Selecting Social Media Responses to News: A Convex Framework Based On Data Reconstruction

Zaiyi Chen[*] Linli Xu[*] Enhong Chen[*] Biao Chang[*] Zhefeng Wang[*] Yitan Li[*]

Abstract

With the explosive growth of social media, it has gained significantly increasing attention from both journalists and their readership in recent years by enhancing the reading experience with its timeliness, high participation, interactivity, etc. On the other hand, the popularity of social media services such as Twitter also leads to the challenge of information overload by generating thousands of responses (tweets) for each article of hot news, which will be overwhelming for readers. In this paper, we address the problem of selecting a representative subset of responses to news in order to deliver the most *important* information. We consider different criteria regarding the importance of the selected subset, and treat the problem from the data reconstruction perspective with concerns for both quality and generalizability of the selection. The intuition behind our work is that a good selection should be relevant from two levels: i) at the message level, it brings readers new information as much as possible or generalizes other people's opinions comprehensively; ii) at the text level, it is able to reconstruct the corpus. Specifically, the task of selecting responses to news can be formulated as a convex optimization problem where sparse non-negative weights are introduced for all the responses indicating whether they are selected or not. Several gradient based optimization and step size selection methods are also investigated in this paper to achieve a faster rate of convergence. More importantly, the proposed framework evaluates the utility of a set of responses jointly and therefore is able to reduce redundancy of the selected responses. We evaluate our approach on real-world data obtained from Twitter, and the results demonstrate superior performance over the state of the art in both accuracy and generalizability.

1 Introduction

As an open platform, social media has provided popular channels for people to share information and convey opinions on a wide range of topics and events. The openness of social media facilitates communication and enhances information discovery and delivery. At the same time, the immediacy of social media which can produce instantaneous responses is highly desirable compared to industrial media whose time lag could be days or weeks. As a consequence, social media is extensively used for real-time broadcast and discussion of important events. On Twitter, a significant proportion of tweets are posted on news events [12]; for instance, over 48 hours in March 2014, more than 37 million people viewed 19.1 million Oscars tweets across Twitter.

On the other hand, the popularity of social media services also leads to huge volume of messages posted in the context of certain news, which are not only overwhelming for readers to track, but also damage the immediacy of social media. This challenge motivates devising methods towards effective selecting and displaying a representative subset of responses to news which readers would like to read.

To achieve a good selection of responses, one needs to first evaluate the quality of the responses. Intuitively, at the message level, the selected messages should contain either new *information* which readers are interested in or *opinions* which they would like to argue for or against. It is worth noting that the significance of information and opinions may vary depending on the nature of the events, and should be considered separately. Hence, the techniques proposed in this area can be roughly categorized into two groups: *information based summarization* which selects the most representative responses of each informative topic of corpus [21, 4, 9]; and *opinion based summarization* which aims at summarizing users' opinions of a specific item or breaking news [14, 18]. However, most of the techniques in these two groups are too different to be put into the same framework, which implies a possible hindrance if one wants to consider the two criteria at the same time. Recently, a diversity maximization based summarization framework has been proposed in [22], which incorporates information and opinion summarization into the objective and solves the problem approximately with a greedy algorithm by exploiting the submodularity of the objective function.

In this paper, we treat the *response selection* task from two levels and propose a new convex optimization

[*]University of Science and Technology of China. czy6516@mail.ustc.edu.cn, {linlixu, cheneh}@ustc.edu.cn, {zhefwang, chbiao, etali}@mail.ustc.edu.cn

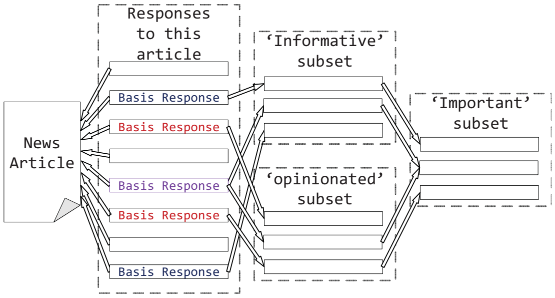


Figure 1: The procedure of response selection for a news article.

based approach. Specifically, by mining the real world events and all the messages in response to them, we observe that people are more inclined to read "*important*" responses which they would like to repost or reply. The "*importance*" should be measured not only from the message level but also from the text level. At the message level, the "*important*" responses should contain either *information* or *opinions* as discussed above. In addition, at the text level, the quality of summary should be taken into consideration since a good selection should provide a global perspective of all responses. To this end, the *response selection* problem can be defined here as follows: given all the responses referring to a certain news article, the system automatically selects a subset of responses that represent the origin corpus comprehensively and the items in the subset should be the most *important* ones. This procedure is also illustrated in Figure 1, where *importance* consists of two separate indicators – *informativeness* and *opinionatedness*, representing information and opinion based summarization respectively. For each indicator, the quality measure of the selection includes two components: the utility scores of *informativeness* or *opinionatedness* at the message level, and the quality of summarization of the corpus at the text level.

To achieve that, we introduce sparse non-negative weights for all the responses indicating whether they are selected or not, and the goal is to efficiently optimize an objective that integrates the message-level utility scores and the text-level quality of summarization simultaneously. The utility scores are learned based on textual and personalized features of individual responses; in the meantime, to measure the quality of summarization, we take a data reconstruction perspective and formulate a convex weighted non-negative linear reconstruction inspired by the methodology proposed for active learning [25]. Importantly, the proposed framework evaluates a set of responses jointly and therefore is able to reduce redundancy of the selected responses. We further investigate different gradient based algorithms and analyze the corresponding convergence behaviors to solve the optimization problem efficiently.

The contributions of this paper are fourfold. i) We analyze users' behavior and redefine the task of selection of responses to news. ii) Unlike the greedy or heuristic algorithms discussed above, the problem is formulated as a convex optimization framework which can be solved efficiently with convergence to global optimum guaranteed. iii) The redundancy is implicitly reduced by considering the responses jointly. iv) A significant improvement in accuracy and quality of summarization on real-world data is achieved.

The remainder of the paper is organized as follows. After discussing related work in Section 2, the proposed method is presented in Section 3 followed by the optimization techniques introduced in Section 4. Experiments are conducted in Section 5 to demonstrate the effectiveness of the proposed method. The paper is concluded in Section 6.

2 Related Work

In this paper we evaluate the quality of summarization at the text level from a data reconstruction perspective, following a recipe proposed for active learning in [25]. This principle is also employed in the document summarization problem [11]. The general methodology can be viewed as a sparse coding process which selects the most representative bases spanning the linear subspace of the dictionary while optimizing the number and positions of non-zeros in the sparse representation to minimize the reconstruction loss.

On the other hand, there exists extensive work on social media sampling and summarization, with the goal to select a representative subset of messages on various items including a given question [1], topic [4], product [19] or event [6]. The idea of importance is originally associated with the work on predicting reposts of messages [16]. A more recent and relevant piece of work is proposed in [22] that summarizes *interesting* messages in response to social media. Interestingness consists of a few message-level indicators as well as a set-level indicator "diversity" which measures the normalized joint entropy of the set. The objective function is then designed as a sum of the message-level utility scores and the set-level diversity. Given the hardness of exhaustively searching the space of all possible message subsets, the submodularity of the objective function is exploited and a greedy algorithm is proposed.

Overall, our work combines aspects of both data reconstruction and social media sampling. The proposed method is not only able to leverage the social information to enhance the performance, but also optimize the objective function globally by translating the original problem to a basis selection problem. In addition, we employ techniques on sentiment analysis [20, 14] and redundancy detection [26] to get rich features.

3 Proposed Framework

We first start with the formal definition of the response selection problem. Given a news article and a set of responses $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ where $\mathbf{x}_i \in R^d$ is a term-frequency vector, we want to find a subset $S = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m] \subseteq X$ of m responses, which are the most "important" to a typical reader of the article.

The *importance* of the selected subset is measured with the *informativeness* and *opinionatedness* indicators as well as the quality of data reconstruction. To achieve that, we impose a utility function for each indicator at the message level while formulating a weighted non-negative linear reconstruction at the text level.

In this section, we will first build a model considering each indicator independently, and then extend it to a framework of selecting "important" responses by integrating the *informativeness* and *opinionatedness* indicators simultaneously.

3.1 Text-Level Data Reconstruction

At the text level, we want to find an optimal subset of representative responses $S \subseteq X$ such that any response in X can be reconstructed with S.

For each indicator, the selected subset of responses S can be denoted by a binary indicator vector $\boldsymbol{\beta} \in \{0,1\}^n$: when $\beta_j = 1$, the jth response will be selected. A given response $\mathbf{x}_i$ can then be represented with a non-negative linear combination of selected responses in S:

$$\mathbf{x}_i = \sum_{j=1}^{n} \mathbf{x}_j \beta_j a_{ij} + \boldsymbol{\epsilon}_i = X \operatorname{diag}(\boldsymbol{\beta})\, \mathbf{a}_i + \boldsymbol{\epsilon}_i \tag{3.1}$$

where $\boldsymbol{\beta}$ is an overall control of whether a response is selected or not, $\mathbf{a}_i \geq 0$ is the column coefficient vector of length n regarding the linear reconstruction for $\mathbf{x}_i$, $\boldsymbol{\epsilon}_i \in R^n$ is assumed to be i.i.d. Gaussian noise. This non-negative linear reconstruction allows only additive combination of the responses, which implicity minimizes redundant information in data representation [17, 23].

We can then learn the selection of responses $\boldsymbol{\beta}$ and the coefficients $\{\mathbf{a}_i\}$ by minimizing the sum of reconstruction errors of all the responses, which can be formulated as

$$\begin{aligned} \min_{A,\boldsymbol{\beta}} \ \mathcal{L}(A,\boldsymbol{\beta}) &= \sum_{i=1}^{n} ||\mathbf{x}_i - X \operatorname{diag}(\boldsymbol{\beta})\, \mathbf{a}_i||^2 \\ &= \|X - X\operatorname{diag}(\boldsymbol{\beta})A^\top\|_F^2 \\ \text{s.t.} \quad \boldsymbol{\beta} \in \{0,1\}^n, &\quad \sum_{i=1}^{n} \beta_i = m \quad \text{and} \quad A \geq 0, \end{aligned} \tag{3.2}$$

where $\| \cdot \|$ is the ℓ_2 norm of a vector, $|| \cdot ||_F$ is the Frobenius norm of a matrix and $A = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n]^\top$. Note that the selection learned by solving (3.2) is independent of the message-level indicators. Next to incorporate the message-level information, we exploit the utility score functions as weighted combinations of all the responses regarding the contributions to the two indicators – *informativeness* and *opinionatedness*.

3.2 Message-Level Utility Scoring

To select the "good" responses in terms of *informativeness* or *opinionatedness*, we maximize the sum of normalized utility scores of the selected responses. For each indicator, the vector of utility scores $\mathbf{u}$ is generated with a scoring function learned from responses and a large number of features. In this paper, we mainly focus on Twitter specific features, however the way we extract them can be generalized to other social media platforms. Table 1 lists all the features used in this paper to compute the utility scores which include:

- *textual features* that capture the linguistic characters which indicate the quality of expression.
- *opinion features* that represent the sentiment orientation of the owner of a tweet.
- *social features* that capture the ability of diffusion of a response and reflect a user's relationship in the social network.

To learn the utility scoring function, given the above features we train Support Vector Regression ($\epsilon - SVR$ [8]) models on tweets manually labeled on the indicators of *informativeness* and *opinionatedness* as well as *importance*. After getting the utility scores $\mathbf{u}$, the message-level objective for an indicator is to select the responses with the highest utility scores:

$$\begin{aligned} &\max_{\boldsymbol{\beta}} \ \mathcal{U}(\boldsymbol{\beta}) = \mathbf{u}^\top \boldsymbol{\beta} \\ &\text{s.t.} \quad \boldsymbol{\beta} \in \{0,1\}^n, \quad \sum_{i=1}^{n} \beta_i = m. \end{aligned} \tag{3.3}$$

To integrate the minimization problem (3.2) and maximization problem (3.3), we first switch the utility score $\mathbf{u}$ to $\mathbf{u'}$ without affecting the solution:

$$u'_i = \frac{u_{\max} - u_i}{u_{\max}} + \sigma \tag{3.4}$$

where $u_{\max}$ is the maximum of vector $\mathbf{u}$ and σ is a small positive constant to ensure that $\mathbf{u'}$ is positive. The optimal selection $\boldsymbol{\beta}$ for an indicator can then be obtained by solving the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta},A} \mathcal{L}(A,\boldsymbol{\beta}) - \lambda\, \mathcal{U}(\boldsymbol{\beta}) \\ &\quad = \|X - X\operatorname{diag}(\boldsymbol{\beta})A^\top\|_F^2 + \lambda \mathbf{u'}^\top \boldsymbol{\beta} \\ &\text{s.t} \quad \boldsymbol{\beta} \in \{0,1\}^n \ , \ \sum_{i=1}^{n} \beta_i = m, \quad A \geq 0 \end{aligned} \tag{3.5}$$

Table 1: The features used in the computation of utility scores

Textual features	Tf-idf score	Average tf-idf score of all words in the tweet, emphasizing rarely-used and penalizing out-of-vocabulary words
	Log-likelihood	Likelihood of the tweet, based on a bigram language model constructed from all of the tweets of the article.
	Number of words	A higher number of words may indicate a tweet with more useful information.
	First person pronoun information	Indicate if the tweet is mainly about the author himself. It's useful when analyzing user's sentiment orientation [15].
	Question	Judge whether a tweet contains questions.
	Quote sharing	Identify whether a tweet has an additional quotation which is likely to bring new information or express opinion.
	Proportion of words, hashtags, capitalized characters	Identify the quality of content.
	Repetitions	The number of repetitions of words. Identify the quality of content.
Opinion features	Proportion of positive and negative words	Identify the sentiment orientation.
	Mixed sentiment score	Identify the polarity of a tweet. It depends not only on the number of sentimental words, but also on their intensity [20].
Social features	Location	Whether a geographic location mentioned in the tweet or the location of user.
	Retweet or reply flag	Whether the tweet is a retweet or a reply.
	Followers	The number of followers and the number of friends. (Two users are friends means they follow each other.)
	Follower-friend ratio	Ratio between number of followers and friends of a user.
	Number of replies, retweets and favourites	The total number of replies, retweets and favourites of a tweet.
	Number of tweets, retweets and favourites	The total number of tweets, retweets and favourites of a user.
	Tweet-retweet ratio	Ratio between tweets and retweets of a user.
	User verified	Indidate if the user is verified by Twitter, which may increase the credibility of the user's posts.

where λ is a parameter controlling the relative significance of the message-level and text-level criteria.

Due to the discrete constraints on $\boldsymbol{\beta}$, the problem (3.5) is still difficult to optimize. However we can relax $\boldsymbol{\beta}$ to be continuous, and reformulate (3.5) as

$$\begin{aligned} &\min_{\boldsymbol{\beta},A} \sum_{i=1}^{n}\{||\mathbf{x}_i - X\mathbf{a}_i||^2 + \sum_{j=1}^{n} \frac{{a_{ij}}^2}{\beta_j}\} + \lambda \mathbf{u}'^{\top}\boldsymbol{\beta} \\ &= ||X - XA^{\top}||_F^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{{a_{ij}}^2}{\beta_j} + \lambda \mathbf{u}'^{\top}\boldsymbol{\beta} \\ &\text{s.t. } \beta_i \geq 0\ ,\ \ A \geq 0. \end{aligned} \tag{3.6}$$

The last term in (3.6) $\mathbf{u}'^{\top}\boldsymbol{\beta}$ works like a weighted ℓ_1 norm of $\boldsymbol{\beta}$ since the utility scores in $\mathbf{u}'$ are positive, enforcing some elements in $\boldsymbol{\beta}$ to be 0. When $\beta_j = 0$, then $a_{1j}, ...a_{nj}$ must be 0, which implies the jth response is not selected.

The problem (3.6) is convex regarding $\boldsymbol{\beta}$ and A [25], which guarantees a global optimal solution. By fixing A and setting the derivative of the objective regarding $\boldsymbol{\beta}$ to be 0, we can get the closed-form solution of $\boldsymbol{\beta}$:

$$\beta_j = \sqrt{\frac{\sum_{i=1}^{n} {a_{ij}}^2}{\lambda\, u_j}}, \tag{3.7}$$

which establishes the connection that β_j is proportional to $||A_{:,j}||_2$ and in inverse proportion to u_j.

3.3 The Unified Framework Integrating All the Indicators

The model formulated in (3.6) deals with one indicator, *informativeness* or *opinionatedness* independently. On the other hand, our major task is to select a subset of responses according to *importance*, which combines aspects described by *informativeness* and *opinionatedness* simultaneously. When modeling *importance*, we need to take the possible relationship among the indicators into consideration. In principle, an *important* response should likely contain both *informative* and *opinionated* characteristics at the same time. To achieve that, we integrate all the indicators by imposing a joint sparsity regularization term.

Specifically, we denote K sets of selections by $\{\boldsymbol{\beta}^1, A^1\}, \{\boldsymbol{\beta}^2, A^2\}...\{\boldsymbol{\beta}^K, A^K\}$. In our case with indicators *importance*, *informativeness*, *opinionatedness*, $K =$

3. For all the indicators we need to first integrate their independent objective functions by $\sum_{k=1}^{K} \mathcal{F}(A^k, \boldsymbol{\beta}^k) = \sum_{k=1}^{K}(\mathcal{L}(A^k, \boldsymbol{\beta}^k) - \lambda\, \mathcal{U}(\boldsymbol{\beta}^k))$. Next to combine the information of responses across all indicators, we apply an ℓ_2 norm over the rows of $B = [\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, ..., \boldsymbol{\beta}^K]$. In the mean time, to achieve that similar selections are produced for all the indicators, we need to promote joint row sparsity of B. Therefore we adopt an $\ell_{1,2}$ norm as a regularization on B, which is convex and known as "group lasso" by facilitating row sparsity – each row is either all zero or mostly non-zero, while the number of non-zero rows is small.

Thus, we arrive at the unified framework of response selection based on Joint Weighted Non-negative Linear Reconstruction (JWNLR):

$$\begin{aligned}
(3.8) \qquad & \min_{\{A^k\},B} \sum_{k=1}^{K} \mathcal{F}(A^k, \boldsymbol{\beta}^k) + \gamma \|B\|_{1,2} \\
&= \min_{\{A^k\},B} \sum_{k=1}^{K} (\mathcal{L}(A^k, \boldsymbol{\beta}^k) - \lambda\, \mathcal{U}(\boldsymbol{\beta}^k)) + \gamma \|B\|_{1,2} \\
&\text{s.t.} \quad A^k \geq 0, B \geq 0
\end{aligned}$$

In this framework, the closed-form solution of $\boldsymbol{\beta}^k$ cannot be easily found. However, the problem is still convex which can be solved globally with iterative algorithms.

4 Optimization Methods

The problem (3.8) can be solved by alternative optimization over $\{A^k\}$ and B. To design an efficient algorithm, we investigate different gradient-based optimization methods and step size selection strategies. We use a Constrained Newton's Method and two gradient based algorithms to update B and A by rows respectively. Analysis and comparison of the convergence behavior of the discussed algorithms is also included in this section.

For convenience, we define $F(*)$ as the value of the objective function $\sum_{k=1}^{K} \mathcal{F}(A^k, \boldsymbol{\beta}^k) + \gamma ||B||_{1,2}$ with respect to *, and ∇F_* as the derivative of the objective function with respect to *; $A \in \{A^k\}$, $\boldsymbol{\beta} \in \{\boldsymbol{\beta}^k\}$, $\mathbf{a}$ and $\mathbf{b}$ denote column vectors of $A^\top$ and $B^\top$ or row vectors of A and B respectively.

The constrained Newton's Method (CNM) [5], is one of the fastest iterative algorithms when the subproblem of finding the quadratic direction is simple and the dimension of variable is small. We first use CNM to find the optimal update for a row vector $\mathbf{b}$ of B as follows:

$$\begin{aligned}
(4.9) \qquad \mathbf{b}^{t+1} = \arg\min_{\mathbf{b} \geq 0} \{ & \nabla F_{\mathbf{b}^t}^\top (\mathbf{b} - \mathbf{b}^t) \\
& + (\mathbf{b} - \mathbf{b}^t)^\top \nabla^2 F_{\mathbf{b}^t} (\mathbf{b} - \mathbf{b}^t) \}
\end{aligned}$$

where $\nabla^2 F_{\mathbf{b}}$ is the Hessian matrix with respect to $\mathbf{b}$. In practice, CNM might converge unacceptably slowly when a good starting point is unknown. As a result, we use the Armijo rules mentioned in Section 4.2 in the first few steps to improve the convergence properties.

Algorithm 1 Response Selection via Weighted Non-negative Linear Reconstruction (JWNLR)

```
Input: Set of all responses $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$; $\lambda$, $\gamma$
Output: Set of selected responses: $S \subseteq X$
Initialize $t$, $c^0$ to 1; $\mathbf{a}^0$ to $[1, 1, ..., 1]^\top$
repeat
    Update $B^t$ according to equation (4.9);
    for all $\mathbf{a}^t$ do
        i) Update $\mathbf{a}^t$ according to equation (4.11);
        ii) Choose $d^t$ by equation (4.12) or (4.13), and
        update $\mathbf{a}^t$ by euqation (4.14);
    end for
until converge;
$S \leftarrow \{\mathbf{x}_i \mid \mathbf{x}_i \in X, \beta_i$ is in top-$k$ non-zero elements
of $\boldsymbol{\beta}\}$;
```

Another issue when optimizing A is that CNM is not very effective for high dimensional problems. Here we take advantage of the Multiplicative Update approach (MU) and widely used Gradient Projection approach (GP) with two strategies of stepsize selection [10] to solve the problem with respect to A. Both approaches guarantee that the limit point of any convergent sequence generated by them is a global minimizer [5].

4.1 Multiplicative Update Algorithm Multiplicative update is widely used to solve the bound-constrained quadratic program (BCQP). The update rule is obtained by analyzing the Karush-Kuhn-Tucker (KKT) conditions [7] of the problem, which are:

$$(4.10) \qquad \nabla F_A = \Theta, \;\; a_{ij} \geq 0, \;\; \theta_{ij} \geq 0 \text{ and } \; \theta_{ij} a_{ij} = 0$$

where $\Theta = [\theta_{ij}]$ is the Lagrange multiplier for A.

Substituting Θ in $\theta_{ij} a_{ij} = 0$ with $\nabla F_A = \Theta$, we can get the MU rule for a_{ij}:

$$(4.11) \qquad {a_{ij}}^{t+1} = \left[\frac{(X^\top X)_{ij}}{(A^t X^\top X + A^t \mathrm{diag}(\boldsymbol{\beta})^t)_{ij}} \right] a_{ij}^t$$

While the multiplicative update rule is parameter free and liberates people from tuning parameters, the convergence rate is not promised although in practice it is acceptable most of the time. In addition, following equation (4.11), entry a_{ij} with $(X^\top X)_{ij} \neq 0$ would get to 0 only after infinite number of iterations, which is unrealistic and would hurt the desired group sparse property seriously, while enforcing entries with small values to 0 straightforwardly might damage the precision in general. On the other hand, the gradient projection approach provides guarantees of convergence rate and has access to group sparsity. However if a bad step size is chosen, it might converge very slowly or even never converge. Here we leverage the recent advances in step size selection and simplify the parameter searching process to achieve a quick convergence.

4.2 Accelerated Gradient Projection Algorithm The subproblem of solving for A in the optimization problem (3.8) can also be tackled by the accelerated gradient projection (AGP) algorithm, which updates the row vectors of A separately. To choose the step size d for gradient descent, we have the following two options.

Option 1: At the iteration t, given the initial value of the step size d_0^t, *Armjio rule* chooses a step size d^t which is the first number in the sequence $d_0^t, \eta d_0^t, \eta^2 d_0^t, ...$ satisfying:

$$\begin{aligned} & F((\mathbf{a}^t - d^t \nabla F_{\mathbf{a}^t})_+) \leq \\ & F(\mathbf{a}^t) - \frac{1}{2} \nabla F_{\mathbf{a}^t}{}^\top (\mathbf{a}^t - (\mathbf{a}^t - d^t \nabla F_{\mathbf{a}^t})_+), \end{aligned} \tag{4.12}$$

where d_0^t is determined by $\frac{(\nabla F_{\mathbf{a}^t})^\top \nabla F_{\mathbf{a}^t}}{(\nabla F_{\mathbf{a}^t})^\top (X^\top X + \text{diag}(\boldsymbol{\beta})^{-1}) \nabla F_{\mathbf{a}^t}}$, and $(\mathbf{a}^t - d^t \nabla F_{\mathbf{a}^t})_+$ denotes the projection of $\mathbf{a}^t - d^t \nabla F_{\mathbf{a}^t}$ in the convex set $\{\mathbf{a} | \mathbf{a} \geq 0\}$.

Option 2: After the first iteration, *Barzilai-Borwein rule* chooses a step size to approximate the Hessian matrix:

$$\begin{aligned} & \boldsymbol{\delta}^t = \mathbf{a}^t - \mathbf{a}^{t-1}, \; \boldsymbol{\zeta}^t = \nabla F(\mathbf{a}^t) - \nabla F(\mathbf{a}^{t-1}) \\ & d^t = \arg\min_d \|\boldsymbol{\delta}^t - d\boldsymbol{\zeta}^t\|^2 = \frac{\|\boldsymbol{\delta}^t\|_2^2}{\langle \boldsymbol{\delta}^t, \boldsymbol{\zeta}^t \rangle} \end{aligned} \tag{4.13}$$

Among the above two options, the Armjio rule provides a reasonable initial value and guarantees that the optimization procedure descents monotonously over steps. The Barzilai-Borwein does not have this property, but the convergence is still promised [2].

We further adopt the Nesterov's acceleration [3] to speed up the procedure of gradient projection:

$$\begin{aligned} \boldsymbol{y}^{t+1} &= (\mathbf{a}^t - d^t \nabla F(\mathbf{a}^t))_+, \\ c^{t+1} &= \frac{1 + \sqrt{1 + 4(c^t)^2}}{2}, \\ \mathbf{a}^{t+1} &= \mathbf{a}^t + \frac{c^t - 1}{c^t} \boldsymbol{y}^k. \end{aligned} \tag{4.14}$$

The accelerated gradient projection above achieves a convergence rate of $O(\frac{1}{t^2})$.

By optimizing A and B alternately, we could find the global optimal solution of problem (3.8). The overall procedure is described in Algorithm 1. In each iteration, given the number of responses n, the complexity of updating $\mathbf{b}$ is $O(K^3)$, and the complexity of updating a certain $\mathbf{a}$ is $O(n^2)$ no matter which optimization method is chosen. Assuming the numbers of iterations for computing A and B are t_1 and t_2 respectively, the total computational cost for Algorithm 1 is $O(t_1 K n(n^2 + t_2 K^2))$.

5 Experiments

In this section we empirically evaluate the performance of the proposed framework on selecting social media responses to news. We obtain the data set with Twitter[1] search API, which consists of 26 news articles from Wall street journal[2], BBC[3], New York Times[4] and 20,609 tweets annotated by editors with the annotation checked repeatedly for consistency. It is worth noting that our method can be applied to selecting responses to news on other platforms without loss of generality.

5.1 Gold Standard Collection and Analysis On the obtained data set, we build a gold standard collection in a consistent way as follows: i) For the *informative* indicator, a tweet is assigned with score 1 by an annotator if it contains new information to which readers may pay attention, 0 otherwise; for the *opinionated* indicator, a tweet is assigned with a score ranging from 3 (the tweet significantly expresses the writer's opinion with strong sentiment words or explanation) through 2 (the tweet expresses the writer's opinion but not strongly enough) to 1 (the tweet is neutral and contains no opinion of the writer); further, we assign score 1 to tweets which annotators think *important*, 0 otherwise. ii) All tweets that are repetitive or irrelevant to the article are removed from the collection.

Overall, for the *informativeness* indicator, about 20% of scores are 1s; while for the *opinionatedness* indicator, around 10%, 15% of the scores are 3, 2 respectively, as shown in Figure 2(a). In the meantime, for *importance*, only less than 10% of the scores are positive, which implies only a small proportion of the tweets are worth reading in practice for the annotators. Figure 2(b) depicts the distribution of tweets labeled as *important*, which shows that only 4% of *important* messages are scored 0 for both *informativeness* and *opinionatedness*, which is reasonable; while a majority of *important* responses are scored positively as *informative* and *opinionated* simultaneously. On the other hand, we observe that *informativeness* and *opinionatedness*

[1] `http://twitter.com`
[2] `http://online.wsj.com`
[3] `http://www.bbc.com`
[4] `http://nytimes.com`

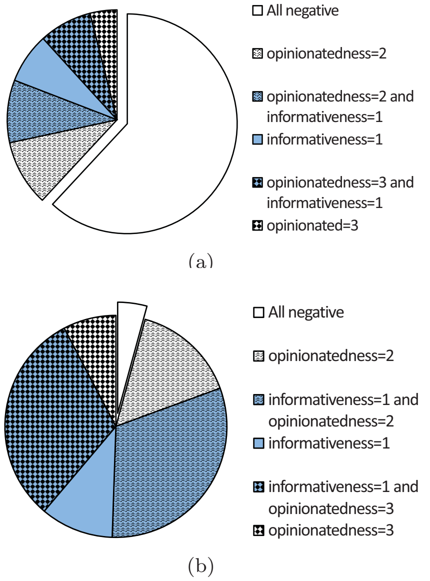


Figure 2: The pie-charts of *informativeness* and *opinionatedness*: (a) Distribution of the whole dataset; (b) Distribution of all the *important* tweets.

make different contributions to *importance* which depends significantly on the nature of the news. For instance, a controversial topic[5] which focuses on a controversial country or people is extremely *opinion* oriented, as 14% of *important* tweets are labeled with 1 for *informativeness* and 93% of them are labeled with 3 for *opinionated*, while a different news topic [6] has 90% and 47% of *important* tweets labeled positive for *informativeness* and *opinionated* respectively. In brief, the analysis above agrees with our underlying intuition that *informativeness* and *opinionatedness* have different impacts on *importance* while still sharing joint patterns when considered together.

5.2 Comparison with Baselines We now empirically investigate the performance of the proposed method (JWNLR) on the obtained data with gold standard annotations, and compare with the state of the art. Experiments are conducted on each subtask of *informativeness* and *opinionatedness*, as well as on the integrated task of *importance*. For *informativeness* and *importance*, positives and negatives correspond to scores 1's and 0's respectively. For *opinionatedness* negatives correspond to score 0's and all others are positive. A 10-fold cross-validation is performed to eliminate contingency, with nine folds for training and one for testing.

Baselines: In the experiments, we compare with two state of the art methods including the previously discussed SVR_ENTROPY method [22] which translates summarization into diversity maximization, and optimizes the objective with a greedy algorithm. We also compare to DWFG [24] which is based on conditional random fields and simultaneously treats messages and web documents as 'wings' in a dual wing factor graph, where factors are assigned to individual tweet and sentence features in the graph. A standard sum-product algorithm which is an approximate and relatively time-consuming inference approach is used to determine the key sentences and the important messages. In addition, we compare with the results obtained by only performing support vector regression ($\epsilon - SVR$) [8] on the utility scores.

Metrics: Two metrics are used to measure the performance. Specifically, we compute the $\mathbf{F}_1$ scores for the *top*10 selections and utilize the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit [13] which has been widely applied in automatic summarization evaluation. ROUGE-N is computed as follows:

(5.15)

$$\mathrm{ROUGE-N} = \frac{\sum_{S \in \mathrm{Ref}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)}{\sum_{S \in \mathrm{Ref}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}(\mathrm{gram}_n)}$$

where n is the length of the n-gram, Ref is the set of reference summaries. $\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)$ is the maximum number of n-grams co-occurring in the candidate summary and the set of reference summaries, while $\mathrm{Count}(\mathrm{gram}_n)$ is the number of n-grams in the reference summaries.

Results: ROUGE can generate three types of scores: precision, recall and F-measure. In this study, we use F-measure to compare our method with baselines on each subtask of *informativeness* and *opinionatedness*, as well as on the main task of *importance* where we employ the integrated framework in (3.8). The results are shown in Table 2, where we can observe that the proposed convex approach based on joint weighted non-negative linear reconstruction (JWNLR) outperforms the other baselines most of the times. In addition, the last column of Table 2 demonstrates a significant advantage of the proposed method over SVR_ENTROPY on *importance*, which implies the suboptimality of the greedy algorithm. The results also show that our method produces higher $\mathbf{F}_1$ scores than $\epsilon - SVR$ which justifies that considering text-level and message-level information simultaneously could integrate more information and improve the performance of summarization.

We also present the results of our method on a specific news article, *Clue to earthquake lightning mys-*

[5] http://www.bbc.com/news/technology-26071818

[6] http://online.wsj.com/news/articles/SB10001424052702304851104579361451951384512

Table 2: The average F-measures of ROUGE-N and ROUGE-L, and F_1 scores of top 10 selections

	Informativeness				Opinionatedness				Importance			
	R-1	R-2	R-L	F1	R-1	R-2	R-L	F1	R-1	R-2	R-L	F1
DWFG	0.211	0.193	0.207	0.119	0.340	0.273	0.338	0.179	0.271	0.194	0.269	0.107
$\epsilon - SVR$	0.161	0.132	0.141	0.136	0.288	0.231	0.267	**0.256**	0.306	0.225	0.297	0.168
SVR_ENTROPY	**0.229**	**0.199**	0.228	0.139	0.345	0.282	0.343	0.252	0.345	0.282	0.343	0.153
JWNLR	0.213	**0.199**	**0.233**	**0.157**	**0.353**	**0.287**	**0.351**	**0.256**	**0.402**	**0.285**	**0.394**	**0.227**

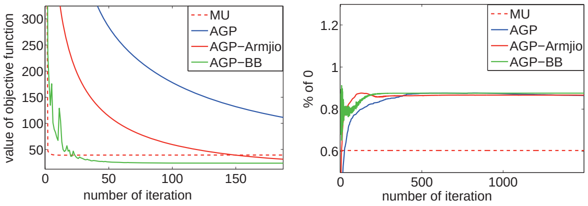
Table 3: Tweets selected by our method in response to a specific news

label	content	label
1	my assumption has been that the glow was generated by friction as the fault lines ground passed each over creating static	1
2	Mulayam singh and kejriwal are two sides of same coin, one spreading gundaraj in UP the other in Delhi	0
3	i did once witness what i assume was a 'seismic glow', accompanied by a deep undulating hum. Any one ells ?	1
4	it would make sense that such a low frequency hum might vibrate soil particles creating the accompanying glow	1
5	Does this mean power stations in the future will just be warehouses of flour rocking back & forth? :)	1

tery[7], to visualize the practical results. Due to the space limit, we exhibit the top 5 tweets selected by our method, as shown in Table 3. It shows that all of them are well expressed and 4 of them are labeled positive as *important*. At the same time, these tweets are semantically different, which indicates that the non-negative reconstruction does work and reduces the redundant information. This suggests the competence of the *Joint Non-negative Linear Reconstruction* method in the social context summarization task.

5.3 Comparison of Optimization Algorithms

To solve the optimization problem (3.8) we propose a multiplicative update (MU) rule and an accelerated gradient projection update with two stepsize selection methods, Armjio rule (AGP-Armjio) and Barzilai-Borwein rule (AGP-BB) respectively. In this subsection, we will empirically investigate and analyze these methods from the perspectives of convergence and sparsity of solution.



(a) (b)

Figure 3: (a) Objective value, (b) Percentage of entries equal to 0 in matrix A, of MU, GP method with fixed stepsize and AGP approach with two stepsize selection strategies (AGP-Armjio and AGP-BB).

As shown in Figure 3(a), the multiplicative update rule converges very fast within the first few steps, however the precision and quality of the solution is lower than the other algorithms. We also try to enforce entries with small values to 0, which may increase the objective value and does not produce much improvement in group sparsity. It is also shown in Figure 3(b) that MU does not perform very well to achieve group sparsity.

On the other hand, the accelerated gradient projection algorithms with Armjio rule and Barzilai-Borwein rule outperform the basic gradient projection approach even if the parameters are not carefully chosen. Remarkably, AGB-BB requires no parameter other than c^0 for acceleration, and it achieves the best performance on sparsity, and is comparable to MU in convergence rate with a lower objective value.

6 Conclusion

In this paper, we revisit the task of selecting responses to news and propose a novel convex optimization based approach to achieve better performance in both prediction accuracy and quality of summarization. We interpret responses in terms of two separate indicators, *informativeness* and *opinionatedness*, which are intuitive and important for human understanding. Furthermore, we consider message-level and text-level information simultaneously, and tackle the task of response selection from a data reconstruction perspective. Remarkably, the proposed framework is able to reduce the redundancy of the selected responses by evaluating the utility of a set of responses jointly. We also investigate different gradient-based optimization algorithms and analyze their convergence performance to solve the optimization problem efficiently. The experimental results demonstrate a significant improvement on real-world data over the state of the art. A valuable direction to pursue for further investigation is to facil-

[7] `http://www.bbc.com/news/science-environment-26462348`

itate personalized recommendation of news responses with the methodology proposed in this paper.

7 Acknowledgements

Research supported by the National Natural Science Foundation of China (No. 61375060), the Fundamental Research Funds for the Central Universities (WK0110000036), the National Science Foundation for Distinguished Young Scholars of China (No. 61325010), and the National High Technology Research and Development Program of China (No. 2014AA015203).

References

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 ACM International Conference on Web Search and Data Mining*, 2008.

[2] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] H. Becker, M. Naaman, and L. Gravano. Selecting quality twitter content for events. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 11, 2011.

[5] D. P. Bertsekas. Nonlinear programming. 1999.

[6] G. Beverungen and J. Kalita. Evaluating methods for summarizing twitter posts. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[7] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[9] 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 *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[10] M. A. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Select. Topics Signal Process*, 1(4):586–597, 2007.

[11] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. Document summarization based on data reconstruction. In *AAAI*, 2012.

[12] 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 International Conference on World Wide Web*, 2010.

[13] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.

[14] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[15] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010.

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

[17] S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive psychology*, 9(4):441–474, 1977.

[18] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[19] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Natural Language Processing and Text Mining*, pages 9–28. Springer, 2007.

[20] A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management*, 2010.

[21] Z. Ren, S. Liang, E. Meij, and M. de Rijke. Personalized time-aware tweets summarization. In *Proceedings of the 36th ACM SIGIR International Conference on Research and Development in Information Retrieval*, 2013.

[22] T. Stajner, B. Thomee, A. Popescu, and A. Jaimes. Automatic selection of social media responses to news. *In Proceedings of the 2013 ACM International Conference on Web Search and Data Mining*, 2013.

[23] E. Wachsmuth, M. Oram, and D. Perrett. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4(5):509–522, 1994.

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

[25] K. Yu, S. Zhu, W. Xu, and Y. Gong. Non-greedy active learning for text categorization using convex ansductive experimental design. In *Proceedings of the 31st ACM SIGIR International Conference on Research and Development in Information Retrieval*, 2008.

[26] F. M. Zanzotto, M. Pennacchiotti, and K. Tsioutsiouliklis. Linguistic redundancy in twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.