Automatic Selection of Social Media Responses to News

Tadej Štajner
Jožef Stefan Institute
Jamova ul. 39
1000 Ljubljana, Slovenia
tadej.stajner@ijs.si

Bart Thomee
Yahoo! Research
Avinguda Diagonal 177
08018 Barcelona, Spain
bthomee@yahoo-inc.com

Ana-Maria Popescu
Research Consulting
Mountain View, CA, USA
anamaria.popescug@gmail.com

Marco Pennacchiotti
eBay, Inc.
2065 Hamilton Ave.
95125 San Jose, CA, USA
mpennacchiotti@ebay.com

Alejandro Jaimes
Yahoo! Research
Avinguda Diagonal 177
08018 Barcelona, Spain
ajaimes@yahoo-inc.com

ABSTRACT

Social media responses to news have increasingly gained in importance as they can enhance a consumer's news reading experience, promote information sharing and aid journalists in assessing their readership's response to a story. Given that the number of responses to an online news article may be huge, a common challenge is that of *selecting* only the most interesting responses for display. This paper addresses this challenge by casting message selection as an optimization problem. We define an objective function which jointly models the messages' utility scores and their entropy. We propose a near-optimal solution to the underlying optimization problem, which leverages the submodularity property of the objective function. Our solution first learns the utility of individual messages in isolation and then produces a diverse selection of interesting messages by maximizing the defined objective function. The intuitions behind our work are that an interesting selection of messages contains diverse, informative, opinionated and popular messages referring to the news article, written mostly by users that have authority on the topic. Our intuitions are embodied by a rich set of content, social and user features capturing the aforementioned aspects. We evaluate our approach through both human and automatic experiments, and demonstrate it outperforms the state of the art. Additionally, we perform an in-depth analysis of the annotated "interesting" responses, shedding light on the subjectivity around the selection process and the perception of interestingness.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Social Media, Microblogging, Sampling, Summarization

1. INTRODUCTION

Social media services are increasingly playing a major role as platforms for people to express and share their opinions on current events. Responses to real-world events often come in the form of short messages, referring to specific topics, content or information sources. Such messages serve several purposes: to share a particular news article, to express an opinion about the ongoing events, or to add or refute information about the topic or the mentioned article. The extensive use of social media during recent major events (e.g. the Arab Spring and the Financial Crisis) shows that its use in these situations has become pervasive. On Twitter, for instance, a significant share of all tweets posted concerns news events [18].

Considering this large volume of messages being posted in the context of news, keeping track of messages that refer to the most popular articles can easily become overwhelming. This information overload motivates the development of automatic systems that select and display only the most interesting messages. This *social media message selection* problem is illustrated in Figure 1: the selection system takes as input a news article and all the social media responses referring to that article, and outputs the most interesting subset of responses. Our work is motivated by a variety of existing and future applications in which interesting social media responses could be automatically coupled with traditional news content, e.g., for displaying social responses near a news article for an enhanced reading experience.

Even though quantifying the interestingness of a selection of messages is inherently subjective, we postulate that an interesting response set consists of a diverse set of informative, opinionated and popular messages written to a large extent by authoritative users. By decomposing the notion of interestingness into these *indicators* we can pose the task of finding an interesting selection of messages as an optimization problem. We aim at maximizing an objective function which explicitly models the relationship among the indica-

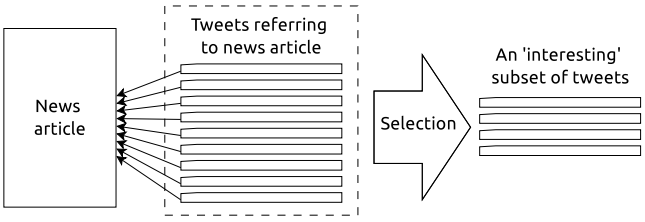

Figure 1: The social media message selection problem: selecting a small subset of messages sent in response to a news article.

tors thus producing selections that the typical person finds most interesting. We expect each indicator to have some influence on the messages ultimately selected to be part of the interesting set. Our method considers multiple content, social and user features to infer the intrinsic level of *informativeness, opinionatedness, popularity and authority* of each message, while simultaneously ensuring the inclusion of *diverse* messages in the final set.

We evaluate our approach through both human and automatic experiments and demonstrate that it outperforms the state of the art. In addition, we perform an in-depth analysis of the human evaluations, shedding light on the subjectivity and perception of interestingness in this particular task. To the best of our knowledge we have created the largest manually annotated news-response dataset to date, consisting of 45 news articles and 28,055 ratings per indicator annotated by 14 professional editors. Even though in this paper we focus on messages posted on Twitter, our method can be applied to responses to news in other platforms, for instance article comments or discussion forums.

The remainder of this paper is organized as follows. In Section 2 we first discuss related work. In Section 3 we present our proposed method, after which in Section 4 we describe our experiments and discuss the results. In Section 5 we conclude the paper with final remarks and future work.

2. RELATED WORK

There is an increasing amount of literature on social media sampling and summarization, where the goal is respectively to select a representative sample of messages on a given topic [4, 10] or to build a summary of an event from social media posts [31, 20, 6]. Closely related is also the task of update summarization, which aims to create a summary of new information given one or more documents that have already been read, focusing mainly on web documents [21] as well as social media, especially within opinions [15].

The idea of interestingness is also related to the work done on predicting re-posts of messages [25], where *retweets* are considered a proxy for interestingness. The authors of the paper use a varied set of features for predicting retweets, including emotion-related information, the presence of lexical indicators of interestingness and mood indicators; bag-of-words features are also explored but revealed low predicting power due to sparsity issues. Interestingness was also explored in the context of whole conversations within social media [11], where it was decomposed into two modalities: the interestingness of both the participants, as well as the conversation. While the former can be characterized by participation in interesting conversations, the latter is characterized by the topic, the communication dynamics of the participants and the conversation itself. Another aspect that we consider as part of interestingness is content quality, one of the most important indicators in question-answering communities [1], where the model combines both the social aspects and the linguistic quality of the messages. In image tagging [13], an interesting object has been defined to occur more frequently during a specific time interval than outside it, corrected for object frequency.

Our method has been inspired by the outcomes of studies analyzing the different types of social media reactions to different types of events, e.g. by the observation that breaking news is often accompanied by "informational" messages, whereas other news items are characterized by more "conversational" reactions [23]. Furthermore, events can elicit different levels of excitement and participation, depending on both the nature of the event (life, work), the characteristics of the participants [30] and even their opinion [22]. These multiple facets have been effectively tackled by multi-objective optimization approaches [9]. Recently, significant attention has been directed at detecting and tracking events as they surface in microblogs [12, 27]. Social media messages have also been used to predict the popularity aspect of news stories [3], as well as assist in their summarization [33]. Our problem statement looks at the reverse perspective by observing responses to news articles describing real-world events.

Overall, results presented in the literature point to a strategy that contextualizes the posted messages as much as possible, generating not just content, social and communication aspects, but also diverse textual and linguistic descriptors. To construct rich individual and collective features we therefore leverage insights from several existing approaches. Primarily, we build on existing work on sentiment and intensity analysis [29, 30], and apply basic redundancy detection [35]. Additionally, we incorporate user authority, both for specific topics as well as in general [2].

3. PROPOSED METHOD

We formulate the social message selection problem as follows.

Problem statement: Given a news article and a set of related messages M, we seek a subset $S \subseteq M$ of k messages which are the most "interesting" to a typical reader in the context of the article.

We represent "interestingness" using a set of *indicators*. We identify four message-level indicators : *informativeness, opinionatedness, popularity, authority*; and one set-level indicator, *diversity*. The intuition is that an interesting selection of messages contains *informative*, *opinionated* and *popular* messages referring to the news article. Furthermore, the messages should be written mostly by users who have *authority* on the topic. Looking beyond single messages, we posit that an interesting selection should contain messages that are *diverse* in content.

Solution: We computationally model the four message-level indicators using an utility function r and the set-level diversity indicator using a normalized entropy function H_0. The solution to the social message selection problem is then to find the subset S^* that maximizes the objective function $g(S)$, which measures the "goodness" of subset S as follows:

$$g(S) = \lambda \sum_{m \in S} r(m) + (1 - \lambda) H_0(S) \quad (1)$$

where $r(m)$ represents the utility score of a message m and $H_0(S)$ is the normalized joint entropy of the entire set of messages S. We balance the effect of the collective diversity indicator on the sampling by specifying a suitable λ, which we will define manually in our experiments. The maximization problem is thus defined as follows:

$$S^* = \underset{S \in 2^M, |S|=k}{\arg\max} \; (g(S)) \quad (2)$$

In Section 3.3 we will describe how we derive r and H_0, while we first focus on how to solve the optimization problem in Sections 3.1 and 3.2.

3.1 Submodularity of the objective function

In order to find an optimal solution for the maximization problem we could exhaustively search the space of all possible message subsets. However this proves computationally prohibitive. We therefore consider the properties of our objective function in order to find a fast, approximate solution to the problem. To this end we prove that $g(S)$ is submodular. This key property allows us to use a greedy algorithm [26] that provides efficient linear solutions within a deterministic error bound from the best possible solution.

A submodular function f on the set Ω is defined as the set function $f : 2^\Omega \rightarrow \Re$, where 2^Ω is the power set of Ω and one of several equivalent conditions are satisfied, amongst which the following:

$$\begin{gathered} \forall X, Y \subseteq \Omega, X \subseteq Y, \forall x \in \Omega \setminus Y : \\ f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y) \end{gathered} \quad (3)$$

Intuitively, this condition specifies that f has a natural diminishing return property, meaning that the difference of the value of the function that a single element addition makes decreases as the size of the input set increases. Here, we attempt to demonstrate that our function $g(S)$ is submodular. Let us refer to $\sum_{m \in S} r(m)$ as $R(S)$ for brevity. We then insert the entire objective function in the submodularity inequality condition mentioned above:

$$\begin{gathered} \lambda R(X \cup \{m\}) + (1 - \lambda) H_0(X \cup \{m\}) \\ -\lambda R(X) - (1 - \lambda) H_0(X) \\ \lambda R(Y \cup \{m\}) + (1 - \lambda) H_0(Y \cup \{m\}) \\ -\lambda R(Y) - (1 - \lambda) H_0(Y) \end{gathered} \quad \geq \quad (4)$$

Since the scoring function is a sum over all $r(m)$, which are independent amongst each other, it holds that $R(S \cup \{m\}) - R(S) = r(m)$ and thus $\lambda R(X \cup \{m\}) - \lambda R(X) = \lambda R(Y \cup \{m\}) - \lambda R(Y)$. Therefore, all the expressions of the scoring function cancel out and we are left with the expression for the submodularity of entropy:

$$H_0(X \cup \{m\}) - H_0(X) \geq H_0(Y \cup \{m\}) - H_0(Y) \quad (5)$$

In order to prove that $g(S)$ is submodular we therefore have to simply show that entropy is submodular. Previous work has already established that the entropy of a sample is a monotonic non-decreasing submodular function [17]. We can thus conclude that $g(S)$ is submodular.

3.2 Approximation algorithm

While maximizing a submodular function is still a computationally difficult problem, maximizing entropy can be solved using a greedy approach, achieving an approximate solution that is within a constant factor from the optimal solution,

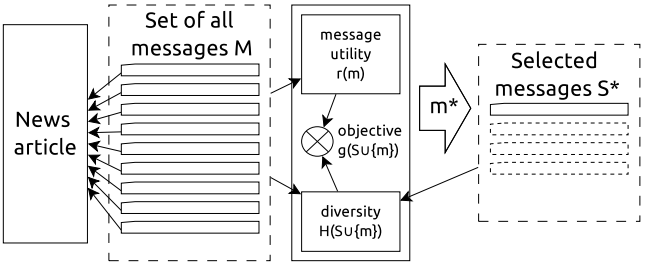

Figure 2: The diagram of the proposed algorithm, showing the two criteria, joint individual message and set diversity.

where no obvious heuristic would reduce this bound [16]. Here, since $H_0(\emptyset) = 0$, it follows that a greedy selection algorithm that selects a set of k elements is at most $1 - \frac{1}{e}$ times worse than the optimal set of examples. Thus, to approximate the optimal solution, we plug $g(S)$ into our greedy algorithm as shown in Algorithm 1 and visualized in Figure 2. Our algorithm initializes by picking the element m_i that ranks highest with the scoring function $r(m_i)$. Then, it iteratively adds the element m_i that would provide the biggest boost in the objective function if it were added to the set S, stopping at the desired sample size k.

Algorithm 1: Greedy iterative sampling algorithm

Input : A collection of messages M, the sample size k
Output: A sample set of messages S^*
$S \longleftarrow \{\underset{m_i \in M}{\arg\max}(r(m_i))\}$
while $|S| < k$ **do**
  $m^* \longleftarrow \underset{m_i \in M \setminus S}{\arg\max}(g(S \cup \{m_i\}))$
  $S \longleftarrow S \cup \{m^*\}$
end
$S^* \longleftarrow S$

3.3 Message scoring and message set selection

The two main components of the target function, r and H_0, are both computed from a large set of features. These features serve as computational proxies for the five indicators mentioned in the previous sections. Intuitively, most features are likely to be specific to the target social message platform; in this paper we therefore focus on Twitter-specific features, though most of them are generalizable.

Features: Table 1 reports the list of features, where superscripts s and d indicate features that are used for computing r and H_0, respectively.

- The *content features* capture the linguistic aspects that indicate how interesting, informative and opinionated a message is.
- The *social features* represent the level of diffusion of a message, reflecting its popularity.
- The *user features* capture the overall and topical authority of a user in the social network.

Individual message scoring $r(m)$: Given a tweet and its corresponding *scoring* feature values (scaled to interval $[0, 1]$), we use a supervised model as the scoring function $r(m)$. Specifically, we use the prediction given by support vector regression (ϵ-SVR [32]), trained using the manually

obtained labels for *interestingness* described in detail in Section 4.2, using only examples that were labeled with 1 and 3, omitting those with the label of 2.

Entropy of message set H_0: Given a message set S and a set of *selection* features used to collectively select messages, we treat these features as binary random variables. Let $H_0(S)$ denote the normalized entropy of the message set S given a probability model that uses the empirical probability of the selection feature values: $H_0(S) = H(S)/\log d$ in which d is the number of binary random variables, and $H(S) = -\sum_{i=0}^{d} p(f_i = 1)\log p(f_i = 1)$, where $p(f_i = 1)$ is the empirical probability that the feature f_i has the value of 1 given all examples in S. The intuition behind maximizing entropy is that it favors adding examples to S with different non-zero features from the ones already in S, since adding an example with different features increases entropy more than adding an example already similar to the ones in S.

Balancing individual and collective scores: Using the parameter λ, we can balance the influence between scoring and entropy: $\lambda = 0$ yields an entropy-only approach, which we denote as ENTROPY, whereas $\lambda = 1$ gives a scoring-only approach that we refer to as SVR. The balanced combination at $\lambda = 0.5$ is denoted as SVR_ENTROPY.

3.4 Comparison with the state of the art

Our research is closely related to the diversity-based sampling method of De Choudhury et al. [10] and the social context summarization method of Yang et al. [33] that currently are considered state of the art. We first briefly describe these two methods and discuss their strengths and weaknesses in order to motivate our approach.

Diversity-based sampling

De Choudhury et al. [10] tackle tweet selection using a greedy approach that iteratively picks tweets that minimize the distortion of entropy of the current subset, given a certain diversity criterion. The approach prescribes building a probabilistic model of tweets given their diversity using social, content and network features. The rationale behind the approach is that samples with a low entropy will be highly homogeneous with respect to the feature space, while those with a high entropy will be highly heterogeneous. In a social media message selection context it is thus preferred for the method to produce a sample with high entropy in order to promote diversity in the selected messages. The sample generation is initiated by adding a random message to the sample, after which new samples are iteratively added, so that the normalized entropy of the sample is closest to the pre-specified diversity parameter. This iterative process is repeated until the desired sample size has been reached. In our evaluations we refer to this method as DIVERSITY.

Social context summarization

Yang et al. [33] propose an approach based on conditional random fields that simultaneously addresses message and document summarization by modeling messages and sentences as connected 'wings' in a factor graph, where factors are assigned to individual message and sentence features, between their instances on a single wing and across both wings. To determine the key sentences and the important messages, an approximate inference approach iteratively updates and propagates the beliefs between factors and labels, where the final labeling indicates which sentences and messages are selected. We refer to this dual-wing factor graph method as DWFG. In the same paper, a simplified model was also presented, implemented as a logistic regression classifier, that enriches each tweet with the features of the sentence most similar to it rather than introducing sentence-tweet factors between similar tweet-sentence pairs as is the case in the original model. In the task of summarizing relevant tweets for a news article this method yielded good performance. We refer to this method as LR+.

Observations

Even though the entropy measure used in the DIVERSITY approach has been demonstrated to be a good proxy for diversity, it does not evaluate individual messages with respect to their individual characteristics. Furthermore, its content probability model assigns each message to a broad category, such as 'politics' and 'sports'; while this works very well for obtaining a diverse sample on broader topics, this may not be a good fit in the case of responses to a single news story on a relatively narrow topic. The dual-wing factor graph and logistic regression methods take a richer set of features into account, but do not explicitly model diversity in the context of how people respond to news. Specifically, connecting messages only through their reply/retweet relations means that highly similar or even duplicate messages – a common occurrence in the context of news – may not be linked to each other, likely resulting in a substantial amount of redundancy being present in the final selection. Our method therefore offers the following improvements:

- We focus on a comprehensive notion of interestingness instead of solely optimizing for message diversity (DIVERSITY) or for message summarization (DWFG and LR+).
- We detect and avoid redundancy between the content of messages by including a richer set of textual features.
- We model the importance of messages beyond their literal content by including features that capture the wider social, conversational and linguistic context, e.g. the amount of opinion, emotion and controversy of a message, the level of authority of the author on the topic, and the user's intent of the message [24].
- We provide a theoretic guarantee of the sample quality.

4. EXPERIMENTS

In this section we describe our experiments for evaluating the performance of our proposed method against the previously mentioned approaches on the problem of selecting the most interesting tweets posted in response to news.

4.1 Dataset

We collected a set of tweets posted on Twitter between February 22, 2011 and May 31, 2011 that were written in the English language and that included a URL to an article published online by news agencies such as Reuters, Huffington Post, India Times and Al Jazeera. We crawled each of these links to obtain the original news article content, discarding redirected and unresolved links. We observed that the news-related tweets frequently had highly similar content to each other; when ignoring retweet markers, links, mentions and hashtags, and only purely looking at the words, just over half of the tweets were identical in our dataset. For each news article, we therefore filtered out all such duplicate tweets and only retained the one with the earliest timestamp. For the experiments we ultimately picked a set of 45 news articles

Table 1: The features used by our methods. Superscripts s and d respectively indicate features used in r and H_0, respectively.

N-grams^d	All unigrams, bigrams, and trigrams taken over the words in the tweet.
Tf-idf score^s	Average tf-idf score of all words in the tweet, emphasizing rarely-used and penalizing out-of-vocabulary words.
Log-likelihood^s	Likelihood of the tweet, based on a bigram language model constructed from all of the tweets of the article.
Number of words^s	A higher number of words may indicate a tweet with more useful content.
Me-information^s	Presence or absence of a first-person pronoun, indicating if the tweet is mainly about the author himself [24].
Question^s	Presence of a question mark in the content.
Quote sharing^s	Presence of a quotation from the news article in the tweet.
Quality^s	An measure of the tweet quality, indicating how well-written the tweet is. We use a supervised approach using on a lexicon of low-quality terms, an English dictionary, and the proportion of words, hashtags, capitalized characters, and repetitions.
Sentiment^s	% of positive and negative words, % of subjective words, and a mixed sentiment score, reflecting the presence of highly emotional or opinionated expressions in a tweet [29].
Intensity^s	A measure indicating the strength of the user's reaction [30].
Explicit controversy^s	A measure reflecting the mention of common controversial issues [29].
Location^d	The geographic location derived from the tweet, either from a mentioned place or the location of the user.
Retweet and reply^{s,d}	A flag indicating whether the tweet is a retweet or a reply.
Followers, friends^s	Number of followers and friends.
Follower-friend ratio^s	Ratio between number of followers and friends.
Number of retweets^s	The total number of retweets of the user's posts over the time of data gathering, signaling user authority [34, 7].
Number of users retweeting^s	The total number of other users that retweeted this user at least once.
Tweet-retweet ratio^s	Ratio between the number of tweets a user posts and the number of retweets received.
User verified^s	A flag indicating if the user is verified by Twitter, which may increase the credibility of the user's posts.
User spam^s	A flag indicating if the user is a spammer, based on a large spam dictionary trained on web page spam.
User authority score^s	A global authority score, computed to have a topic-independent estimate of the user's overall importance.
User topic authority^s	We consider a user u authoritative for the article's topic, when previous tweets on the same topic were often retweeted. We extract the three most relevant named entities from the article using our "aboutness" system [28] and consider these entities as a synthetic summary of the topic of the article. For each of these entities we then extract the set of tweets of the user that mention the entity and look up the number of times any tweet in in this set was retweeted by other users. The topic authority score is then computed as a combination of the relevance of the entities for the article and the number of retweets.

that spanned a wide range of topics, including business, politics, technology, weather, disasters, and warfare. We ensured each article had on the order of 100 unique tweets associated with it, so that there would not only be sufficient tweets for the evaluated methods to work with, but also so that the tasks would not overly exhaust the annotators that created the gold standard.

4.2 Gold standard collection

We asked a team of 14 annotators to create a gold standard collection by: a) rating the individual tweets sent in response to an article on the *informative* and *opinionated* indicators, and b) selecting approximately 10 tweets they found the most *interesting* as a set in the context of the news article. We did not ask the annotators to rate tweets for the *authority* and *popularity* indicators due to the difficulty of assessing these values at the tweet level and without context. Instead, we opted for a simple mechanical model for deriving the values of these indicators, as we will explain later in this section.

For the *informative* and *opinionated* indicators, the annotators were asked to assign a score to each tweet on a scale from *1* (the tweet decidedly does not exhibit the indicator) through *2* (the tweet somewhat exhibits the indicator) to *3* (the tweet decidedly exhibits the indicator), while allowing them to skip tweets that were either spam or not relevant to the article. To strengthen the gold standard collection, each task was assigned to three different annotators. Due to time constraints, the number of ratings performed by annotators varied. In total, 28,055 ratings were gathered, of which approximately 70% were *1*s, 15% were *2*s, and 15% were *3*s. In other words, in aggregate, 70% of the tweets were not considered informative nor opinionated by the annotators.

As previously mentioned, we did not ask annotators to rate tweets for the *authority* and *popularity* indicators. First, the authority indicator is not straightforward to assess without context: while identifying major news outlets as authoritative (e.g., AP, Reuters, BBC, etc.) might be trivial, doing so for individuals is more nuanced. Second, the popularity indicator is directly related to how often a tweet is retweeted and replied to, which is difficult to assess when seeing a tweet by itself. Therefore, we opted for a simple mechanical model for both: for the authority indicator we combined the values of user authority and topic authority features of the users writing the tweets, and for the popularity indicator combines the retweet and reply counts for each tweet.

4.2.1 Inter-annotator agreement

We measured the inter-annotator agreement for all three tasks using Cohen's linearly weighted kappa statistic [8]. We averaged the pairwise kappa values of all possible combinations of annotators that had overlapping tweets they had rated, obtaining the overall kappa. Because the annotators had different scoring behaviors, i.e. they used different amounts of *1*s, *2*s and *3*s from each other, the maximum attainable kappa was less than 1.0. To place the kappa into context, we normalized it by dividing the obtained kappa by the maximum obtainable kappa. Even though the defi-

nition of what constitutes a 'good' kappa varies across the literature, a higher kappa always indicates higher agreement. Overall, the annotators had a relatively high agreement on *opinionated* tweets with $\kappa_{cwopi} = 0.61$, while only fair agreement of $\kappa_{cwint} = 0.35$ for *interesting* tweets and just slight agreement of $\kappa_{cwinf} = 0.20$ for *informative* tweets. The agreement on the latter two indicators indicates a high subjectivity. The annotators approached the content with different background knowledge, which may have contributed to what they themselves considered informative. As an example of inter-annotator (dis)agreement, we show three tweets sent in response to an article about Japanese citizens returning cash they discovered in the rubble after their town was hit by a tsunami:

> *"I am continually amazed at the honor and trustworthiness of the Japanese people. Truly astounding and heartening to see!"*
>
> *"Japanese citizens turning in cash found in tsunami zone - #cnn"*
>
> *"Giving respect to the Japanese people."*

The annotators agreed that the first tweet was interesting, possibly due to the personal touch of this well-written tweet. They also agreed that the second tweet was not interesting, presumably because it repeated the title of the news article. However, the annotators disagreed on the last tweet, where one annotator may have agreed with the conveyed idea, whereas another may have considered it someone's personal opinion and therefore not of interest. The latter case clearly exemplifies that rating tweets is a subjective matter.

4.3 Experimental evaluation

Using the gold standard annotations, we evaluate the methods on predicting interestingness. We also evaluate the performance on opinionatedness and informativeness in order to estimate their difficulty.

Methods: We evaluated the baseline methods DIVERSITY, DWFG and LR+ as presented in Section 3.4 against our proposed method SVR_ENTROPY and its two variations SVR and ENTROPY. We also experimented with other scoring mechanisms for individual messages, such as using the likelihood estimates from a logistic regression classifier for probabilistic scoring, and with other diversity mechanisms at the collective level, such as using more sophisticated natural language processing methods, like group-wise textual entailment. However, these alternatives did not show any significant benefit and we therefore do not include them in our evaluations.

Training: The methods that required training were trained with negative and positive judgments. For interestingness positive examples are those included in the interesting set by the annotators for each article, and all others are negative. For opinionatedness and informativeness positives and negatives are respectively ratings *1* and *3* from all annotators on all articles. We randomly sampled the tweets to ensure a balanced distribution of positive and negative judgments. For the DWFG we labeled all sentences of the article as *positive*, since we can consider a news article to be a summary of a newsworthy event.

Evaluation measures: To measure performance, we computed both the ROUGE-2 [19] and F_1 scores for the *informative*, *opinionated* and *interesting* indicators. These scores measure how closely the output of a method is to the "ideal" output, i.e. those scored positively by the annotators. While F_1 only considers exact tweets to be correct, ROUGE-2 counts tweets with similar content as a partial match, reporting the recall of word bigrams of the obtained sample and the gold standard examples. We performed 10-fold cross-validation, using nine folds for training (when needed) and one for testing.

Results: Table 2 contains the summary of our experimental results. While our proposed methods outperform the DIVERSITY, DWFG and LR+ baselines on the main task of predicting interestingness, the two evaluation metrics show different rankings of the methods.

- For *interestingness*, the best ROUGE-2 performance was obtained by methods that have an entropy selection component (ENTROPY, SVR_ENTROPY), while F_1 favors SVR-based approaches.
- For *opinionatedness*, we observed that while the ROUGE-2 scores were very similar across ENTROPY and SVR_ENTROPY, F_1 scores were again higher among SVR-based approaches. Although the margin was not statistically significant, the LR+ baseline was able to achieve the highest ROUGE-2 performance, likely due to enriching the feature set with the features of the most similar sentence. However, even though ROUGE-2 measured that the lexical overlap with the correct messages was high, F_1 scores show that the actual correct messages were not selected as often.
- For *informativeness*, the overall results were lower than for other indicators. This result matched the lower inter-annotator agreement of the ratings. Upon inspection we noticed that for the most part, informative tweets did not contain article snippets, but instead contained novel text. We also noted the same pattern of ROUGE-2 favoring ENTROPY, while F_1 favoring SVR.

Table 2: Comparison of the tweet sets generated by the various methods in terms of their informativeness, opinionatedness and interestingness as measured by ROUGE-2 and F_1 at $k = 10$. The asterisk marks the case where the measurement was significantly higher than all of the three baselines, having a T-test p-value below 0.05.

	Informative		Opinionated		Interesting	
	R-2	F_1	R-2	F_1	R-2	F_1
DIVERSITY	0.178	0.050	0.270	0.108	0.280	0.109
DWFG	0.170	0.048	0.259	0.056	0.235	0.051
LR+	0.175	0.032	**0.355**	0.132	0.264	0.054
ENTROPY	**0.246***	0.124*	0.345	0.184*	**0.357***	0.216*
SVR	0.212*	**0.132***	0.326	**0.217***	0.336*	**0.239***
SVR_ENTROPY	0.232*	0.123*	0.343	0.211*	0.346*	0.224*

To summarize, the ENTROPY approach excels at content fragment retrieval (ROUGE-2), whereas approaches combining a scoring and an entropy selection model are able to provide better performance on F_1, offering a tunable trade-off between diversity and individual tweet scores.

We also noticed some key differences in behavior of the baselines we compared against. Whereas the DWFG method ensures content diversity by encoding the diffusion network into the factor graph, so that retweets do not get selected when the original tweet does, in the news domain many of the near-duplicates tweets were actually not retweets at all: when responding on the same article, several users posted very similar responses independently, resulting into lower

performance on ROUGE-2. While this is not captured by the dual-wing factor graph design, it is implicitly taken care of in DIVERSITY, where it diversifies the tweet content represented with a set of topics [10]. However, the original DIVERSITY scenario focused on sampling from tweet sets that are retrieved via a hashtag query, which are still diverse enough to be accurately described by topics. On the other hand, news responses tend to be topically narrower, requiring finer-grained content representation, favoring our diversity representation.

4.3.1 Understanding tweet "interestingness"

Given the gold standard annotations, we aimed to infer to what extent the individual *interestingness* of a tweet can be decomposed into the individual indicators and how much influence each of the indicators had. In this investigation, we averaged the ratings received or computed for the individual indicators. We used the gold standard labels for the *informative*, *opinionated* and *interestingness* indicators as well as the automatically derived labels for the *popularity* and *authority* indicators in the same fashion as in Section 4.2. *Authority* was estimated by combining the features of topic authority and user authority. *Popularity* was derived by combining the number of retweets and replies for a given tweet. After normalizing indicators to the same interval, we fitted a least-squares linear model and obtained the following linear combination, supported by a coefficient of determination R^2 of 0.38, having a correlation coefficient of 0.62:

$$int = 0.60 \cdot inf + 0.29 \cdot opi + 0.03 \cdot pop + 0.10 \cdot aut \quad (6)$$

While this simplified model explains only 38% of the variance, it gives us a glimpse of what constitutes "interestingness". The primary driver is *informativeness*: while our annotators had low agreement on how informative a tweet is, having at least one of them mark a tweet as such was sufficient for the signal to get picked up. The second driver was *opinionatedness* followed by *authority* and *popularity*. The low coefficient for popularity can also be explained by the fact that it strongly correlates with the authority indicator, meaning that the majority of the effect is already explained by authority.

4.4 Preference judgment collection

While we already have shown a promising approach compared to existing literature, we conducted further experiments in order to better understand the differences within the variations of the proposed approach and to understand the extent to which certain factors contribute to the interestingness of a set of tweets. We performed a second evaluation in which we used another team of annotators, none of whom had previously participated in the gold standard collection, using the crowdsourcing service CrowdFlower.

We evaluated the satisfaction of the annotators with the sets of tweets ultimately produced for the same collection of 45 news articles. Instead of asking the annotators to assign a numeric score to each set of tweets, we paired the methods and placed them in a head-to-head competition, where we requested the annotators to express preference for either set or to indicate whether they were equivalently interesting. This approach allows us to extrinsically evaluate the quality of the sets produced by each of the methods, rather than having the annotators score each set separately or pick from multiple sets at once, both of which could have given ambiguous results. Unlike ratings, comparative judgments generate highly reliable responses and clear individual differences. In order to limit the number of pairwise comparison evaluations, we chose to compare our preferred method SVR_ENTROPY against DIVERSITY and DWFG, the principal methods proposed by the authors in their respective papers.

We requested at least five judgments for each task and appropriately mixed in trivial gold standard scenarios to filter out malicious annotators. We selected judgments from annotators who had at least 75% precision on the gold filter questions. After expressing a preference, the annotators were presented with a follow-up question to indicate why they had chosen one set of tweets over the other, allowing us to compare their preferences with the ratings the previous annotators had given to see what kind of, if any, correlations exist between the individual tweet and collective selection indicators, and the interestingness of the sample.

4.5 Preference judgment analysis

Once again, we first analyze the agreement between annotators and then look at the outcome of the evaluation to see which method performed better than the others. Finally, we take a closer look at what constitutes an "interesting" selection of tweets. We have structured the analysis in the following two scenarios:

Combining diversity and scoring: We are interested in understanding whether diversity alone or scoring alone is enough for achieving an interesting selection.

Comparison against baselines: Using our best performing method so far, we are interested in knowing how it compares to existing baselines on the same problem.

4.5.1 Method comparison

We aggregated the preference counts of the annotators and show the results in Table 3. We conducted a χ^2 statistical test over the aggregate counts with the null hypothesis that both of the methods in a given pair are equivalent. We observed that a combination of diversity and scoring (SVR_ENTROPY) produced better selections than diversity (ENTROPY) or scoring (SVR) alone. When evaluated separately, scoring (SVR) was better than diversity (ENTROPY); considering that in our dataset the tweets were already relatively diverse due to the prefiltering of duplicate tweets, scoring proved to be more important. Even though SVR_ENTROPY was consistently good (but not the best) within the gold standard evaluation, in light of the preference judgment results we consider it as the preferred option in further comparisons. Looking at the head-to-head comparisons against the baselines, we observed that the DIVERSITY approach performed similarly to the DWFG approach, whereas both are outperformed by SVR_ENTROPY.

Table 3: Comparison of tweet selection methods based on annotator preference votes.

Method A	Method B	Votes A	Votes B	None	p-value
Combining entropy with scoring					
SVR	SVR_ENTROPY	151	**194**	16	0.02
SVR	ENTROPY	**193**	155	9	0.04
ENTROPY	SVR_ENTROPY	114	**147**	13	0.04
Comparison against baselines					
DIVERSITY	SVR_ENTROPY	79	**158**	3	0.00
DWFG	SVR_ENTROPY	99	**149**	5	0.00
DWFG	DIVERSITY	**118**	94	10	0.10

4.5.2 Inter-annotator agreement

We measured the inter-annotator agreement for the preference judgments in two ways: we first used Cohen's unweighted kappa statistic to calculate all pairwise kappas when there was some overlap in the tasks between two annotators and we ultimately averaged them to obtain the overall κ_{cu} of 0.50, having a maximum κ_{cumax} of 0.81, thus yielding a mean normalized κ_{cunorm} of 0.62, showing moderate agreement. However, since every annotator only judged a limited number of examples and the overlaps were relatively sparse, we also measured a multi-rater Fleiss kappa [14]. Even though the overall κ_f was fair at 0.21, the results varied across different comparisons. While the Cohen and Fleiss kappas are not directly comparable, both point to fair-to-moderate overall agreement.

Table 4: Inter-annotator agreement via Fleiss' kappa κ_f. Scores from 0.00 to 0.40 indicate slight agreement, while 0.40 to 0.75 indicate good agreement.

Method A	Method B	κ_f
SVR	SVR_ENTROPY	0.35
ENTROPY	SVR_ENTROPY	0.06
ENTROPY	SVR	0.25
DIVERSITY	SVR_ENTROPY	0.04
DWFG	SVR_ENTROPY	0.09
DIVERSITY	DWFG	0.48

Agreement measurements in Table 4 show that in the cases with higher agreement, the results of the method comparison were also more clear (with some exceptions). For instance, ENTROPY vs. SVR_ENTROPY demonstrated low agreement, but still had a significant difference in overall votes, meaning that the ratio of votes was more constant over all articles, as opposed to there being a particular article where one method would consistently dominate. The cases with moderate agreement, such as SVR vs. SVR_ENTROPY and DIVERSITY vs. DWFG highlight situations where one method can objectively produce better samples.

4.5.3 Understanding set "interestingness"

After the annotators had completed a pairwise comparison task, we posed a follow-up question by asking the annotators what they considered the reason for their preference. They answered a multiple-choice question selecting the indicators they considered that contributed to interestingness. Popularity and authority were omitted from the selection given that users cannot be expected to infer them from the situation at hand, so both were calculated using the same process as in Section 4.2. If one sample had at least one more popular or authoritative tweet than the other, that indicator was automatically counted as a reason. While this does not enable comparison between importance of individual indicators, it does allow for comparisons across various method pairs. Users were also asked to supply additional reasons beyond the pre-defined ones.

Table 5 shows the number of times the annotators stated a particular reason for preferring the selection produced by one method over the other, grouped by the preferred method. The most interesting observation is the prominence of opinionatedness in some scenarios: this was the primary driver between distinguishing SVR and ENTROPY, demonstrating the effectiveness of having the scoring operate on sentiment and intensity features. On the other hand, informativeness was more difficult to predict, making the samples harder to distinguish on informativeness than on opinionatedness, where differences are more obvious. Authority was less prominent in the aggregate. In some cases, the annotators remarked that they knew the author, suggesting that the social proximity of the author of the tweet to the annotator (and subsequently any reader of the produced selections) serves as a potential personalization opportunity. Other comments given by the annotators outside of the pre-defined choices noted humor as an important factor for preferring one set over the other, which was not modeled by any of the methods, as well as several instances of a set not being preferred due to containing profanity or too much controversy, indicating the importance of these features.

Table 5: Distribution of preference reasons for the tested method pairs.

Method A	Method B	inf.	opi.	aut.	pop.
SVR	SVR_ENTROPY	163	197	59	49
ENTROPY	SVR	139	226	43	54
ENTROPY	SVR_ENTROPY	160	125	24	13
DIVERSITY	SVR_ENTROPY	147	122	66	68
DIVERSITY	DWFG	166	109	66	50
DWFG	SVR_ENTROPY	146	134	85	68

5. CONCLUSIONS

We proposed an optimization-driven method to solve the social message selection problem for selecting the most interesting messages posted in response to an online news article. Our method first learned the utility of individual responses in isolation and then produced a diverse selection of interesting responses by maximizing an objective function that jointly models the responses' utility scores and the entropy. We proposed a near-optimal solution to the optimization problem that leveraged the submodularity property of the objective function. The intuition behind our work was that an interesting selection of messages contains diverse, informative, opinionated and popular messages referring to the news article, written mostly by users that have authority on the topic. Our intuitions were embodied by a rich set of content, social and user features capturing these aspects.

We compared three variations of our method against state of the art approaches (DIVERSITY and DWFG) in two experimental settings: first, using a gold standard consisting of messages rated by professional annotators and second, using pair-wise comparative judgments obtained via crowdsourcing. We found that our method, SVR_ENTROPY obtains the best overall performance among the tested approaches.

Furthermore, we investigated how various indicators affect *interestingness* of a message and a message set. Our preliminary results based on the collected gold standard data found that for individual messages,*informativeness* is most important but simultaneously the most difficult to judge, followed by *opinionatedness* and *popularity*. We also found that *message set interestingness* was tied most strongly to *opinionatedness* and then to *informativeness*.

There are many directions of future work: incorporating additional message-level or author-level indicators, or (motivated by the low inter-annotator agreement for some indicators) focusing on personalized models based on the users' topical interests or their social circles [5]. Another possible direction is moving from extractive sampling to using methods based on abstractive summarization, rephrasing the contents of the social media posts and enabling shorter,

more concise summaries. The real-time nature of social media message streams can also be viewed as an incremental sampling setting, where we cannot assume that entire sample is available ahead of time. Our proposed approach is amenable to extension to online sampling, provided some criterion for substituting already chosen messages with new ones.

6. ACKNOWLEDGMENTS

This research is partially supported by European Commission Seventh Framework Programme FP7/2007-2013 under the ARCOMEM, SOCIAL SENSOR and RENDER (ICT-257790-STREP) projects, by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, 'Social Media" and the Slovenian Research Agency.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 1st ACM WSDM*, pages 183–194, 2008.

[2] E. Bakshy, J. Hofman, W. Mason, and W. D.J. Everyone's an influencer: quantifying influence on Twitter. In *Proceedings of the 4th ACM WSDM*, pages 65–74, 2011.

[3] R. Bandari, S. Asur, and B. Huberman. The pulse of news in social media: forecasting popularity. In *Proceedings of the 6th AAAI ICWSM*, 2012.

[4] H. Becker, M. Naaman, and L. Gravano. Selecting quality Twitter content for events. In *Proceedings of the 2011 AAAI ICWSM*, 2011.

[5] M. Bernstein, L. Hong, S. Kairam, H. Chi, and B. Suh. A torrent of tweets: managing information overload in online social streams. In *In Workshop on Microblogging: What and How Can We Learn From It?(CHI'10*. Citeseer, 2010.

[6] G. Beverungen and J. Kalita. Evaluating methods for summarizing Twitter posts. In *Proceedings of the 5th AAAI ICWSM*, 2011.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in Twitter: the million follower fallacy. In *Proceedings of the 2010 AAAI ICWSM*, 2010.

[8] J. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[9] O. Dalal, S. H. Sengemedu, and S. Sanyal. Multi-objective ranking of comments on web. In *Proceedings of the 21st ACM WWW*, pages 419–428, 2012.

[10] M. De Choudhury, S. Counts, and M. Czerwinski. Find me the right content! Diversity-based sampling of social media spaces for topic-centric search. In *Proceedings of the 5th AAAI ICWSM*, 2011.

[11] M. De Choudhury, H. Sundaram, A. John, and D. Seligmann. What makes conversations interesting?: Themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th ACM WWW*, pages 331–340, 2009.

[12] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using Twitter data. In *Proceedings of the 19th ACM WWW*, pages 331–340, 2010.

[13] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Transactions on the Web (TWEB)*, 1(2):7, 2007.

[14] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[15] K. Ganesan, C. Zhai, and E. Viegas. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st ACM WWW*, pages 869–878, 2012.

[16] A. Kelmans and B. Kimelfeld. Multiplicative submodularity of a matrix's principal minor as a function of the set of its rows and some combinatorial applications. *Discrete Mathematics*, 44(1):113–116, 1983.

[17] C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th ACM WWW*, pages 591–600, 2010.

[19] C.-Y. Lin. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the 2004 ACL Workshop on Text Summarization Branches Out*, pages 74–81, 2004.

[20] F. Liu, Y. Liu, and F. Weng. Why is "SXSW" trending? Exploring multiple text sources for Twitter topic summarization. In *Proceedings of the 11th ACL HLT*, pages 66–75, 2010.

[21] C. Long, M. Huang, X. Zhu, and M. Li. A new approach for multi-document update summarization. *Journal of Computer Science and Technology*, 25(4):739–749, 2010.

[22] Z. Luo, M. Osborne, and T. Wang. Opinion retrieval in Twitter. In *Proceedings of the 6th AAAI ICWSM*, 2012.

[23] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[24] M. Naaman, J. Boase, and C. Lai. Is it really about me? Message content in social awareness streams. In *Proceedings of the 2010 ACM CSCW*, pages 189–192, 2010.

[25] N. Naveed, T. Gottron, J. Kunegis, and A. Alhadi. Bad news travel fast: a content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Conference on Web Science*, 2011.

[26] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[27] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM, 2012.

[28] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of the 18th ACM CIKM*, pages 365–374, 2009.

[29] A. Popescu and M. Pennacchiotti. Detecting controversial events from Twitter. In *Proceedings of the 19th ACM CIKM*, pages 1873–1876. ACM, 2010.

[30] A. Popescu and M. Pennacchiotti. "Dancing with the Stars", NBA games, politics: an exploration of Twitter users' response to events. In *Proceedings of the 5th AAAI ICWSM*, 2011.

[31] B. Sharifi, M. Hutton, and J. Kalita. Experiments in microblog summarization. In *Proceedings of the 2nd IEEE International Conference on Social Computing*, pages 49–56, 2010.

[32] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[33] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li. Social context summarization. In *Proceedings of the 34th ACM International Conference on Research and Development in Information Retrieval*, pages 255–264, 2011.

[34] S. Ye and S. F. Wu. Measuring message propagation and social influence on Twitter.com. In *Proceedings of the 2nd International Conference on Social Informatics*, pages 216–231, 2010.

[35] F. Zanzotto, M. Pennacchiotti, and K. Tsioutsiouliklis. Linguistic rudundancy in Twitter. In *Proceedings of the 2011 EMNLP*, 2011.