Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment

Yuheng Hu[†] **Fei Wang**[§] **Subbarao Kambhampati**[†]
[†] Department of Computer Science, Arizona State University, Tempe, AZ 85281
[§] IBM T. J. Watson Research Lab, Hawthorne, NY 10532
[†]{yuhenghu, rao}@asu.edu [§]fwang@us.ibm.com

Abstract

Individuals often express their opinions on social media platforms like Twitter and Facebook during public events such as the U.S. Presidential debate and the Oscar awards ceremony. Gleaning insights from these posts is of importance to analyzing the impact of the event. In this work, we consider the problem of identifying the segments and topics of an event that garnered praise or criticism, according to aggregated Twitter responses. We propose a flexible factorization framework, SOCSENT, to learn factors about segments, topics, and sentiments. To regulate the learning process, several constraints based on prior knowledge on sentiment lexicon, sentiment orientations (on a few tweets) as well as tweets alignments to the event are enforced. We implement our approach using simple update rules to get the optimal solution. We evaluate the proposed method both quantitatively and qualitatively on two large-scale tweet datasets associated with two events from different domains to show that it improves significantly over baseline models.

1 Introduction

Given the ubiquity and immediacy of social media, individuals often express their opinions on Twitter and Facebook, in particular during live or breaking public events such as the U.S. Presidential debate and Apple products press conference. While viewers can see opinions one by one when watching, the collection of these posts provides an opportunity to understand the overall sentiment of people during the event. Gleaning insights from those posts is of increasing importance to many businesses. Recent studies have revealed that a massive number of people, news media, companies and political campaigns turn to social media to collect views about products and political candidates during and after the event. This guides their choices, decision-making, voting, and even stock market investments [Bollen *et al.*, 2011].

In this work we are interested in analyzing public events by automatically characterizing segments and topics of that event in terms of the aggregate sentiments (positive [+] *v.s.* negative [-]) they elicited on Twitter (see Fig. 1). Classifying the sentiment behind textual content has received considerable attention during recent years. A standard approach would be to manually label comments (e.g., tweets) with their sentiment orientation and then apply off-the-shelf text classification techniques [Pang *et al.*, 2002]. However, such a solution is inapplicable to our problem due to three reasons. First, manually annotating the sentiment of a vast amount of tweets is time consuming and error-prone, presenting a bottleneck in learning high quality models. Besides, sentiment is always conveyed with highly domain-specific contextual cues, and the idiosyncratic expressions in tweets may rapidly evolve over time, especially when tweets are posted live in response to the event. It can cause models to potentially lose performance and become stale. Last and most importantly, this approach is unable to relate aggregated Twitter sentiment to segments and topics of the event. One may consider enforcing tweets' correlation with the segment and topics from the event that occur within fixed time-windows around the tweets' timestamps [Shamma *et al.*, 2009; O'Connor *et al.*, 2010] and classify the sentiment based on that. However, as pointed by our recent work [Hu *et al.*, 2012a], this assumption is often not valid: a segment of the event can actually be referred to by tweets at any time irrespective of whether the segment has already occurred or is occurring currently or will occur later on.

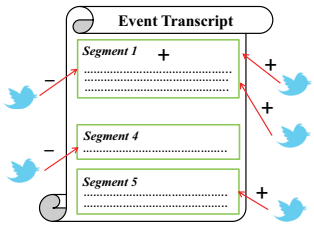


Figure 1: *Problem Setup*

The weaknesses discussed in the foregoing motivate the need for a fully automated framework to analyze events via aggregated twitter sentiment, with (1) little or no manual labeling of tweet sentiment, (2) ability to align tweets to the event, and (3) ability to handle the dynamics of tweets. While such a framework does not exist, the literature does provide partial solutions. For example, our recent work [Hu *et al.*, 2012b] provides an effective unsupervised framework called ET-LDA for aligning tweets and events by jointly modeling both the tweets and events in a latent topic space. Similarly, while manual annotation of all tweets is infeasible, it is often possible to get sentiment labeling for small sets of tweets. Finally, there also exist domain-independent sentiment lexicons such as MPQA corpus [Wilson *et al.*, 2009].

We propose a flexible framework, named SOCSENT, for event analytics via Twitter sentiment that leverages these partial solutions. Specifically, our framework seeks low-rank representations of the Twitter sentiment and its correlations

to the event by factorizing an input tweet-term matrix into four factors corresponding to tweets-segment, segment-topic, topic-sentiment and sentiment-words. The ET-LDA approach can be seen as providing the initial information ("prior knowledge") on the tweet-segment and segment-topic factors. Similarly, the availability of labeled tweets can be used to constrain the product of tweet-segment, segment-topic and topic-sentiment matrices. Finally, the sentiment lexicon is used to regulate the sentiment-words matrix. We pose this factorization as an optimization problem where, in addition to minimizing the reconstruction error, we also require that the factors respect the prior knowledge to the extent possible. We derive a set of multiplicative update rules that efficiently produce this factorization, and provide empirical comparisons with several competing methodologies on two real datasets, covering one recent U.S. presidential candidates debate in 2012 and one press conference. We examine the results both quantitatively and qualitatively to demonstrate that our method improves significantly over baseline approaches.

2 Related Work

Sentiment analysis has achieved great success in determining sentiment from underlying text corpora like newspaper articles [Pang *et al.*, 2002] and product reviews [Hu and Liu, 2004]. Various approaches, mostly learning-based, have been proposed, which include classification using sentiment lexicons [Wilson *et al.*, 2009], topic sentiment mixture model [Mei *et al.*, 2007], and nonnegative matrix factorization [Li *et al.*, 2009]. Recently, there has been increasing interest in applying sentiment analysis to social media data like tweets such as [Bollen *et al.*, 2011; O'Connor *et al.*, 2010]. Some works also consider incorporating external social network information to improve the classification performance ([Tan *et al.*, 2011; Hu *et al.*, 2013]).

Our work is also inspired by the research in characterizing events by the tweets around them. These works include inferring structures of events using Twitter usage patterns [Shamma *et al.*, 2009], exploring events by the classification of audience types on Twitter [Vieweg *et al.*, 2010], sentiment analysis of tweets to understand the events [Diakopoulos and Shamma, 2010] and modeling the behavioral patterns between events and tweets [Hu *et al.*, 2012a].

The focus of the above work is mostly classifying sentiments of document sources or processing the tweets around the event. However, they do not provide insights into how to characterize the event's segments and topics through the aggregated Twitter sentiment, which is the main contribution of this work. Perhaps the closest work to us is [Diakopoulos and Shamma, 2010]. However, it depends on completely manual coding (via Amazon Mechanical Turk) to determine the sentiment. In contrast, we provide a fully automated and principled solution which can be used to handle the vast amount of tweets posted around an event.

3 SocSent Framework

In this section, we first present the basics of our proposed framework SOCSENT. We then describe how to obtain and leverage prior knowledge. Table 1 lists the notation used in this paper. Note that although the primary sentiment we focus on is binary: positive or negative, our model can be easily extended to handle multiple types of sentiment.

Table 1: Notation

Notation	Size	Description
$\mathbf{X}$	$n_t \times N$	Tweet-Term matrix
$\mathbf{G}$	$n_t \times n_s$	Tweet-Segment matrix
$\mathbf{T}$	$n_s \times K$	Segment-Topic matrix
$\mathbf{S}$	$K \times 2$	Topic-Sentiment matrix
$\mathbf{F}$	$N \times 2$	Term-Sentiment matrix
$\mathbf{G}_0$	$n_t \times n_s$	Prior knowledge on Tweet-Segment
$\mathbf{F}_0$	$N \times 2$	Prior knowledge on Term-Sentiment
$\mathbf{R}_0$	$n_t \times 2$	Prior knowledge on Tweet-Sentiment

3.1 Basic Framework

Let a public event be partitioned into n_s sequentially ordered segments, each of which discusses a particular set of topics. A segment consists of one or more coherent paragraphs available from the transcript of the event (we will discuss the segmentation in Section 3.2). There are also n_t tweets posted by the audience in response to the event, contributing to a vocabulary of N terms. As mentioned earlier, our goal is to identify segment and topics of the event that gained praise or criticism, according to how people reacted and appreciated them on Twitter. Accordingly, our basic framework takes those n_t tweets in terms of tweet-vocabulary matrix $\mathbf{X}$ as input and decomposes into four factors that specify soft membership of tweets and terms in three latent dimensions: segment, topic, and sentiment. In other words, our basic model tries to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{G},\mathbf{S},\mathbf{F}} \quad & \left\|\mathbf{X} - \mathbf{GTSF}^\top\right\|_F^2 \\ s.t. \quad & \mathbf{G} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{F} \geqslant 0 \end{aligned} \tag{1}$$

where $\mathbf{G} \in \mathbb{R}^{n_t \times n_s}$ indicates the assignment of each tweet to the event segments based on the strength of their topic associations. That is, the i-th row of $\mathbf{G}$ corresponds to the posterior probability of tweet i referring to each of the n_s segments of the event. Similarly, $\mathbf{T} \in \mathbb{R}^{n_s \times K}$ indicates the posterior probability of a segment s belonging to the K topic clusters. Also, $\mathbf{S} \in \mathbb{R}^{K \times 2}$ encodes the sentiment distribution of each topic k. Finally, $\mathbf{F} \in \mathbb{R}^{N \times 2}$ represents the binary sentiment for each term in the vocabulary of tweets. Note that the non-negativity makes the factorized factors easy to interpret.

As a result of this factorization, we can readily determine whether people appreciate the segments or topics of the event or dislike them. For example, from topic-sentiment matrix $\mathbf{S}$ we can directly obtain the crowd's opinion on each topic covered in the event. In addition, from segment-sentiment matrix $\mathbf{Q}$ (where $\mathbf{Q} = \mathbf{T} \times \mathbf{S}$), we can distill sentiment regarding each segment of the event. Finally, it is also feasible to characterize the sentiment for each tweet, through the new tweet-sentiment matrix $\mathbf{R}$ where $\mathbf{R} = \mathbf{G} \times \mathbf{T} \times \mathbf{S}$.

Conceptually, our basic matrix factorization framework is similar to the probabilistic latent semantic indexing (PLSI) model [Hofmann, 1999] and the non-negative matrix Trifactorization model (NMTF) [Ding *et al.*, 2006]. In PLSI and NMTF, $\mathbf{X}$ is viewed as the joint distribution between words and documents, which is factorized into three components: $\mathbf{W}$ is the word class-conditional probability, $\mathbf{D}$ is the document class-conditional probability and $\mathbf{S}$ is the class probability distribution. These methods provide a simultaneous solution for the word and document class conditional distribution. Our model goes beyond that by providing simultaneous solutions for projecting the rows and the columns of $\mathbf{X}$ onto three latent dimensions.

3.2 Constructing Prior Knowledge

So far, our basic matrix factorization framework provides potential solutions to infer the aggregated Twitter sentiment regarding the segment and topics of the event. However, it largely ignores a lot of prior knowledge on the learned factors. Previous literature (see [Pang *et al.*, 2002]) shows that leveraging such knowledge can help regulate the learning process and enhance the framework's performance (which is empirically verified in Section 4). Accordingly, we first show how to construct three types of prior knowledge: (a) sentiment lexicons of terms, (b) sentiment labels of tweets, and (c) alignment of tweets to the segment of the event. We then incorporate them into our framework in Section 3.3.

Sentiment Lexicon

Our first prior knowledge is from a sentiment lexicon, which is publicly available as a part of the MPQA corpus[1]. It contains 7,504 representative words that have been human-labeled as expressing positive or negative sentiment. In total, there are 2,721 positive (e.g., "*awesome*") and 4,783 negative (e.g., "*sad*") unique terms. It should be noted, that this list was constructed without any specific domain in mind; this is further motivation for using training examples and unlabeled data to learn domain specific connotations. To overcome the irregular English usage and out-of-vocabulary words in Twitter, we apply a lexicon normalization technique [Han and Baldwin, 2011] for the terms in our sentiment lexicon. This involves detecting ill-formed words and generates correction candidates based on morphophonemic similarity. As a result, "*happppppppppy*" is seen as a correct variant of "*happy*" thus sharing the same sentiment. We use those candidates to expand the original lexicon, making it adaptive to Twitter-related linguistic styles. Besides, we also add popular abbreviations and acronyms on Twitter such as "*smh*" (shake my head, negative) and "*lol*" (positive) to the lexicon. Eventually, we have 5,267 positive and 8,701 negative unique terms in the lexicon. We encode it in a term-sentiment matrix $\mathbf{F}_0$, where $\mathbf{F}_0(i,1) = 1$ if the a word i has positive sentiment, and $\mathbf{F}_0(i,2) = 1$ for negative sentiment.

Sentiment Label of Tweets

In addition to the lexicon, our second prior knowledge comes from human effort. We ask people to label the sentiment for a few tweets (e.g., less than 1000) for the purposes of capturing some domain-specific connotations, which later leads to a more domain-adapted model. The partial labels on documents can be described using a tweet-sentiment matrix $\mathbf{G}_0$ where $\mathbf{G}_0(i,1) = 1$ if the tweet expresses positive sentiment, and $\mathbf{G}_0(i,2) = 1$ for negative sentiment. One can use soft sentiment labeling for tweets, though our experiments are conducted with hard assignments.

Alignment of Tweets to the Event Segments

Our last prior knowledge focuses on the alignment between the event and the tweets which were posted in response to it. Like the sentiment label of tweets, this prior also tries making the model more domain-specific. To overcome the inherent drawbacks of the fixed time-window approach, we apply the ET-LDA model from our previous work [Hu *et al.*, 2012b]. ET-LDA is a hierarchical Bayesian model based on Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003]. It aims to model: (1) the event's topics and their evolution (event segmentation), as well as (2) the associated tweets' topics and the crowd's tweeting behaviors. The model has two major components with each capturing one perspective of the goals. Both parts have the LDA-like model, and are connected by the link which captures the topical influences from the event on its Twitter feeds. In practice, ET-LDA takes an event's transcript and all the event-related tweets and then concurrently partitions the speech into a number of homogeneous segments and aligns each tweet to event segments based on the strength of their topical associations. We encode the alignment results in a tweet-segment matrix $\mathbf{R}_0$ where its rows represent n_t tweets and its columns represent n_s segments of the event. As the content of $\mathbf{R}_0$ is the posterior probability of a tweet referring to the segments, we have $\sum_{1 \le j \le n_s} \mathbf{R}_0(i,j) = 1$ for each tweet i.

[1] http://mpqa.cs.pitt.edu/

3.3 Incorporating Prior Knowledge into Our Framework

After defining and constructing the three types of prior knowledge, we can incorporate them into our basic factorization framework as supervision (see Eq. 2). We later demonstrate in Section 4 that such supervision provides better regularization to the learned factors and significantly enhances the model's performance.

$$\begin{aligned} \min_{\mathbf{F},\mathbf{T},\mathbf{G}} \quad & \mathcal{J} = \left\| \mathbf{X} - \mathbf{GTSF}^\top \right\|_F^2 \\ & + \alpha Tr\left((\mathbf{F} - \mathbf{F}_0)^\top \mathbf{\Lambda} (\mathbf{F} - \mathbf{F}_0) \right) \\ & + \beta Tr\left((\mathbf{GTS} - \mathbf{R}_0)^\top \mathbf{\Theta} (\mathbf{GTS} - \mathbf{R}_0) \right) \\ & + \gamma Tr\left((\mathbf{G} - \mathbf{G}_0)^\top \mathbf{\Gamma} (\mathbf{G} - \mathbf{G}_0) \right) \\ s.t. \quad & \mathbf{F} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{G} \geqslant 0, \mathbf{S} \geqslant 0 \end{aligned} \tag{2}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters which determine the extent to which we enforce $\mathbf{F} \approx \mathbf{F}_0$, $\mathbf{G} \approx \mathbf{G}_0$ and the multiplication $\mathbf{G} \times \mathbf{T} \times \mathbf{S} \approx \mathbf{R}_0$, respectively. $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$, $\mathbf{\Theta} \in \mathbb{R}^{n_t \times n_t}$ and $\mathbf{\Gamma} \in \mathbb{R}^{n_t \times n_t}$ are diagonal matrices, indicating the entries of $\mathbf{F}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$ that correspond to labeled entities. The squared loss terms ensure that the solution for $\mathbf{F}$, $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$, in the otherwise unsupervised learning problem, be close to the prior knowledge $\mathbf{F}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$.

It is worth noting the benefit of coupling $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$ in Eq. 2. One may consider applying regularization to each of them. However, this will add additional computational cost during the model inference since the model gets more complex. In contrast, the supervision from $\mathbf{R_0}$ on the joint of $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$ can achieve the equivalent enforcement (while $\mathbf{G}$ is individually constrained).

The above model is generic and it allows flexibility. For example, in some cases, our prior knowledge on $\mathbf{F}_0$ is not very accurate and we use smaller α so that the final results are not dependent on $\mathbf{F}$ very much. In addition, the introduction of $\mathbf{G}_0$ and $\mathbf{R}_0$ allows us to incorporate partial knowledge on tweet polarity and assignment information.

3.4 Model Inference

To infer the solutions for factors $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$ in the framework, we first rewrite Eq. 2 as:

$$\begin{aligned}\mathcal{J} =& Tr\left(\mathbf{X}^\top\mathbf{X} - 2\mathbf{X}^\top\mathbf{GTSF}^\top\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTSF}^\top\right)\\ &+ \alpha Tr\left(\mathbf{F}^\top\mathbf{\Lambda F} - 2\mathbf{F}^\top\mathbf{\Lambda F}_0 + \mathbf{F}_0^\top\mathbf{\Lambda F}_0\right)\\ &+ \beta Tr\left(\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{\Theta GTS} - 2\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{\Theta R}_0 + \mathbf{R}_0^\top\mathbf{\Theta R}_0\right)\\ &+ \gamma Tr\left(\mathbf{G}^\top\mathbf{\Gamma G} - 2\mathbf{G}^\top\mathbf{\Gamma G}_0 + \mathbf{G}_0^\top\mathbf{\Gamma G}_0\right) \quad (3)\end{aligned}$$

The coupling between $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$ makes it difficult to find optimal solutions for all factors simultaneously. In this work, we adopt an alternative optimization scheme [Ding *et al.*, 2006] for Eq. 3, under which we update $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$ alternatingly with the following multiplicative update rules.

First, for the tweets-segment matrix $\mathbf{G}$, we have:

$$G_{ij} \leftarrow G_{ij}\sqrt{\frac{\left[\mathbf{XFS}^\top\mathbf{T}^\top + \beta\mathbf{\Theta R}_0\mathbf{S}^\top\mathbf{T}^\top + \gamma\mathbf{\Gamma G}_0\right]_{ij}}{\left[\mathbf{GTSF}^\top\mathbf{FS}^\top\mathbf{T}^\top + \beta\mathbf{\Theta GTSS}^\top\mathbf{T}^\top + \gamma\mathbf{\Gamma G}\right]_{ij}}} \quad (4)$$

Next, for the tweets-segment matrix $\mathbf{T}$, we have:

$$T_{ij} \leftarrow T_{ij}\sqrt{\frac{\left[\beta\mathbf{G}^\top\mathbf{\Theta R}_0\mathbf{S}^\top + \mathbf{G}^\top\mathbf{XFS}^\top\right]_{ij}}{\left[\mathbf{G}^\top\mathbf{GTSF}^\top\mathbf{FS}^\top + \beta\mathbf{G}^\top\mathbf{\Theta GTSS}^\top\right]_{ij}}} \quad (5)$$

In addition, for the tweets-segment matrix $\mathbf{S}$, we have:

$$S_{ij} \leftarrow S_{t_{ij}}\sqrt{\frac{\left[\mathbf{T}^\top\mathbf{G}^\top\mathbf{XF}\right]_{ij}}{\left[\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTSF}^\top\mathbf{F} + \mathbf{T}^\top\mathbf{G}^\top\mathbf{\Theta GTS}\right]_{ij}}} \quad (6)$$

Last, for the tweets-segment matrix $\mathbf{F}$, we have:

$$F_{ij} \leftarrow F_{ij}\sqrt{\frac{\left[\mathbf{X}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}_0\right]_{ij}}{\left[\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}\right]_{ij}}} \quad (7)$$

Our learning algorithm consists of an iterative procedure using the above four rules until convergence. The outline of the specific steps is shown below.

Algorithm 1: Factorization with Prior Knowledge

input : α, β, γ
output: $\mathbf{G}, \mathbf{T}, \mathbf{S}, \mathbf{F}$

Initialize $\mathbf{G} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{S} \geqslant 0, \mathbf{F} \geqslant 0$
while *Algorithm Not Converges* **do**
  Update $\mathbf{G}$ with Eq.(4) while fixing $\mathbf{T}$,$\mathbf{S}$,$\mathbf{F}$
  Update $\mathbf{T}$ with Eq.(5) while fixing $\mathbf{G}$,$\mathbf{S}$,$\mathbf{F}$
  Update $\mathbf{S}$ with Eq.(7) while fixing $\mathbf{G}$,$\mathbf{T}$,$\mathbf{F}$
  Update $\mathbf{F}$ with Eq.(6) while fixing $\mathbf{G}$,$\mathbf{T}$,$\mathbf{S}$
end

Computational complexity The tweet-term matrix $\mathbf{X}$ is typically very sparse with $z \ll n_t \times N$ non-zero entries. Also, K and n_s are typically also much smaller than n_t and N. By using sparse matrix multiplications and avoiding dense intermediate matrices, the updates can be very efficiently and easily implemented. In particular, updating $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{F}$ each takes $O(C^2(n_t+n_s+N)+Cz)$ time per iteration which scales linearly with the dimensions and density of the data matrix. C is a constant. Empirically, the number of iterations before practical convergence is usually very small (less than 350). Thus, our approach can scale to large datasets.

3.5 Algorithm Correctness and Convergence

The correctness and convergence of Algorithm 1 can be guaranteed by the following two theorems.

Theorem 1. *The limiting solutions of the updating rules Eq.(4), Eq.(5), Eq.(7), Eq.(6) satisfy the Karush–Kuhn–Tucker (KKT)[Nocedal and Wright, 2000] conditions for minimizing $\mathcal{J}$ in Eq.(3) under the nonnegativity constraints.*

Proof. We prove the theorem for updating $\mathbf{F}$ here, all others can be proved in the same way. The Lagrangian for $\mathbf{F}$ is

$$\begin{aligned}\mathcal{L} =& \mathcal{J} - \mathbf{\Gamma S}^\top\\ =& \left\|\mathbf{X} - \mathbf{GTSF}^\top\right\|_F^2 + \alpha Tr\left((\mathbf{F}-\mathbf{F}_0)^\top\mathbf{\Lambda}(\mathbf{F}-\mathbf{F}_0)\right)\\ &- \mathbf{\Psi F}^\top + C \quad (8)\end{aligned}$$

where the Lagrangian multipliers $\mathbf{\Psi}_{ij}$ enforce the nonnegativity constraint on $\mathbf{F}_{ij}$, and we use C to represent the terms irrelevant to $\mathbf{F}$. The gradient of $\mathcal{L}$ with respect to $\mathbf{F}$ is

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial\mathbf{F}} =& 2\left(\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}\right)\\ &-2\left(\mathbf{X}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}_0\right) - \mathbf{\Psi}\end{aligned}$$

From the complementary slackness condition, we can obtain

$$\begin{aligned}&\left(2\left(\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}\right)\right.\\ &\left.-2\left(\mathbf{X}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}_0\right)\right)_{ij}\mathbf{F}_{ij} = \mathbf{\Lambda}_{ij}\mathbf{F}_{ij} = 0 \quad (9)\end{aligned}$$

This is the fixed point equation that the solution of $\mathbf{G}_s$ must satisfy at convergence. Actually, for Eq.(7), we have the following condition at convergence

$$F_{ij} = F_{ij}\sqrt{\frac{\left[\mathbf{X}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}_0\right]_{ij}}{\left[\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}\right]_{ij}}} \quad (10)$$

which is equivalent to

$$\begin{aligned}&\left(2\left(\mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}\right)\right.\\ &\left.-2\left(\mathbf{X}^\top\mathbf{GTS} + \alpha\mathbf{\Lambda F}_0\right)\right)_{ij}\mathbf{F}_{ij}^2 = 0 \quad (11)\end{aligned}$$

Eq. (11) is identical to Eq. (9). Both equations require that at least one of the two factors is equal to zero. The first factor in both equations are identical. For the second factor $\mathbf{F}_{ij}$ or $\mathbf{F}_{ij}^2$ if $\mathbf{F}_{ij} = 0$ then $\mathbf{F}_{ij}^2 = 0$, and vice versa. Thus if Eq. (9) holds, Eq. (11) also holds and vice versa. □

Theorem 2.. *The updating rules Eq.(4), Eq.(5), Eq.(6), Eq.(7) will finally converge to a stationary point*.

This theorem can be proved with the auxiliary function method as in [Li *et al.*, 2009]. Due to the space limit, we omit the proof details here.

4 Experiments

In this section, we examine the effectiveness of our proposed framework SOCSENT against other baselines. Three sentiment classification tasks are undertaken on: 1) event segments, 2) event topics, and 3) tweets sentiment. We also evaluate the robustness of our framework with respect to various sizes of training data and different combinations of the prior knowledge.

Datasets and Experimental Setup

We use two large scale tweet datasets associated with two events from different domains: (1) the first U.S. Presidential debate on Oct 3, 2012 and (2) President Obama's Middle East speech on May 19, 2011. The first tweet dataset consists of 181,568 tweets tagged with "#DenverDebate" and the second dataset consists of 25,921 tweets tagged with "#MEspeech". Both datasets were crawled via the Twitter API using these two hashtags. In the rest of this paper, we use the hashtags to refer to these events. We obtained the transcripts of both events from the New York Times, where DenverDebate has 258 paragraphs and MESpeech has 73 paragraphs. Preprocessing operations, such as stemming and stopwords elimination, are applied to both tweets and transcripts. Furthermore, we split both tweet datasets into a 80-20 training and test sets.

For ET-LDA, we use the implementation from [Hu *et al.*, 2012b]. Its parameters are set using the same procedure described in [Hu *et al.*, 2012b]. Coarse parameter tuning for our framework SOCSENT was also performed. We varied α, β and γ and chose the combination which minimizes the reconstruction error in our training set. As a result, we set $\alpha = 2.8$, $\beta = 1.5$, $\gamma = 1.15$. All experimental results in this section are averaged over 20 independent runs.

Establishing Ground Truth: To quantitatively evaluate the performance of our framework, we need the ground truth of the sentiment for event segments, event topics and tweets. At first, we asked 14 graduate students in our school (but not affiliated with our project or group) to manually label the sentiment (i.e., positive or negative) of 1,500 randomly sampled tweets for each dataset. We then applied ET-LDA model to segment two events and establish the alignment between the labeled tweets and the event segments. So for each segment, we label its sentiment according to the majority aggregated Twitter sentiment that correlated to it. For example, if 60 out of 100 tweets that refer to segment S are positive, then S is considered to have received positive sentiment since people showed their appreciation for it. In addition, the top-5 topics of S (these top topics were also learned by ET-LDA) are also labeled as having positive sentiment. We aggregate this sentiment across all the event segments and assign the majority sentiment to each topic of the event. Finally, we obtained 35 segments of DenverDebate, where 20 segments were labeled as negative. Also, 62% labeled tweets and 12 out of 20 topics were negative. For MEspeech, we have 6 of 9 segments, 13 out of 20 topics, and 72% tweets marked as negative. Such negativity on Twitter is not surprising, it actually conforms to the findings in [Diakopoulos and Shamma, 2010].

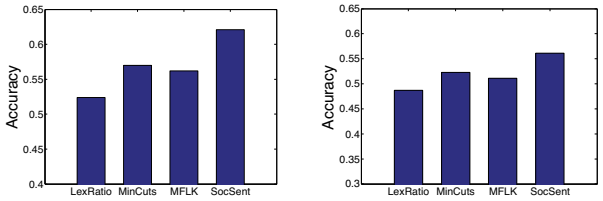
Baselines

To better understand the performance of SOCSENT, we implemented some competitive baseline approaches:

- *LexRatio*: This method [Wilson *et al.*, 2009] counts the ratio of sentiment words from OpinionFinder subjectivity lexicon[2] in a tweet to determine its sentiment orientation. Due to its unsupervised setting, we ignore the tweets which do not contain any sentiment words.
- *MinCuts*: This method [Pang and Lee, 2004] utilizes contextual information via the minimum-cut framework to improve polarity-classification accuracy. We used MinCuts package in LingPipe[3].
- *MFLK*: This is a supervised matrix factorization method which decomposes an input term-document matrix into document-sentiment and sentiment-terms matrices. Supervision from a sentiment lexicon is enforced [Li *et al.*, 2009]. We implemented it with our Twitter lexicon.

Classification of Sentiment of the Event Segment

We first study the performance of SOCSENT on classifying the segments' sentiment for the two events via aggregated Twitter responses against the baseline methods. Note these baselines are inherently unable to relate their Twitter sentiment classification results to the event segments. To remedy this, we take a two-step approach. First, we split the whole event into several time windows (10-min. in our experiment). Then, we enforce the segments' sentiment to be correlated with the inferred tweet sentiment that occur within the time-windows around the tweets's timestamps. Figure 2 presents the classification results where accuracy is measured based on the manually labeled ground truth. It is clear that SOCSENT can effectively utilize the partially available knowledge on tweet/event alignment from ET-LDA to improve the quality of sentiment classification in both events. In particular, it improves other approaches in the range of 7.3% to 18.8%.

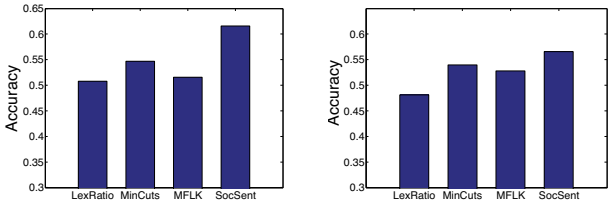


(a) DenverDebate (b) MEspeech

Figure 2: *Classification of Sentiment of the Event Segment.*

Classification of Sentiment of the Event Topics

Next, we study sentiment classification of the topics covered in the event. The results are shown in Figure 3. Similar to the last task, we again use time window approach to correlate the event topics with the sentiment of tweets. Not surprisingly, SOCSENT improves the three baselines with a range of 6.5% to 17.3% for both datasets.



(a) DenverDebate (b) MEspeech

Figure 3: *Classification of Sentiment of the Event Topics.*

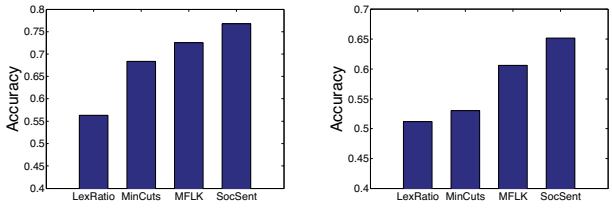
Classifying the Sentiment of the Tweets

In the third experiment, we evaluate the prediction accuracy of Twitter sentiment. Figure 4 illustrates the results for tweets posted in response to DenverDebate and MEspeech. As

[2]http://mpqa.cs.pitt.edu/opinionfinder/

[3]http://alias-i.com/lingpipe/

we can see, SOCSENT greatly outperforms other baselines on both datasets. In fact, it achieves the largest performance improvement margin (compared to results in Figure 2 and 3). We believe this is because SOCSENT adopts the direct supervision from the pre-labeled tweet sentiment. We also observe that all the methods have better performance on DenverDebate than on MEspeech (see Figure 2, 3 and 4). This is mainly because DenverDebate attracted a significant larger number of tweets than MEspeech. Therefore, the DenverDebate dataset is likely to be less sparse in the sense that more words from the sentiment lexicon can also be found in the training set. As a result, the effect of sentiment lexicon is fully utilized thus producing better results than on MEspeech.



(a) DenverDebate (b) MEspeech

Figure 4: *Classification of Sentiment of the Tweets.*

Varying Training Data Size

In Table 2, we show the performance of classifying segments' sentiment using various methods with respect to different size of training data. Note that LexRatio is an unsupervised approach so its performance is unchanged in this experiment. It is clear that the other three methods achieve better performance when more training data is supplied. Besides, on both DenverDebate and MEspeech datasets, we find that SOCSENT is more stable over other methods with various sizes of training data from 10% to 100%. In other words, SOCSENT does not show dramatic changes when the size of the training data changes. This demonstrates that our proposed method is robust to training data sizes.

Table 2: *Classification accuracy on Segment's sentiment vs. Training data sizes. Notations:* ***LR*** *is for LexRatio,* ***MC*** *is for MinCuts,* ***MF*** *for MFLK, and* ***SS****, for our method* SOCSENT .

DenverDebate

	$T_{10\%}$(gain)	$T_{25\%}$(gain)	$T_{50\%}$(gain)	$T_{100\%}$(gain)
LR	0.524	0.524	0.524	0.524
MC	0.538 (+2.7%)	0.563 (+7.4%)	0.568 (+8.4%)	0.574 (+9.5%)
MF	0.532 (+1.5%)	0.536 (+2.3%)	0.558 (+6.5%)	0.562 (+7.3%)
SS	0.588 (+12.2%)	0.595 (+13.5%)	0.613 (+17.0%)	0.621 (+18.5%)

MESpeech

	$T_{10\%}$(gain)	$T_{25\%}$(gain)	$T_{50\%}$(gain)	$T_{100\%}$(gain)
LR	0.487	0.487	0.487	0.487
MC	0.502 (+3.1%)	0.520 (+6.8%)	0.521 (+6.9%)	0.523 (+7.4%)
MF	0.488 (0.2%)	0.504 (+3.5%)	0.509 (+4.5%)	0.511 (+4.9%)
SS	0.541 (+11.1%)	0.549 (+12.7%)	0.558 (+14.6%)	0.561 (+15.2%)

Effectiveness of Prior Knowledge

Finally, given the available three types of prior knowledge – sentiment lexicon, tweet labels and tweet/event alignment by ET-LDA, it is interesting to explore their impact on the performance of SOCSENT. Table 3 presents the evaluation results on two datasets, where we judge SOCSENT on three aforementioned classification tasks with respect to different combinations of its prior knowledge. For each combination, we come up with separate update rules which have the similar form as Eq. 4-Eq. 6. Besides, we find optimal parameters using the same procedure described above in the setup of experiment. Several insights are gained here: First, using single type of prior knowledge is less effective than combining them. Especially, combining all three types of prior knowledge leads to the most significant improvement (an average of 29.8% gain over the baseline *N.A* on two datasets). Second, domain-specific knowledge (tweet labels, event/tweet alignment) is more effective than domain-independent knowledge (sentiment lexicon) in all three prediction tasks. Last, domain-specific knowledge is particulary helpful in its corresponding task. For example, having tweet/event alignment (denoted as $\mathbf{G}_0$ in Table 3) achieves more accurate results in classifying the sentiment of the event segments than without having it. For example, combinations with this prior knowledge such as $\mathbf{F}_0 + \mathbf{G}_0$ or $\mathbf{R}_0 + \mathbf{G}_0$ have better performance than $\mathbf{F}_0 + \mathbf{R}_0$ with 6.5% and 8.3% improvement, respectively. These insights demonstrate the advantage of SOCSENT's ability to seamlessly incorporate prior knowledge.

Table 3: *Combinations of prior knowledge vs. Accuracy. Notations:* $\mathbf{F}_0$ *for sentiment Lexicon,* $\mathbf{R}_0$ *for tweets labels,* $\mathbf{G}_0$ *for prior tweet/event alignment knowledge from ET-LDA.* ***N.A*** *refers to the basic framework without any constraints.*

DenverDebate

	Segment (gain)	Topics (gain)	Tweets (gain)
N.A	0.486	0.502	0.498
$\mathbf{F}_0$	0.523 (+7.5%)	0.542 (+7.9%)	0.545 (+9.4%)
$\mathbf{R}_0$	0.532 (+9.5%)	0.548 (+9.2%)	0.578(+16.1 %)
$\mathbf{G}_0$	0.484 (-0.01%)	0.504 (+0.02%)	0.491 (-1.4 %)
$\mathbf{F}_0$+$\mathbf{R}_0$	0.572 (+17.7%)	0.564 (+12.4%)	0.735 (+47.8 %)
$\mathbf{F}_0$+$\mathbf{G}_0$	0.604 (+23.6%)	0.605 (+20.5%)	0.68(+36.5%)
$\mathbf{R}_0$+$\mathbf{G}_0$	0.612 (+25.7%)	0.612 (+21.9%)	0.687(+37.9%)
$\mathbf{F}_0$+$\mathbf{R}_0$+$\mathbf{G}_0$	0.618 (+27.2%)	0.628 (+25.1%)	0.768 (+54.2 %)

MEspeech

	Segment (gain)	Topics (gain)	Tweets (gain)
N.A	0.472	0.498	0.512
$\mathbf{F}_0$	0.493 (+4.4%)	0.503 (+1.1%)	0.557 (+8.7%)
$\mathbf{R}_0$	0.502 (+6.3%)	0.512 (+2.8%)	0.566 (+10.5%)
$\mathbf{G}_0$	0.467 (-1.1%)	0.494 (-0.8%)	0.515 (+0.5%)
$\mathbf{F}_0$+$\mathbf{R}_0$	0.542 (+14.8%)	0.552 (+10.2%)	0.606 (+18.3%)
$\mathbf{F}_0$+$\mathbf{G}_0$	0.568 (+20.3%)	0.578 (+16.0%)	0.632 (+23.4%)
$\mathbf{R}_0$+$\mathbf{G}_0$	0.578 (+22.4%)	0.588 (+18.1%)	0.642 (+25.3%)
$\mathbf{F}_0$+$\mathbf{R}_0$+$\mathbf{G}_0$	0.588 (+24.5%)	0.598 (+20.1%)	0.652 (+27.3%)

5 Conclusion

In this paper, we have described a flexible factorization framework, SOCSENT that characterizes the segment and topics of an event via aggregated Twitter sentiment. Our model leverages three types of prior knowledge: sentiment lexicon, manually labeled tweets and tweet/event alignment from ET-LDA, to regulate the learning process. We evaluated our framework quantitatively and qualitatively through various tasks. Based on the experimental results, our model shows significant improvements over the baseline methods. We believe that our work presents the first step towards understanding complex interactions between events and social media feedback and reveals a perspective that is useful for the extraction of a variety of further dimensions such as polarity and influence prediction.

Acknowledgements
This research is supported in part by ONR grants N000140910032 and N00014-13-1-0176, NSF grant IIS201330813 and a Google Research Award. The authors would like to thank the students from the IR course at Arizona State University for providing the ground truth labels for this study.

References

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bollen *et al.*, 2011] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[Diakopoulos and Shamma, 2010] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the 28th international conference on Human factors in computing systems CHI 10*, page 1195, 2010.

[Ding *et al.*, 2006] Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135, 2006.

[Han and Baldwin, 2011] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 368–378, 2011.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[Hu and Liu, 2004] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[Hu *et al.*, 2012a] Y. Hu, A. John, D.D. Seligmann, and F. Wang. What were the tweets about? topical associations between public events and twitter feeds. *Proceedings from ICWSM, Dublin, Ireland*, 2012.

[Hu *et al.*, 2012b] Y. Hu, A. John, F. Wang, and S. Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[Hu *et al.*, 2013] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. *Proceedings of WSDM*, 2013.

[Li *et al.*, 2009] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 244–252. Association for Computational Linguistics, 2009.

[Mei *et al.*, 2007] Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.

[Nocedal and Wright, 2000] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2000.

[O'Connor *et al.*, 2010] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[Pang and Lee, 2004] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[Pang *et al.*, 2002] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[Shamma *et al.*, 2009] D.A. Shamma, L. Kennedy, and E.F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10. ACM, 2009.

[Tan *et al.*, 2011] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. *arXiv preprint arXiv:1109.6018*, 2011.

[Vieweg *et al.*, 2010] S. Vieweg, A.L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088. ACM, 2010.

[Wilson *et al.*, 2009] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.