take appropriate actions for each of them, such as addressing complaints from troubled customers and encouraging loyal customers to spread good words. As social streams are of huge volume, a majority of posts are irrelevant to the brand, and even fewer contain opinions worth responding to, social stream monitoring tools often allow agents to filter posts by specifying keywords (e.g. "delta" for Delta Airlines) so as to reduce the information overload.

For large consumer brands, filtering social streams in such a keyword-based approach could require a huge amount of manual effort. For instance, Dell Computer was reported to have a team of more than 70 agents whose jobs are to track about 25,000 Dell related posts per day across various social streams[1]. A key reason behind the huge manual effort is the ineffectiveness of simple keyword-based filters. For brands with non-unique names such as Delta Airlines, filtering for a single word includes many irrelevant posts (e.g. filtering for "delta" includes posts about "alpha delta phi" and "Nile delta"), while a stricter keyword set easily misses many relevant posts (e.g. requiring both "delta" and "airline" would miss the post "I flew to Seattle on Delta"). Furthermore, even among posts that indeed refer to the brand, many are side comments that contain no brand relevant opinion (Jansen et al. 2009), and therefore are usually not worth an agent's attention.

A natural solution is to create intelligent filters using expert-crafted keyword-based rules (Pollock 1988) or machine learning (Paek et al. 2010). However, in both cases, the filters will need to be brand-specific, and require brand-specific knowledge or ground truth. For instance, while for an airline brand the filter may need the word "flight" to indicate relevance and the phrase "lost luggage" to indicate negative opinion, the filter for a digital camera brand would need different words and phrases like "lens" and "fuzzy picture". In practice, we are aware of no social stream monitoring tool that contains such intelligent filters, presumably because vendors of these tools cannot afford heavy per-brand effort, e.g. researching new algorithms or collecting training data for every possible brand. Instead, these tools used simple keyword-based filters, which is a weaker but more general solution.

In this work we propose a strong and general solution. We present CrowdE, a novel filtering system that helps agents find opinionated customers on Twitter through two brand-specific intelligent filters, one for brand relevance, and one for presence of opinion. To minimize per-brand effort in data collection, algorithm design, and performance tuning, we introduce a common crowd-enabled process that creates the filters through machine learning over crowd-labeled tweets. The intuition is simple: crowd workers can likely label the relevance and the opinion across brands, thus serving as a universal source of ground truth. We validate our overall approach using Delta Airlines and Hertz Rent-a-car as example brands. We use the following four research questions to guide our validation process:

[1] http://tech.fortune.cnn.com/2012/07/23/tweetbots/

Q1) Generalness and Per-Brand Effort: How can we build many brand-specific intelligent filters with minimal per-brand effort? How much effort is needed for system designers to create filters for a new brand?

Q2) Quality of Crowd-Labeled Ground Truth: Can crowd workers reliably label the relevance and the opinion of tweets toward brands? What quality control measures are necessary to ensure quality?

Q3) Effectiveness of Filter Algorithms: How effective are the filter algorithms learned from crowd-labeled ground truth? How do they compare to alternatives?

Q4) Usefulness in Filtering Tasks: Are the intelligent filters more useful for finding engagement targets in comparison to keyword filters in current monitoring tools?

We offer both system and empirical contributions. On the system side, we present a novel crowd-enabled system to support a filtering task of practical importance. On the empirical side, we thoroughly validate the system through various aspects, demonstrate the system's advantage over existing solutions, and derive insights for future crowd sourcing and intelligent filtering systems.

The rest of the paper is structured as follows. First, we discuss prior research that influenced our work. To answer Q1, we introduce the design of the CrowdE system, and the common process for creating brand-specific filters. We then evaluate the crowd-labeled ground truth to answer Q2, evaluate the filter algorithms to answer Q3, and report a user study to answer Q4. In the end, we discuss the practical impact, design implications and future work.

Related Work

CrowdE is motivated by prior research on consumer opinion and customer relationship management on social streams. Jansen et al. (2009) studied mentions of 50 brands across 12 industry sectors on Twitter over 13 weeks, and concluded that the prevalence of brand-related opinion makes Twitter a powerful media for spreading consumer opinion. Hennig-Thurau et al. (2010) discussed the impact of social media from the angle of customer relationship management, arguing the benefits of direct engagements.

As a filtering system, CrowdE draws from decades of intelligent filtering research. While early work in the area explored manually-crafted rules (Pollock 1988), later research focused on learning filters from ground truth of various sources. On social streams, Chen et al. (2011) leveraged users' tweets and social network structure to recommend conversations on Twitter. Bernstein et al. (2010) provided topic-based filtering to Twitter users by generating topics from users' historical tweets and search engine-based term expansion. Paek et al. (2010) collected user ratings on Facebook posts and trained support vector machines to filter away low-rated posts. To CrowdE and other social stream monitoring tools, the core challenge lies in the fact that system designers usually do not have access to the necessary brand-specific ground truth.

Crowd workers from platforms like Amazon Mechanical Turk (MTurk) have often been used to provide ground truth for text data, even though their reliability is often questioned. Snow et al. (2008) used crowd workers for five different word-level labeling tasks, and showed that after using majority voting the collected labels were as accurate as expert labels. Hsueh et al. (2009) used crowd workers to label the sentiment of blog snippets, and found that the quality of labels depends on noisy level, sentiment ambiguity and lexical uncertainty. Diakopoulos et al. (2010) analyzed tweet sentiment using crowd workers and implemented several routines to improve label quality. Paul et al. (2011) used crowd workers to label question tweets. They reported that only a minority of workers provided correct labels, and suggested that their finding may be due to the vagueness of their task instruction. In contrast to these prior efforts that treated crowd workers as ad-hoc data sources for manual research analysis, CrowdE treats crowd workers as an integral part of a common reusable process that helps construct running systems.

As a crowed-enabled system, CrowdE is inspired by Turkit (Little et al. 2010) for introducing crowd-sourcing as algorithm components for system building, and by Quinn et al. (2011) for arguing the combination of crowd-sourcing and machine learning for better efficiency. In contrast to these prior efforts that focus on introducing ideas and techniques for artificial problems, we strive to design and evaluate our system in context of a practical problem.

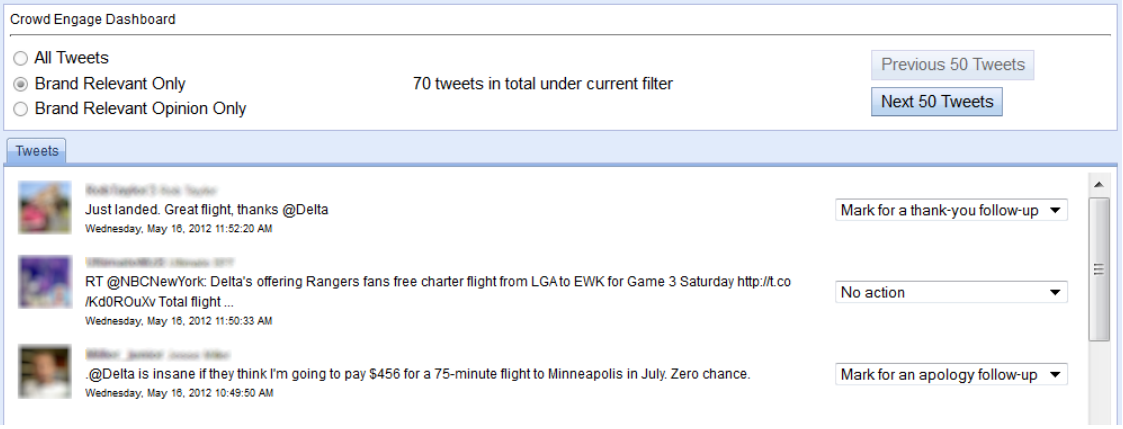
Our work is also related to active learning systems such as that of Brew et al. (2010). The difference is that while active learning systems learn from end users, CrowdE learns from strangers who are completely ignorant of the system that they help to build.

The CrowdE System

Current social stream monitoring tools (e.g. Hootsuite) often allow customer relationship agents to track tweets through keywords. For example, if an agent from Delta Airlines specifies a single keyword "delta", the tool would display all recent tweets mentioning the word "delta" in a list, and update the list periodically as new tweets arriving. In this way, the tool is brand-agnostic, and can be used for agents from all different brands.

However, such keyword-based filters are often overly *inclusive* or *strict*. For finding tweets relevant to Delta Airlines, filtering for all posts including the word "delta" is overly *inclusive*, because it contains not only all tweets about Delta Airlines, but also irrelevant posts that mention fraternities and the Nile Delta. In contrast, requiring both "delta" and "airline" results in an overly *strict* filter, because it misses tweets that refer to Delta Airlines without using the word "airline", e.g. "I flew to Seattle on Delta". Filtering for opinion through keywords can be even more challenging due to different expressions of sentiment and opinion. It is therefore often difficult for agents to create effective keyword filters for their brands.

To address the above problem, we remove from agents the burden of creating effective filters. Instead, before being used for a brand, the CrowdE system first goes



Crowd Engage Dashboard

○ All Tweets
● Brand Relevant Only
○ Brand Relevant Opinion Only

70 tweets in total under current filter

[redacted]
Just landed. Great flight, thanks @Delta
Wednesday, May 16, 2012 11:52:20 AM
Mark for a thank-you follow-up ▾

[redacted]
RT @NBCNewYork: Delta's offering Rangers fans free charter flight from LGA to EWK for Game 3 Saturday http://t.co /Kd0ROuXv Total flight ...
Wednesday, May 16, 2012 11:50:33 AM
No action ▾

[redacted]
.@Delta is insane if they think I'm going to pay $456 for a 75-minute flight to Minneapolis in July. Zero chance.
Wednesday, May 16, 2012 10:49:50 AM
Mark for an apology follow-up ▾

Figure 1. Frontend User Interface of CrowdE.

through a filter creation process, in which we setup a brand-specific inclusive keyword filter (e.g. "delta" for Delta Airlines) and generate two brand-specific intelligent filters on top of the inclusive keyword filter.

The CrowdE User Interface

Figure 1 shows the frontend user interface of CrowdE, using Delta Airlines as the example brand.

At the top panel, users can select between three intelligent filter options. By default, the option *All Tweets* is selected, and the system displays at the bottom all tweets that satisfy a preset inclusive keyword filter (e.g. tweets mentioning the word "delta"). If the user selects *Brand Relevant Only*, the system will only display tweets that pass an intelligent filter for brand relevance (e.g. tweets about Delta Airlines). If the user selects *Brand Relevant Opinion Only*, the system will only display tweets that pass both the brand relevance filter and another filter for the presence of brand-related opinion (e.g. tweets about Delta Airlines and having related opinion). Whenever the filter option is changed, the system shows how many tweets are displayed under the current filter. For instance, in Figure 1, the system indicates that with the brand relevance filter in effect the system is showing 70 tweets in the list below.

The list of tweets at the bottom are displayed in reverse chronological order. The list is static here for our user study; in practice, agents can either manually refresh the list to get all tweets posted since the last refresh, or set up auto-refreshes, e.g. refresh the list every 10 minutes to get tweets posted in the last 10 minutes. The display order of tweets and the refresh mechanism are designed following the search interface on the Twitter website and the norm of current social stream monitoring tools.

For each tweet, the system displays its content, author, the time that the tweet was posted, and possible actions that agents can take. By default, all tweets are marked with no action, and agents can mark tweets with either thank-you follow-ups or apology follow-ups as necessary. In practice, tweets that are marked with actions will later be routed to other agents in marketing or customer support departments for making the actual engagements.

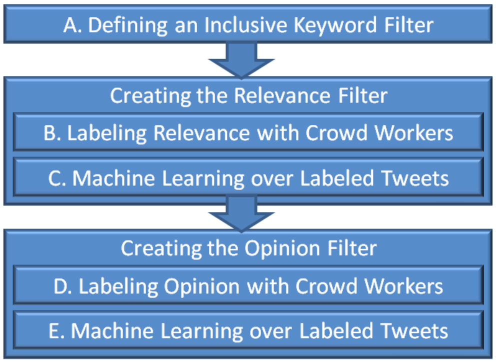

A. Defining an Inclusive Keyword Filter
Creating the Relevance Filter
B. Labeling Relevance with Crowd Workers
C. Machine Learning over Labeled Tweets
Creating the Opinion Filter
D. Labeling Opinion with Crowd Workers
E. Machine Learning over Labeled Tweets

Figure 2. Stages for Creating Brand-Specific Filters.

Creating Intelligent Filters with the Crowd

The CrowdE frontend needs two brand-specific intelligent filters, one for brand relevance, and one for presence of opinion. To build these brand-specific filters with minimal brand-specific effort in data collection, algorithm design, and performance tuning, the system used a common crowd-enabled process that creates the filters through machine learning over crowd-labeled tweets.

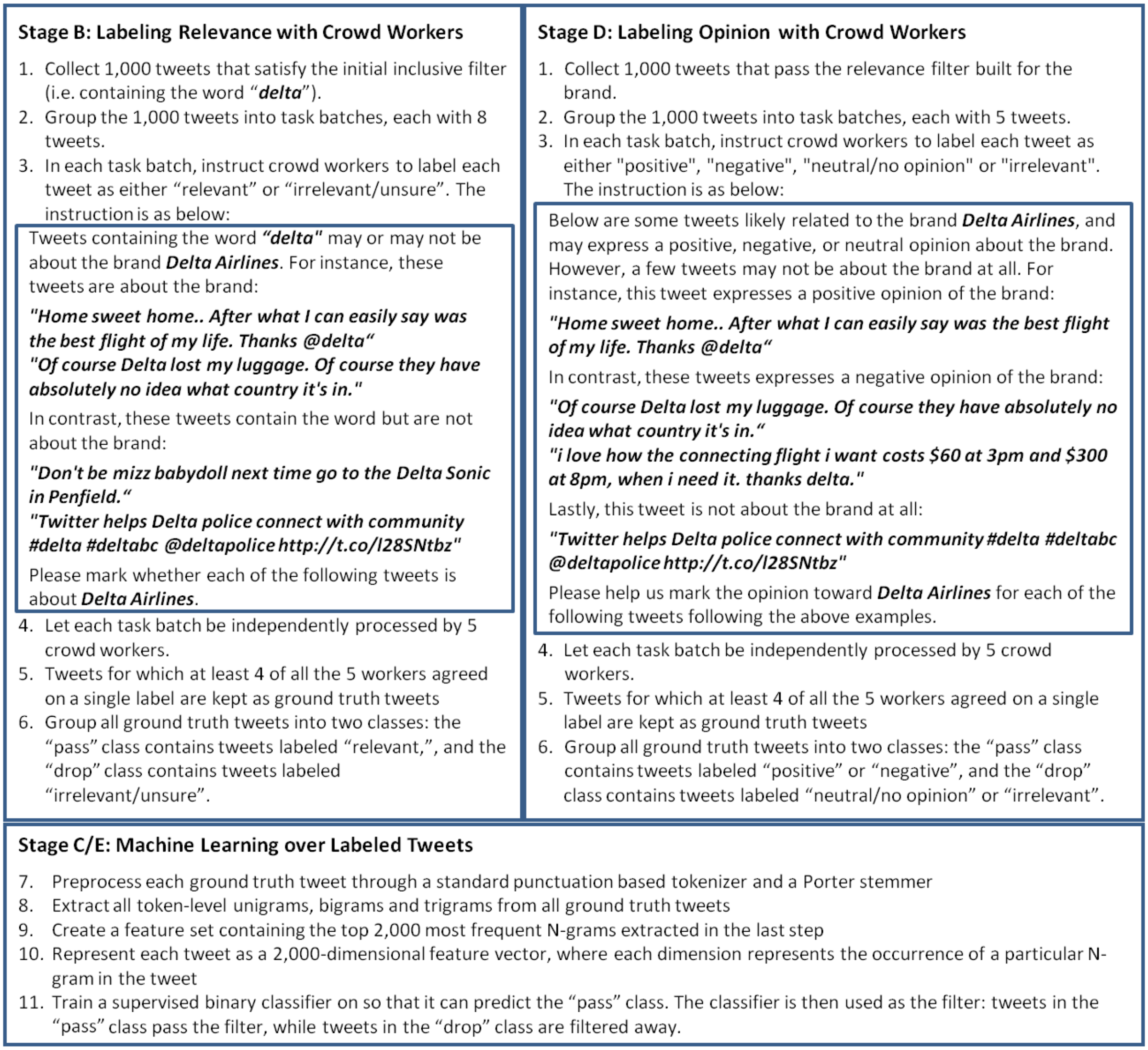
Stage A: Defining an Inclusive Keyword Filter

Figure 2 presents the whole filter creation process in five stages. In Stage A we set up an inclusive keyword filter for the brand, e.g. using the word "delta" to track tweets that are relevant to Delta Airlines. More generally, the inclusive keyword filter may need one or more keywords depending on the brand. For instance, for Hertz Rent-a-car, we used the word "hertz" to track all tweets relevant to the brand, just like the case of Delta. In contrast, for tracking tweets about Apple Inc., we used several keywords including "apple", "mac", "ipod" "iphone", and "ipad". Overall we found that such inclusive filters were easy to setup for most brands we considered, and had little impact on the effectiveness of the two intelligent filters we build later.

In the rest of this section we introduce Stage B, C, D, E using Delta Airlines as the example brand (Figure 3).

Stage B: Labeling Relevance with Crowd Workers

In Stage B we let crowd workers label tweets for building the relevance filter. The filter detects whether a tweet that satisfies the inclusive keyword filter is relevant to the



Stage B: Labeling Relevance with Crowd Workers

1. Collect 1,000 tweets that satisfy the initial inclusive filter (i.e. containing the word "***delta***").
2. Group the 1,000 tweets into task batches, each with 8 tweets.
3. In each task batch, instruct crowd workers to label each tweet as either "relevant" or "irrelevant/unsure". The instruction is as below:

> Tweets containing the word ***"delta"*** may or may not be about the brand ***Delta Airlines***. For instance, these tweets are about the brand:
>
> ***"Home sweet home.. After what I can easily say was the best flight of my life. Thanks @delta"***
> ***"Of course Delta lost my luggage. Of course they have absolutely no idea what country it's in."***
>
> In contrast, these tweets contain the word but are not about the brand:
>
> ***"Don't be mizz babydoll next time go to the Delta Sonic in Penfield."***
> ***"Twitter helps Delta police connect with community #delta #deltabc @deltapolice http://t.co/l28SNtbz"***
>
> Please mark whether each of the following tweets is about ***Delta Airlines***.

4. Let each task batch be independently processed by 5 crowd workers.
5. Tweets for which at least 4 of all the 5 workers agreed on a single label are kept as ground truth tweets
6. Group all ground truth tweets into two classes: the "pass" class contains tweets labeled "relevant,", and the "drop" class contains tweets labeled "irrelevant/unsure".

Stage D: Labeling Opinion with Crowd Workers

1. Collect 1,000 tweets that pass the relevance filter built for the brand.
2. Group the 1,000 tweets into task batches, each with 5 tweets.
3. In each task batch, instruct crowd workers to label each tweet as either "positive", "negative", "neutral/no opinion" or "irrelevant". The instruction is as below:

> Below are some tweets likely related to the brand ***Delta Airlines***, and may express a positive, negative, or neutral opinion about the brand. However, a few tweets may not be about the brand at all. For instance, this tweet expresses a positive opinion of the brand:
>
> ***"Home sweet home.. After what I can easily say was the best flight of my life. Thanks @delta"***
>
> In contrast, these tweets expresses a negative opinion of the brand:
>
> ***"Of course Delta lost my luggage. Of course they have absolutely no idea what country it's in."***
> ***"i love how the connecting flight i want costs $60 at 3pm and $300 at 8pm, when i need it. thanks delta."***
>
> Lastly, this tweet is not about the brand at all:
>
> ***"Twitter helps Delta police connect with community #delta #deltabc @deltapolice http://t.co/l28SNtbz"***
>
> Please help us mark the opinion toward ***Delta Airlines*** for each of the following tweets following the above examples.

4. Let each task batch be independently processed by 5 crowd workers.
5. Tweets for which at least 4 of all the 5 workers agreed on a single label are kept as ground truth tweets
6. Group all ground truth tweets into two classes: the "pass" class contains tweets labeled "positive" or "negative", and the "drop" class contains tweets labeled "neutral/no opinion" or "irrelevant".

Stage C/E: Machine Learning over Labeled Tweets

7. Preprocess each ground truth tweet through a standard punctuation based tokenizer and a Porter stemmer
8. Extract all token-level unigrams, bigrams and trigrams from all ground truth tweets
9. Create a feature set containing the top 2,000 most frequent N-grams extracted in the last step
10. Represent each tweet as a 2,000-dimensional feature vector, where each dimension represents the occurrence of a particular N-gram in the tweet
11. Train a supervised binary classifier on so that it can predict the "pass" class. The classifier is then used as the filter: tweets in the "pass" class pass the filter, while tweets in the "drop" class are filtered away.

Figure 3. Detailed Steps for Creating Filters Specific to Delta Airlines.

Details specific to Delta Airlines, i.e. the initial keyword filter and the example tweets in instructions, were shown in bold Italic. These were the only details that need to be changed to adapt the whole process for another brand, such as Hertz Rent-a-car and Apple Inc.

brand, e.g. whether a tweet containing the word "delta" is indeed relevant to Delta Airlines.

We collected tweets using Twitter Streaming API, and recruited crowd workers from Amazon Mechanical Turk (MTurk). The MTurk transactions were automated through the Turkit toolkit (Little et al. 2010).

We group tweets into task batches to increase efficiency. Each task batch contained 8 tweets, took 15~20 seconds to finish, and had a compensation of US$0.03, equivalent to ~US$6 per hour.

We took the majority label among 5 crowd workers, and removed ambiguous tweets on which workers disagree (i.e. 3vs2), as both measures can improve label quality for machine learning purposes (Snow et al. 2008, Hsueh et al. 2009).

Under this setting, all tweets for a brand could usually be labeled within a few hours, with a total cost of ~US$20. For brands we have tested, usually 20%~25% of the tweets were deemed ambiguous and thus removed.

Stage D: Labeling Opinion with Crowd Workers

In Stage D we let crowd workers label tweets for building the opinion filter. The filter detects whether a tweet satisfying the brand relevance filter contains any positive or negative opinion about the brand, e.g. whether a Delta Airlines related tweet contained any opinion about Delta Airlines. Steps in Stage D are similar to those in Stage B, except for two changes:

Firstly, while in Stage B we labeled tweets satisfying the inclusive keyword filter (e.g. "delta"), here we label tweets satisfying the relevance filter. This change was to avoid wasting workers' time on many tweets irrelevant to the brand, and to increase the ratio of opinionated tweets in the tweet collection for the benefit of machine learning. As the relevance filter we relied on may make mistakes, we have included "irrelevant" as a possible label here as well.

Secondly, as labeling opinion is more time consuming than labeling relevance, for each task batch we reduced the number of tweets to 5, and increased the compensation to US$0.05. Each task batch took 20~25 seconds to finish, making the equivalent wage to ~US$6 per hour again.

All tweets for a brand could usually be labeled within a few hours, with a total cost of ~US$50. For brands we have tested, usually 20%~25% of the tweets were deemed ambiguous and thus removed.

Stage C/E:Machine Learning over Labeled Tweets

Stage C and E share exactly the same machine learning process, the only difference being the ground truth used.

We used the 2,000 most frequent N-grams as the feature set. We experimented with the feature set size ranging from 1,000 to 5,000, and found no substantial difference in our results. We picked 2,000, because larger numbers showed no improvement in classification performance but required substantially longer training time for the machine learning classifiers.

We have also attempted representing tweets using LDA, but obtained inferior performance, consistent with prior results on Twitter (Bernstein et al. 2010, Chen et al. 2011). Fully exploring different topic models is beyond the scope of this work.

We experimented with various supervised classifiers from the Weka machine learning package (Witten et al. 2011), including logistic regression, naive Bayesian, several support vector machines and several decision tree-based classifiers. We chose a decision tree classifier (REPTree) for relevance and a support vector machine classifier (SMO) for opinion, as the two classifiers consistently outperformed other Weka classifiers in their respective classification tasks.

Adapting the Process for Other Brands

The only brand-specific details in the process are the keyword filter in Stage A (e.g. "delta" for Delta Airlines), the example tweets in Stage B (2 relevant and 2 irrelevant), and the example tweets in Stage D (1 positive, 2 negative, and 1 irrelevant). A system designer/vendor or an agent can obtain these details for a new brand within 15 minutes by using the search function on the Twitter website.

After obtaining these brand-specific details, the CrowdE system can be built automatically without any extra effort from system designers: the CrowdE frontend is general across brands; the process for creating intelligent filters is automated using Turkit and Weka; the setup and parameters in the process, such as the number of tweets to be collected and the size of feature set for machine learning, are robust and can be used across brands. The only cost in this process is the ~US$70 for the crowd workers.

We have therefore answered Q1. That is, by following the crowd-enabled process in Figure 2 and 3, we can create intelligent filters for customer engagement with minimal per-brand effort.

Tweet Collection	Agreement Between the Three Authors	Agreement Between the Three Authors and the Crowd Labels
Delta Relevance	.93	.90
Delta Opinion	.76	.77
Hertz Relevance	.84	.83
Hertz Opinion	.75	.77

Table 1. Crowd Labels vs. Authors' Labels.
Agreement was measured as the mean pair-wise Cohen's kappa.

Evaluating Crowd-Labeled Ground Truth

In the rest of this work, we evaluate the system built from the crowd-enabled process, and answer Q2, Q3, and Q4. We use Delta Airlines and Hertz Rent-a-car as example brands, because the two brands are mentioned frequently on Twitter for their services, and because their brand names are difficult to be filtered with simple keyword filters alone (i.e. "delta" and "hertz").

We first answer Q2 by evaluating the quality of ground truth tweets across the two example brands. Following the steps in Figure 3, we collected labels for two collections of tweets per brand, one for relevance and one for opinion. From each collection we randomly selected 100 tweets, and had each of the tweets labeled independently by three authors of the paper.

As shown in Table 1, for all four tweet collections, the mean pair-wise Cohen's kappa among the three authors was comparable to the mean pair-wise Cohen's kappa between the three authors and the crowd labels. In other words, labels obtained through steps in Figure 3 were of comparable quality to labels from the authors.

The process we used to collect the ground truth was the simplest that we could design that ensures the above label quality. Omitting any key component from the process (e.g. the example tweets, the use of majority label from multiple workers, or the removal of ambiguous tweets) would have noticeably reduced label quality.

The process is otherwise robust. We varied the example tweets used in the instructions, and found no substantial change in the resulting labels, as long as the tweets fit to the instruction (e.g. the relevant tweet example is indeed relevant to the brand). We also attempted reducing the compensation by half to ~US$3 per hour. In this case, the quality of the resulting labels remained the same, but the labeling process took a few days, much longer than the time taken under our original compensation level.

Evaluating Filter Algorithms

We answer Q3 by conducting algorithm evaluations on the intelligent filters for the two example brands, Delta and Hertz. The evaluation metrics are precision, recall and F1-score on the ground truth tweets. In our context, high precision means that most tweets that pass the filter are indeed relevant (or opinionated), high recall means that most relevant (or opinionated) tweets can indeed pass the filter. F1-score is the harmonic mean of precision and recall.

	Delta Airlines			Hertz Rent-a-car		
Relevance Filter	Precision	Recall	F1	Precision	Recall	F1
Original Keyword	9.3%	100%	17.0%	38.2%	100%	55.3%
Additional Keyword	96.3%	22.2%	36.1%	94.0%	56.1%	70.3%
Relevance Score	47.6%	51.3%	49.4%	72.9%	73.1%	73.0%
Expert Created Rules	83.6%	92.0%	87.6%	87.7%	87.7%	87.7%
Crowd + REPTree	93.5%	72.5%	81.7%	91.0%	80.4%	85.4%
Opinion Filter	Precision	Recall	F1	Precision	Recall	F1
No Filter	54.0%	100%	70.1%	41.5%	100%	58.7%
Sentiment Score	75.0%	74.9%	74.9%	64.8%	65.5%	65.1%
Crowd + SMO	83.4%	84.2%	84.8%	83.3%	81.0%	82.1%

Table 2. Evaluation of Filter Algorithms.

We compare against simple keyword filters to represent the state-of-art in current social stream monitoring tools, and against a few practical alternatives that can be easily implemented by software vendors without brand-specific ground truth. We do not include any advanced supervised algorithms for comparison, as our goal is not to discover the best possible supervised algorithm, but to understand the benefit of having brand-specific ground truth available.

Besides crowd workers, we have also considered semi-supervised learning as a possibility for compensating the lack of brand-specific ground truth. However, we are unaware of any prior work that applies semi-supervised learning on our filtering problem, or any prior algorithm that could create an effective baseline from the mere four example tweets that we gave to crowd workers in Figure 3. We therefore omit semi-supervised methods in this work and leave them for future exploration.

Evaluating the Relevance Filters

We consider the following methods for comparison:

Original Keyword: A baseline where no filter is used beyond the initial keyword filter. The filter includes all tweets containing the word "delta" for Delta, and includes all tweets containing the word "hertz" for Hertz. This baseline is to represent the case that an agent uses a simple inclusive keyword filter in a social stream monitoring tool.

Additional Keyword: A baseline where we require an additional keyword. For Delta we require each tweet to contain the words "delta" and "airline", and for Hertz we require each tweet to contain the words "hertz" and "car". This baseline is to represent the case of using an overly strict keyword filter in a social stream monitoring tool.

Relevance Score: A baseline that uses text similarity to bootstrap from a few brand-relevant words. We illustrate its procedure using Delta Airlines: We first collected about 40,000 tweets that contains the word "delta". From the collection we created a reference tweet by concatenating all tweets that contain both "delta" and "airline". From the collection we also extracted all stemmed words that appear at least twice, and removed stop words using a standard English stop word list. We can then represent each tweet as a bag-of-words vector, where each dimension of the vector represents the tweet's TF*IDF score on a particular word (Salton et al. 1988). The relevance score of an arbitrary tweet is then calculated as the cosine similarity between the bag-of-words vector of the tweet and the bag-of-words vector of the reference tweet. This score represents the content similarity between the tweet in question and a subset of tweets that were highly likely relevant to Delta Airlines, and a higher score may therefore suggest higher relevance to Delta Airlines. For filtering we ranked candidate tweets according to this score, included the top 9% for Delta Airlines, and included the top 38% for Hertz Rent-a-car. We chose these cut-off numbers from the actual ratio of relevant tweets in the ground truth tweets; in practice it would be determined through trial-and-error. We include this baseline as a practical alternative, because it requires no additional ground truth, has minimal brand-specific details, and can be easily implemented by vendors.

Expert Created Rules: A baseline using manually crafted rules. One author manually created 73 regular expression rules for Delta Airlines and 90 regular expression rules for Hertz Rent-a-car by looking at 75% of the ground truth tweets. The rules were then tested on the remaining 25% of the ground truth tweets. We include this baseline as it is perhaps the most straightforward way for implementing a one-time solution for a brand. It is nonetheless quite costly: for each brand, while our crowd-enabled method requires merely 15 minutes for setup and ~US$70 for the crowd workers, creating the rules requires hours of an expert's time, which is usually much more expensive.

Crowd + REPTree: Our crowd-enabled method. We used the REPTree classifier from Weka, as it consistently outperformed other Weka classifiers across brands. Evaluation metrics were computed using 10-fold cross validation on the ground truth tweets.

The top half of Table 2 shows the performance of all five methods. The original keyword method had low precision (i.e. included many irrelevant tweets) while the additional keyword method had low recall (i.e. missed many relevant tweets), showing the deficiency of simple keyword filters in current social stream monitoring tools. The relevance score method was balanced but substantially inferior to the crowd-enabled method, demonstrating the deficiency of bootstrapping from small amount of brand-relevant information. The expert created rules performed the best but was costly to create, requiring hours of expert effort per brand. In contrast, the crowd-enabled method was close to the top in terms of performance and can be adapted for various brands with minimal effort.

Evaluating the Opinion Filters

We consider the following methods for comparison:

No Filter: A baseline where no additional filter is used. This baseline includes all tweets that pass the relevance filter, and represent the case that no opinion-based filter is available.

Sentiment Score: A baseline that counts sentiment words. This method is based on SentiWordNet (Baccianella et al. 2010), a lexical thesaurus that scores words on positivity and negativity. Following Baccianella et al. (2010), for each tweet we computed the positivity score and the negativity score for all the words that can be mapped onto SentiWordNet, and assigned zero positivity and negativity to all other words in the tweet. The sentiment score of an arbitrary tweet is then calculated as the sum of positivity scores and negativity scores of all its words, divided by the total number of words in the tweet. Intuitively, this score attempts to represent the subjectivity of each tweet through its word usage, and therefore may suggest the presence of opinion in tweets. For filtering we ranked candidate tweets according to this score, included the top 54% for Delta Airlines, and included the top 42% for Hertz Rent-a-car. We chose these cut-off numbers from the actual ratio of opinionated tweets in all brand relevant tweets; in practice it would be determined through trial-and-error. We include this baseline, because it requires no additional ground truth, and because counting known sentiment words has been a prevalent way for tracking opinionated tweets in prior work (e.g. O'Connor et al. 2010).

Crowd + SMO: Our crowd-enabled method. We used the SMO classifier from Weka, as it consistently outperformed other Weka classifiers. Evaluation metrics were computed using 10-fold cross validation on the ground truth tweets.

The bottom half of Table 2 shows the performance of all three methods. The no filter method had low precision (i.e. included many non-opinion tweets). Between the rest two, the crowd-enabled method was better than the sentiment score method. Consistent with prior research on opinion mining (Pang et al. 2008), this result shows the superiority of domain-specific training data for opinion mining.

Overall, we could answer Q3 by stating that the crowd-enabled filter algorithms were effective, and were superior to a number of practical alternatives due to the availability of brand-specific ground truth.

User Evaluation

The algorithm comparison in the previous section does not take into the account the interactive nature of keyword-based tweet filtering in social stream monitoring tools. While in algorithm comparison we have only tested a few keyword filters, in reality agents can try many different keyword filters, and iteratively improve their keyword selection by manually scanning the resulting tweets.

As a result, to answer Q4 and understand the practical usefulness of CrowdE, we compare the CrowdE UI against a keyword-based filtering UI in an in-lab user study. Subjects were asked to find engagement targets using both UIs within time limits. The evaluation was based on both objective measures and subjective assessments.

Data Collection

We set up a single-word filter on a full Twitter live feed using the word "delta", and collected 600 consecutive tweets. We sorted the 600 tweets chronologically, and formed a dataset Delta-A with the 300 odd-indexed tweets, and another dataset Delta-B with the 300 even-indexed tweets (i.e. put the first tweet in Delta-A, the second in Delta-B, the third in Delta-A, the fourth in Delta-B, etc.). Delta-A therefore contains consecutive tweets from a 50% live Twitter feed, while Delta-B contains tweets from a separate 50% feed over the same time period. We then created two other datasets Hertz-A and Hertz-B in a similar fashion. This process gave us in total four datasets for use in the study, each simulating a sample Twitter segment that agents could have encountered in reality. Table 3 reports how these tweets passed the crowd-enabled filters we built.

Measuring Correct Targets

The decision of making a follow-up could sometimes be subjective and vary by individuals. For example, while most would agree that a customer complaint about service delays should be marked for an apology follow-up, it is debatable if a retweet of a news article on the service delay problem should also receive an apology. To account for the subjectivity of such decisions, three authors scanned through the four datasets and independently identified all the follow-ups according to their best judgment. We consider a tweet for which all three authors agreed on the same follow-up as a *correct* target, as the strong agreement indicates that most people would likely expect the same follow-up as well. Most of such correct targets passed the two filters (Table 3). During the user study, a subject's follow-up is considered *correct* if it matches the three authors' consensus. Follow-ups that do not match the consensus are considered *subjective*, as they are not necessarily wrong.

Study Subjects

We recruited 12 subjects from our organization. Two of the subjects were public relation agents, whose jobs include monitoring Twitter and engaging people on behalf of the company brand. The remaining ten subjects were researchers, all of which had prior experience with Twitter and visited Twitter at least a few times a week. We paid each subject with a $5 lunch coupon, and awarded a $50 gift card to the best performing subject (explained below).

The qualitative feedback and quantitative performance of the two agents were not substantially different from the other subjects during the study. As a result, we believe our findings can likely generalize to other agents.

Study Procedure

The user study was of within-subject design, where each subject is exposed to two UIs. The first UI is the CrowdE UI as shown in Figure 1. The second UI is referred to as the keyword UI, modeled after keyword-based filter UIs in current social stream monitoring tools. To maintain visual consistency, the keyword UI uses the same layout as the CrowdE UI, but replaces the radio buttons in the upper-left with a text box and a submit button. Subjects can type whitespace-separated keywords in the text box. Whenever the subject presses the submit button, the list of tweets

Dataset	All Tweets			Correct Targets		
	Total	Passing Relevance Filter	Passing Both Filters	Total	Passing Relevance Filter	Passing Both Filters
Delta-A	300	62	26	11	9	8
Delta-B	300	70	27	11	10	9
Hertz-A	300	75	32	24	23	20
Hertz-B	300	74	30	20	18	18

Table 3. Datasets Used in User Evaluation.

below is filtered so that only tweets that include all the keywords in the text box are displayed.

At the beginning of the study each subject was introduced to the two UIs, and was allowed to practice with the two UIs on a separate practice dataset as much as desired.

Each subject then ran through four study sessions, and in each session used one UI to work on the 300 tweets from one of the four datasets we prepared. Subjects were asked to mark tweets from complaining customers for apology follow-ups, and mark tweets from appreciative customers for thank-you follow-ups. We limited each session to 6 minutes, so as to simulate the scenario where the list of tweets auto-refreshes every 6 minutes with 300 new tweets.

To encourage realistic and active use of filters, we instructed subjects on two points: 1) the study was also a competition, and the subject who marked the highest number of correct follow-ups would obtain a prize at the end of the study; and 2) as 6 minutes is far from sufficient to read all 300 tweets, using the filters appropriately is a key for winning the prize.

To ensure that each UI was first exposed to exactly half of the subjects, each dataset was seen by each subject exactly once, and each dataset was used with each UI in exactly half of the cases, we evenly divided the subjects into two groups and designed the sessions in the two groups as in Table 4.

At the end of the study, subjects reported subjective ratings for both UIs on aspects including efficiency, confidence, completeness, difficulty and tediousness, where each aspect is represented by a statement about the UI (Table 5). The ratings used a 9-point Likert scale, with 1 meaning "strongly disagree" with the statement and 9 meaning "strongly agree". The choice of the aspects and the scales of subjective ratings was guided by prior work (Bernstein et al. 2010). Finally, we led subjects through semi-structured interviews to gain qualitative insights about their experiences.

Results: Objective Measurements

On average, each subject marked 16 follow-ups for Delta and 25 follow-ups for Hertz when using the CrowdE UI. Each subject marked 11 follow-ups for Delta and 17 follow-ups for Hertz when using the keyword UI.

	Group 1	Group 2
Session 1	CrowdE UI on Delta-A	Keyword UI on Delta-A
Session 2	CrowdE UI with Hertz-A	Keyword UI on Hertz-A
Session 3	Keyword UI with Delta-B	CrowdE UI on Delta-B
Session 4	Keyword UI with Hertz-B	CrowdE UI on Hertz-B

Table 4. UIs and Datasets in User Evaluation Sessions

In both UIs, the subjects fully utilized the filters in marking these follow-ups: In a 6-minute session on the CrowdE UI, an average subject spent merely 42 seconds without filters, spent 2:32 on the brand relevance filter, and spent 2:46 on the brand relevant opinion filter. In a 6-minute session on the keyword UI, an average subject spent merely 63 seconds without filters, and spent the remaining 4:57 on keyword filtered tweets, using 15 different keyword queries. Most of these queries consisted of one or two words.

We used a Linear Mixed Model to analyze how the two UIs affected the number of correct follow-ups marked by subjects. We modeled the number of correct follow-ups as the dependent variable, the UI (CrowdE vs. keyword) as a fixed effect, and both the UI and the dataset keyword (Delta vs. Hertz) as repeated effects grouped within each subject. This model compensated for individual differences between subjects, the interdependence among observations of the same subject, and the inherent difference between the two dataset keywords. We also created a similar model for subjective follow-ups.

As shown in Figure 4, the CrowdE UI significantly increased the number of correct follow-ups compared to the keyword UI (F(1, 16.3)=27.0, p<.001), and did not significantly affect the number of subjective follow-ups (F(1, 11.6)=0.22, p=.65). Unsurprisingly, we also found that subjects marked more correct follow-ups for Hertz than for Delta (F(1, 15.3)=105.7, p<.001), as there were more tweets worth marking in the two Hertz datasets.

Results: Subjective Assessment

We used the nonparametric Wilcoxon signed ranks test for paired samples to compare the Likert scale ratings across the two UIs, as the distribution of ratings was not normal. Table 5 shows that overall subjects felt that finding targets through the CrowdE UI was significantly more efficient, more complete, less difficult, and less tedious. The CrowdE UI also gave subjects significantly more confidence in their actions.

For the CrowdE UI, a majority of subjects praised the brand relevant opinion filter for its accuracy. "Most of the stuff was relevant", one subject said, "it really picked up the subset that was of high quality for me to work on."

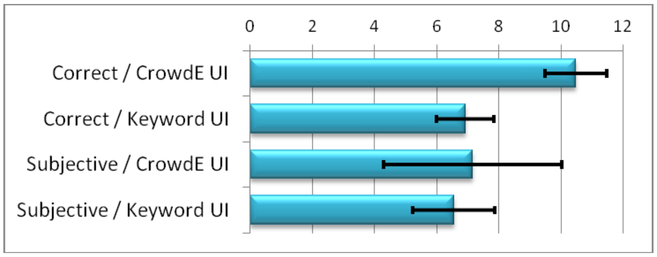


Figure 4. Follow-ups Marked Using the Two UIs. Shows the average number of correct and subjective follow-ups marked with the two UIs per user evaluation session. Error bars show 95% confidence intervals.

Aspect	Statement	CrowdE UI	Keyword UI	p-value
Efficiency	I can quickly identify targets using this UI.	7.8	4.2	< .01
Confidence	I am confident that all the targets I identified are appropriate using this UI.	7.4	6.5	< .05
Completeness	It is unlikely that I will miss targets using this UI	4.9	1.8	< .01
Difficulty	The task seems difficult using this UI	2.0	6.0	< .05
Tediousness	The task seems tedious using this UI	3.8	6.5	< .05

Table 5. Subjective Ratings of CrowdE UI and Keyword UI for Finding Engagement Targets
Each UI was rated on each aspect on a 9-point Likert scale, with 1 meaning "strongly disagree" with the statement and 9 meaning "strongly agree". Numbers shown in the table are mean Likert ratings across the 12 subjects.

Another subject concurred: "It is motivating to read through most of the posts, because they are about airlines and cars, and have opinion. When you back up to brand relevance only and all the tweets, there is just a lot of noise." Meanwhile, a few subjects did not use the opinion filter and relied mostly on the brand relevance filter. They cited two reasons for this decision: 1) they suspected that the opinion filter may miss targets (and they were right); and 2) given that the brand relevance filter could filter the 300 tweets in a session down to fewer than 100 tweets, they felt they could review all brand relevant tweets within the time limit and make their own judgments on the sentiment of each tweet.

For the keyword UI, many subjects reported difficulty in specifying the right keywords. One subject said: "Often I put in a keyword, it still ended up showing too many things that were off." Another subject seconded: "You have to figure out what sentiment words are, and you may be surprised that words in your vocabulary are not actually used in this context. There is luck and chance that you find the right word." A few subjects mentioned a strategy they discovered to alleviate this problem: they scanned for relevant tweets themselves, picked out words that signaled the relevance, and then tried filtering by these words to see if they also appeared in many other tweets. This strategy is interesting in that it follows the same rationale behind CrowdE, except that in our case we used the crowd to identify relevant tweets and used machine learning to pick out the useful words.

When asked to compare the two UIs directly, a majority of subjects clearly preferred the CrowdE UI for the reasons above. Only two subjects, one public relation agent and one researcher, dissented. Both of the subjects were skilled at coming up with keywords quickly, and argued that the keyword UI was comparable to the CrowdE UI in terms of helpfulness while offering a different trade-off. One of them explained: "Filtering by keywords involves more work, but it also gives me control. The automatic filter is only accurate to a point, and since I don't know what keyword it uses, I don't know if it does what I want." However, note that despite their subjective opinion, their objective performance under the CrowdE UI was still substantially better as measured by the number of correctly marked follow-ups.

Discussion

Below we discuss the practical impact of the system and a few design implications from our evaluation.

Practical Impact

The CrowdE system can serve as a practical solution for filtering opinionated customers across brands. Compared to keyword-based filters in current social stream monitoring tools, our intelligent filters are stronger and more helpful to agents. Meanwhile, with the common crowd-enabled process, creating filters for a new brand requires merely 15 minutes for setup and ~US$70 for the crowd labels. While this work only reports thorough evaluations of Delta and Hertz, following the same process we were able to build effective intelligent filters for brands in various domains, ranging from fast food to consumer electronics. The CrowdE system can also be used for other social stream platforms (e.g. Facebook, Google+ and Tumblr).

The CrowdE system likely works best when there are many tweets and a tight time constraint. For instance, if a brand gets 24,000 social media posts a day using a keyword filter and can fund 8 hours of work daily on finding engagement targets, the workload would be 3,000 posts per hour, or 300 posts per 6 minutes, as in our study. Intelligent filters show value here but a fast reader can get by with only the relevance filter. If the volume is lower or more manpower is available, filtering can become unnecessary; on the other hand, if the volume is higher or less manpower is available, even a fast reader may find both filters indispensible.

A limitation of our user evaluation is that subjects only learned and used the UIs in a short period of time, and their task proficiency may improve with more usage. However, as our study reported similar results for both experienced agents and novice users, it is unlikely that longitudinal usage would greatly change our overall finding.

Design Implications

Prior machine learning work has often used crowd workers as ad-hoc sources of ground truth. In this work, we took a novel design paradigm by specifying a concrete reusable process for collecting crowd labels and building classifiers. This design paradigm served us well for minimizing per-brand effort in data collection, algorithm design, and performance tuning. The same design paradigm may be used to save system building effort in other situations as well, such as creating intelligent filters on a topic in many different languages at the same time.

The crowd workers can be further integrated into the filter creation process, beyond simply providing labels. For brand relevance, the crowd-enabled decision tree had a lower recall than the expert created rules (Table 2). Comparing the internals of the two methods revealed one key reason: while the decision tree can only learn from the labels, people can judge the tweet content and make generalizations. For instance, the decision tree cannot reliably judge the phrase "1st class" as relevant to airlines unless the phrase appears at least a few times with the right label. In contrast, people can immediately judge the phrase as relevant, even if it appears only once or had the wrong label. People can also generalize the idea and add phrases such as "economy class" to the rule set. Cues such as "1st class" could be critical for the recall, as they can often be the only signal of relevance in a tweet. Since we have already involved crowd workers, we may improve the situation by asking crowd workers to do a bit more. That is, besides asking for the relevance of a tweet, we can also ask crowd workers to indicate relevant phrases from the tweet and/or to make generalizations. In addition, we can provide a keyword UI to crowd workers, so that they can validate the phrases they find, like some of the subjects were able to do in our study.

Our user study revealed a trade-off between the keyword UI and the CrowdE UI. Compared to the keyword UI, users of the CrowdE UI give up direct manipulation of keywords, and instead manipulate higher level concepts, i.e. relevance and opinion. This trade-off is comparable to the trade-off between direct manipulation and intelligent agents (Shneiderman et al. 1997), and we have indeed observed comparable results, including a performance improvement due to additional machine intelligence and user suspicion due to loss of full control. Future work could explore combinations of the two UIs that provide keyword manipulations in intelligent filters. Future work could also improve the transparency of the intelligent filters to increase users' trust, such as providing keyword-based explanations for the intelligent filters.

Conclusion

Filtering tweets for direct customer engagements has been a common need in social stream monitoring. In this work, we introduced a Twitter-based filtering system that helps agents find opinionated customers through brand-specific intelligent filters, and introduced a common filter creation process to minimize per-brand effort in creating the filters. We evaluated the system from various angles, and showed that the system is superior to the keyword-based filters used in current social monitoring tools.

We revealed several future directions, including deeper integration of crowd workers for building better filters, and deeper exploration on the trade-off between intelligence and user control in the UI of filtering systems.

Acknowledgement

Research was sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C-0028. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Defense Advanced Research Projects Agency or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Baccianella, S., Esuli, A. and Sebastiani, F. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC'10.

Bernstein, M., Suh, B., Hong, L., et al. 2010. Eddi: Interactive topic-based browsing of social status streams. In UIST'10.

Brew, A., Greene, D., and Cunningham, P. 2010. Using crowdsourcing and active learning to track sentiment in online media. In ECAI'10.

Chen, J., Nairn, R., Chi, E.H. 2011. Speak little and well: recommending conversations in online social streams. In CHI'11.

Diakopoulos, N. A., and Shamma, D. A. 2010. Characterizing debate performance via aggregated Twitter sentiment. In CHI'10.

Hennig-Thurau, T., Malthouse, E., Friege, C., et al. 2010. The impact of new media on customer relationships. Journal of Service Research, 13 (3). 311-33.

Hsueh, P., Melville, P., and Sindhwani, V. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In HLT'09.

Jansen, B. J., Zhang, M., Sobel, K., et al. 2009. Twitter power: Tweets as electronic word of mouth. Journal of American Society for Information Science & Technology, 60(11), 2169-2188.

Little, G., Chilton, L. B., Goldman, M., et al. 2010. TurKit: Human computation algorithms on Mechanical Turk. In UIST'10.

O'Connor, B., Balasubramanyan, R., Routledge, B., et al. 2010. From Tweets to polls: Linking text sentiment to public opinion time series. In ICWSM'10.

Paek, T., Gamon, M., Counts, S., et al. 2010. Predicting the importance of newsfeed posts and social network friends. In AAAI'10.

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval (2).

Paul, S. A., Hong, L., and Chi, H. 2011. What is a question? Crowdsourcing tweet categorization. In CHI'11 Workshop.

Pollock, S. 1988. A rule-based message filtering system. ACM Trans. Inf. Syst. 6, 3, 232-254.

Quinn, A.J., and Bederson, B.B. 2011. Human-machine hybrid computation. In CHI'11 Workshop.

Salton, G. and Buckley, C. 1998. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 (5): 513-523.

Shneiderman, B. and Maes, P. 1997. Direct manipulation vs. interface agents. Interactions, 4, 6, 42-61.

Snow, R., O'Connor, B., Jurafsky, D., et al. 2008 Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks. In EMNLP'08.

Witten, I.H., Frank, E., and Hall, M.A. 2011. Data mining: Practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann.