Sentiment Expression Conditioned by Affective Transitions and Social Forces

Moritz Sudhof* Andrés Goméz Emilsson† Andrew L. Maas* Christopher Potts‡
*Computer Science †Psychology ‡Linguistics
Stanford University
{sudhof, nc07agom, amaas, cgpotts}@stanford.edu

ABSTRACT

Human emotional states are not independent but rather proceed along systematic paths governed by both internal, cognitive factors and external, social ones. For example, anxiety often transitions to disappointment, which is likely to sink to depression before rising to happiness and relaxation, and these states are conditioned by the states of others in our communities. Modeling these complex dependencies can yield insights into human emotion and support more powerful sentiment technologies.

We develop a theory of conditional dependencies between emotional states in which emotions are characterized not only by valence (polarity) and arousal (intensity) but also by the role they play in state transitions and social relationships. We implement this theory using conditional random fields (CRFs) that synthesize textual information with information about previous emotional states and the emotional states of others. To assess the power of affective transitions, we evaluate our model in a collection of 'mood' updates from the Experience Project. To assess the power of social factors, we use a corpus of product reviews from a website in which the community dynamics encourage reviewers to be influenced by each other. In both settings, our models yield improvements of statistical and practical significance over ones that classify each text independently of its emotional or social context.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Discourse; Text analysis

Keywords

multidimensional sentiment analysis; sentiment as social; sentiment transitions

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623687.

1. INTRODUCTION

Human emotional states are not independent, nor are they experienced in isolation. Rather, they proceed along systematic paths and cohere into psychologically important groups. Anxiety often transitions to disappointment, which is likely to sink to depression, often a long-lasting state. Happiness amplifies to elation or settles down to contentment. States like anxiety, optimism, and worry are inherently transitional; one's optimism might be vindicated, leading to positive states, or crushed, leading to disappointment. All these states are also conditioned by the states of others in our communities; complex social forces can push our emotions closer to those around us or pull them farther away.

Our goal is to show that modeling dependencies between emotional states can yield insights into human emotion and support more powerful sentiment technologies. In our first set of experiments (Section 3), we concentrate on the affective transitions that individuals experience. Using a collection of users' 'mood' updates from the Experience Project,[1] we develop a simple probability model for capturing the conditional dependencies between emotional states. The ideas are founded in dimensional models of emotion [10, 28, 29, 34] based in valence (polarity) and arousal (intensity), but we emphasize the role that emotions play in transitions between states. We implement this theory using conditional random fields (CRFs; [13, 32]) that synthesize textual information with information about temporally organized sequences of emotions, and we show that this delivers improvements of statistical and practical significance over models that classify each text independently of its emotional or social context. The improvements are largest for transitional emotional states, which are often particularly important for industry applications, since moments of emotional inflection often coincide with changing opinions or attitudes.

We then extend our model to social influences (Section 4). Using a corpus of product reviews from a website with rich community interactions [8, 18, 19], we present evidence that the current reviewer for a product is influenced by the sequence of reviews already posted about that product. Here, the sequences come, not from a specific individual, but rather from group-level actions and reactions. Once again, CRFs incorporating this sequence information out-perform simple classifiers by a large margin. This is an important counterpart to our moods study, not only because of its social dimension, but also because it highlights the value of transition information for common polarity-based sentiment tasks.

[1] `http://www.experienceproject.com`

2. RELATED WORK

Our goal is to understand and make use of the dependencies between emotional states, which derive from both personal and social factors. Our work thus finds precedent in diverse areas of natural language processing, social network analysis, and cognitive psychology, including models of social influence and multidimensional sentiment expression. This section provides a high-level summary of these connections with the existing literature.

Our view of sentiment is a *dimensional* one [10, 28, 29, 34], in which emotions are characterized by a small number of more basic dimensions, usually valence and arousal but sometimes much more abstract ones (e.g., [15]). Valence roughly corresponds to polarity (like/dislike, pleasure/pain, pleasant/unpleasant), and arousal corresponds to intensity of experience, mirroring the amount of energy exhibited, spent, or available during a given emotional state. For example, ecstatic and serene are both emotions of positive valence but opposite arousal levels, just as angry and sad are both negative valence emotions of opposite arousal levels. Our own model can be seen as having valence and arousal as its foundation, but we focus on the patterns found in the transition dynamics between emotions through time.

Our moods data take us well beyond the basic polarity classifications or scales that dominate in computational sentiment analysis [11, 25, 26, 27, 30] and are thus more reminiscent of the multidimensional categories of [2, 14, 16, 22, 31, 35]. Even our star-ratings-based experiments with product reviews (Section 4) highlight the importance of having a 'neutral' category along the valence dimension [12], especially in communities that might be polarized in their opinions but nonetheless seek cohesion, a social force that can push evaluations away from the extremes.

We also study the social factors that influence the attitudes and emotions of community members. This can be thought of as part of a growing literature that seeks to understand online, user-provided metadata as both reflecting and shaping complex social processes concerning influence, group cohesion, individual assertion of identity, and sentiment diffusion in networks [1, 5, 8, 20, 21, 38]. Our general approach is also influenced by [24], who use social variables together with textual ones to predict expressed preferences in political speeches. A related model is used by [33] to capture the ways in which social relationships pattern with attitudes and evaluations. Our CRFs are conceptually different from the graph-based models in these papers, but they aim to capture similar insights about how sentiment predictions should be made partly on the basis of high-level emotional and contextual cues (see also [4]).

Our quantitative evaluations use conditional random fields (CRFs; [13, 32]), which are discriminative models in which it is possible to model complex dependencies not only between the input and output variables, as in typical classification problems, but also among the output variables themselves. General CRFs can model essentially any kind of graphical structure, but we concentrate on the more tractable special case of a linear-chain CRF, in which the output variables are arranged in a linear sequence, creating a discriminative counterpart to the generative Hidden Markov Model. Linear-chain CRFs have been used for a wide variety of sequence-labeling phenomena relating to linguistic structure, and they have been successfully applied to more social and contextual tasks of the sort we study here, including opinion-source identification [6], sentiment transitions inside documents [17], and social network extraction [7].

3. AFFECTIVE TRANSITIONS

We now address the personal, internal factors that shape the temporal flow of emotions people experience.

3.1 Experience Project Moods

Our 'moods' dataset is derived from community-member updates at the Experience Project (EP) social networking site. When logged in, EP members can post short texts describing their moods and also choose from a variety of different mood labels, which we use to classify the texts into emotional categories. Mood status updates are visible on member profiles. Table 1 gives an example of a sequence of three posts by the same user; the corpus contains additional metadata, but we use only the information in this table.

We work with a sample of about 2 million anonymized mood posts with unique author identifiers and 174 different mood labels for emotional, evaluative, and attitudinal states. Figure 1 summarizes the distribution of the top 20 moods by update frequency. This fragment of the full dataset shows that the corpus can be used to study sentiment analysis in the broadest terms, since individual pairs of labels or clusters of labels can be used to develop familiar polarity models [27], polarity models with different levels of intensity [3, 35, 36], models of sentiment based in social emotions like sympathy and solidarity [16, 31], and many others. For our purposes, the mix of valence and arousal with transitional emotions like anxiety and optimism is the most important aspect of the data set. In the next section, we study the entire distribution, seeking to motivate a high-level classification of the moods into distinct categories based on their transition relationships with other moods.

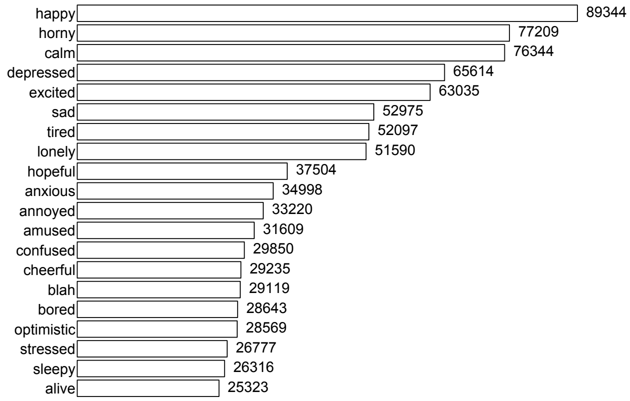


Figure 1: Top 20 mood labels by frequency, accounting for about 40% of the updates in our sample.

3.2 Mood Transitions

One's current emotional state will be heavily influenced by one's previous emotional state. The state before that will also exert an influence on the present, but less so. More generally, we expect previous states to influence current ones, with the influences weakening (becoming less direct) the farther back in time we go.

Time	Mood	Text
2013-07-28 11:56:56	sad	no one wants me . feeling sad cause i dont want me either
2013-07-28 22:41:40	lonely	Laying in this hospital bed I thought I wanted to be here I don't , take me home
2013-07-29 02:32:01	depressed	im sorry i need someone to talk to i need to not be a sub for 5 mins i just need a friend. please
		⋮

Table 1: A partial sequence of mood updates from a single user.

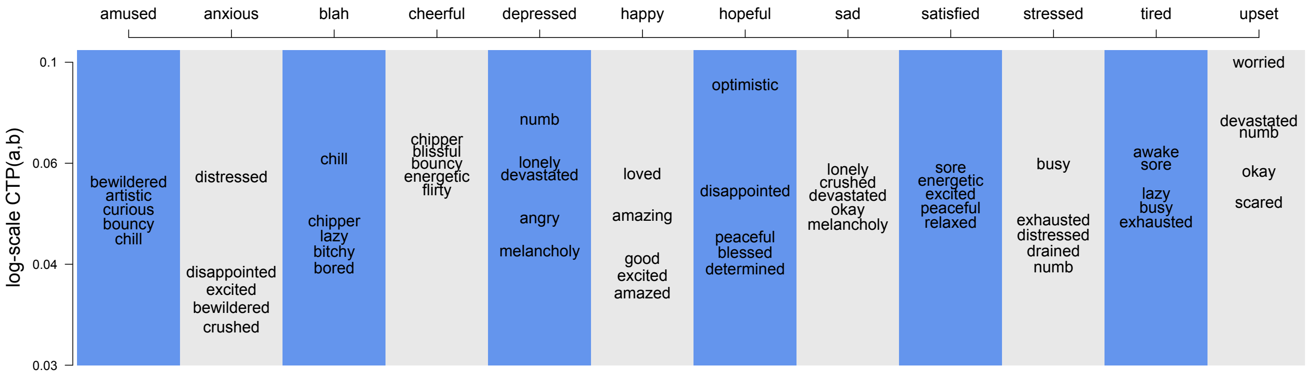


Figure 2: Mood compressed transition probabilities (*CTP* values). Each column labeled with emotion a shows the emotions b with largest *CTP*(a, b), as defined in equation 2.

To capture this pattern of influences, we first define emotional transition probabilities in a way that encompasses the full set of historical relationship between emotions. Let E be the full set of emotions, and let $C(a, t, b)$ be the number of times a user posted emotion $a \in E$ and then posted emotion $b \in E$ exactly t days later. Crucially, these transitions are counted whether or not there is an update between a and b. Then the conditional probability of b given a after t days is given by equation 1.

$$P\left(b \mid a, t\right) = \frac{C(a, t, b)}{\sum_{b' \in E} C(a, t, b')} \qquad (1)$$

Using these values, we then define the *compressed transition probability* between emotions a and b as follows:

$$CTP\left(a, b\right) = \left(c - 1\right) \sum_{t=0}^{\infty} \frac{P\left(b \mid a, t\right)}{c^{t+1}} \qquad (2)$$

Here, c is a dampening constant that gives a higher weight to recent transitions. In practice, c can range between 1.5 to 2.5 without significantly changing CTP values; in this paper, we use $c = 2$. We choose an exponential decay function so that distant transitions affect the CTP value without outweighing more recent transitions. Other dampening functions can also be used; the exact method seems not to matter as long as recent transitions are preferentially weighted and distant transitions are allowed to exert some influence.

The CTP measure has a second noteworthy advantage. There is significant diversity in the way EP members interact with the moods update feature. Some users post updates at roughly regular intervals, whereas others do so in less predictable patterns. In taking into account the temporal distance between updates, the CTP captures the way influence varies with time, making it more robust to this kind of behavioral variation than interval-blind methods are.

Figure 2 depicts a sample of $CTP(a, b)$ values. The conditioning emotions a are given along the top. For each, the emotions b with highest CTP to them are given in rank order, according to the log $CTP(a, b)$ values, normalized by the frequency of b to correct for differing usage levels. Self-transitions generally have dramatically larger CTP values than do other emotions, so those are not depicted. Frequency normalization and the log-scale of the y-axis are intended to make the plot more readable and to facilitate comparisons within and across columns.

The patterns in Figure 2 are intuitive, highlighting valence, arousal, and transition as important ingredients. For instance, the likely next states after a simple positive state like 'happy' have the same polarity, differing largely in intensity (and other social components). In contrast, the transition state 'hopeful', while positive in its own right, has a much more mixed set of later states associated with it, reflecting the uncertainty of this emotion. The emotions 'sad' and 'anxious' are rough negative duals of these: 'sad' emerges as a fairly standard negative valence category, whereas 'anxiety' is itself negative but transitions to both positive and negative outcomes. There are numerous other patterns like these in the data, reflecting not only emotional but also interactional and real-world factors (e.g., 'bored' to 'bitchy'; 'tired' to 'sore'; 'stressed' to 'busy').

To try to get a comprehensive, high-level view of how emotions relate to each other, we can also represent our entire data set as a directed graph based on the full matrix of CTP values, as in Figure 3. For any two emotions a and b, the edge from a to b has weight $CTP(a, b)$. The resulting graph represents the flow between emotional states of the mood updates population, thereby describing the affective space as a dynamic system. This abstraction is helpful for revealing the dynamic structure of the space. In particular, it is possible to use this graph to identify clusters of emotions that are densely connected with each other.

In order to cluster the emotions into disjoint partitions, we employ a variant of weighted label propagation [37] that

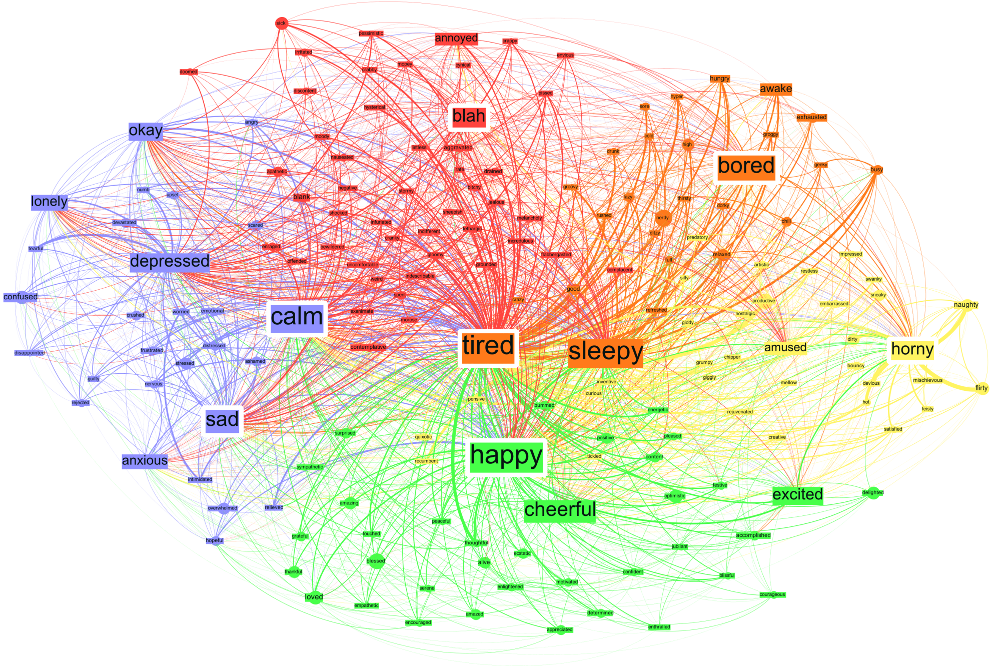


Figure 3: The moods represented as a weighted, directed graph in which edge labels are determined by *CTP* values, and cluster assignments (indicated by color) are also made based on those values. Node sizes are determined by the node's PageRank. White bounding boxes indicate prominent cluster moods.

is tailored specifically for highly dense graphs (by taking the average edge weight from each cluster instead of its sum). The main idea underpinning the algorithm is to identify groups of emotions with a higher than expected within-group transition probability. Unlike other weighted label propagation algorithms, our variant predetermines the number of clusters, and each emotion is iteratively assigned to the cluster from which the average CTP to it is the highest (as opposed to obtaining the label of the single heaviest incoming edge). The algorithm works as follows:

1. Initialize n clusters C_i randomly.
2. For each cluster C_i and each emotion e_j, compute the mean of the values $CTP(e_i, e_j)$ for $e_i \in C_i$ and assign e_j to the cluster with the highest such value.
3. Repeat 2 until the cluster assignments stabilize.

Self-transitions are dampened by multiplying them by a small constant s (here, 0.005) and normalizing accordingly. In our experience, large values of s (e.g., 0.01) generally lead to faster and more reliable convergence, but the resulting clusters are less compelling, whereas smaller values of s deliver substantially more plausible clusters but sometimes lead to non-convergence due to cyclic label assignments. We ran the algorithm 100 times using the parameters from Figure 3 ($s = 0.005$, 5 clusters) and it converged 67% of the time, with convergent runs taking an average of 25.40 iterations. Given that our primary goal is to explore the structure of the data set, these rates are more than acceptable.

Figure 3 uses color to represent a clustering of the mood space into five cells. Node sizes are scaled by a node's PageRank, which correlates with frequency but reduces the effect of self-transitions. Given the way in which the algorithm is defined, clusters have a higher than expected within-cluster transition probability. Thus, they can be interpreted as enduring emotional states. If we label them based on their prominent members (by PageRank), we get clusters 'tired/bored' (orange), 'horny' (yellow), 'sad/calm' (purple), 'happy' (green), and 'blah' (red), which is an intuitive set of groups based primarily on valence but further structured by other social and emotional factors. By varying the number of clusters, we achieve different levels of resolution. For example, with three clusters, the partitions show a near perfect division between negative, neutral/mixed, and positive valence emotions. With ten clusters, rarefied clusters emerge relating to erotic states, lack of energy ('lethargic', 'exhausted'), high-valence positive emotions ('ecstatic', 'delighted', 'blissful'), and so forth.

The directed nature of the graph also highlights the role emotions play in affective transitions. At a high level, we see the densest pathways between 'sad/calm' and 'blah' (purple, red), between 'blah' and 'tired/bored' (red, orange), and between 'happy' and 'horny' (green, yellow). The pathways are more asymmetrical between 'horny' and 'tired/bored', typically running from the first to the second rather than the other way around. Some pairwise paths are extremely unlikely in either direction. For instance, direct 'happy' to 'sad/calm' (green to purple) transitions are rare; emotional paths that begin with 'happy' and end with 'sad/calm' are likely to travel through other emotional states along the way. These patterns match well with our intuitions about human experiences, and they further support our contention that emotions are defined as much by their transitions as they are by their inherent properties.

3.3 Models

We apply CRFs as a modeling technique to capture emotion structure of the sort we just identified in our moods data. Specifically, we use linear-chain CRF models, as they efficiently represent temporal relationships among emotion labels as well as the mapping from text features of a document to its emotion label. Formally, we have a collection of document sequences $\mathcal{D}$, where each document sequence $d \in \mathcal{D}$ is a sequence of tuples, $d = [(x_1, e_1), (x_2, e_2), \ldots, (x_T, e_T)]$. Each tuple (x_t, e_t) is a vector of document features x_t along with the associated emotion label e_t. The sequence length T can vary for each sequence.

We construct a linear-chain CRF with potential functions $\phi_{j,k}(x_t, e_t)$ between an input feature $x_{t,j}$ and possible emotion $e_{t,k}$. Each $\phi_{j,k}$ is a binary indicator function, with $\phi_{j,k}(x_t, e_t) = 1$ when both feature j and label k are present in a document and 0 otherwise. This type of potential function is equivalent to the binary presence feature functions often used in sentiment classification [27]. However, the CRF goes beyond relationships between a single document's text features and its label. A second set of potential functions $\tau_{l,k}(e_{t-1}, e_t)$ serve as binary indicators for whether emotion l was present in the previous document at time $t - 1$ and emotion k is present in the current document at time t.

We use a standard log-linear parameterization, which makes the learning problem convex. Having defined our potential functions and chosen a log-linear CRF approach, our model derivation exactly follows those of other linear chain CRF models in the literature [13, 32]. We do not restate the likelihood and learning problem here; see [32] for the full details.

To evaluate the effect of modeling temporal emotion label relationships, we compare our CRF approach to a time-independent classification approach. This model treats each tuple (x_t, e_t) as independent from other documents in the sequence and estimates the emotion label probability $P(e_t \mid x_t)$. To estimate this probability we use a maximum entropy (MaxEnt) classifier. The MaxEnt model is a standard choice for classification tasks in sentiment analysis and other NLP tasks. Furthermore, the MaxEnt model is a special case of our linear chain CRF approach where we remove the potential functions $\tau_{l,k}(e_{t-1}, e_t)$. In removing these potential functions from our CRF, we are left with only potential functions relating the current text features x_t to the emotion label e_t. Comparing these two models allows us to identify the value in including sequence information.

We use the CRFsuite software package for both the CRF and MaxEnt models [23]. For both models, we use ℓ_2 regularization. We choose the regularization penalty by cross-validating over possible values and evaluating development set performance. We then run 20 trials in which we randomly split the data 80%/20% into training and testing sets and evaluate performance for both models. We use the non-parametric Wilcoxon rank-sums test to measure the significance of the difference between CRF and baseline performance. In order to focus on the effects of sequence structure, we restrict our textual features to simple unigrams.

3.4 Experimental Set-up

We conducted experiments on two different subsets of our moods data: a simple valence/polarity subset and a multidimensional subset involving valence/polarity, arousal, and affective transitions. The polarity experiments are relatively easy to interpret and permit general comparisons with more familiar sentiment tasks, while the multidimensional experiments highlight the nuances of our moods data and reveal more of the power of affective transitions.

We built a polarity label set by clustering high-volume moods with unambiguous valence, drawn from the green and purple clusters in Figure 3:

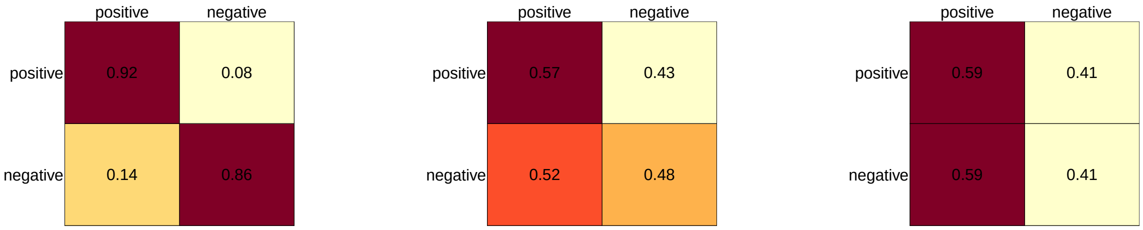
- 'positive' = {'happy', 'excited', 'thankful'}
- 'negative' = {'sad', 'lonely', 'depressed', 'angry'}

Figure 4 summarizes the transition structure of the polarity data set. Each square depicts the probability of seeing a sequence consisting of the row polarity and then the column polarity. The overall structure of the polarity transitions is clear: polarity states are generally enduring, with a slight trend towards positivity. One concern one might raise is that this trend could reflect, not affective transitions, but rather an overall tendency for individuals to be consistently positive or consistently negative. With Figure 4(b), we seek to address this by limiting attention just to high-variance sequences, defined as the top 20% of polarity sequences by label diversity. Thus, these sequences definitely come from users with variable emotional states. The same structure is evident here, though with the additional finding that, for these variable individuals, transitions from 'negative' to 'positive' are more likely than the other way around. As another point of comparison, we include, in Figure 4(c), a version of these heatmaps in which the mood sequences have been randomly shuffled. This destroys all transition structure, revealing instead only the percentages of 'positive' and 'negative' items in the data set.

For our multidimensional experiments, we would ideally study the entire set of moods, since Figure 3 suggests that they have intricate internal structure. However, the resulting models would be intractable to analyze and impractical to report on. Thus, we hand-selected six moods, with the goal of capturing valence, arousal, and the transitional characteristics of emotions, thereby preserving many of the rich inter-emotion dynamics of the full set. The moods we chose for this are 'cheerful', 'satisfied', 'hopeful', 'anxious', 'stressed', and 'depressed'.

Although we focus only on a subset of the moods presented in Figure 2, this visualization provides insights into the transition structure we hypothesize that our CRF can leverage. Overall, since some emotions are more common than others, the label distribution for the multidimensional sentiment classification experiments are not uniform. The emotions 'depressed', 'hopeful', and 'stressed' are the most common, and 'satisfied', 'anxious', and 'cheerful' are the least common. These differing relative frequencies probably reflect deeper underlying facts about the frequency or duration of these experiences.

In Section 3.2, we argued that one's current emotion is affected by one's previous emotions, with influences diminishing rapidly over time. The structure of our CRF model forces us to make a simplifying Markov assumption that only the previous state influences the current one. Since this discretizes the time element, we impose an additional restriction that any two consecutive mood posts in mood sequences can be no more than 24 hours apart, to ensure that we model true affective transitions rather than individuals' general behavior on the site (e.g., users who log on infrequently and only when anxious). After enforcing this restriction, our polarity data set consists of 30,000 sequences containing approximately 70,000 posts, and our multidimensional mood

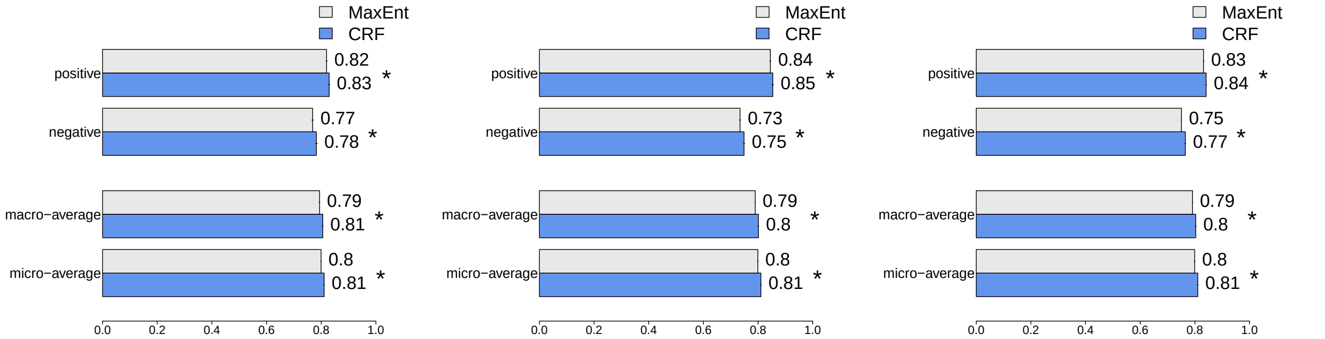


(a) All actual sequences.

(b) High-variance sequences.

(c) Randomized sequences (averaged over 20 randomizations).

Figure 4: Mood polarity transition probabilities, from the row state to the column state.

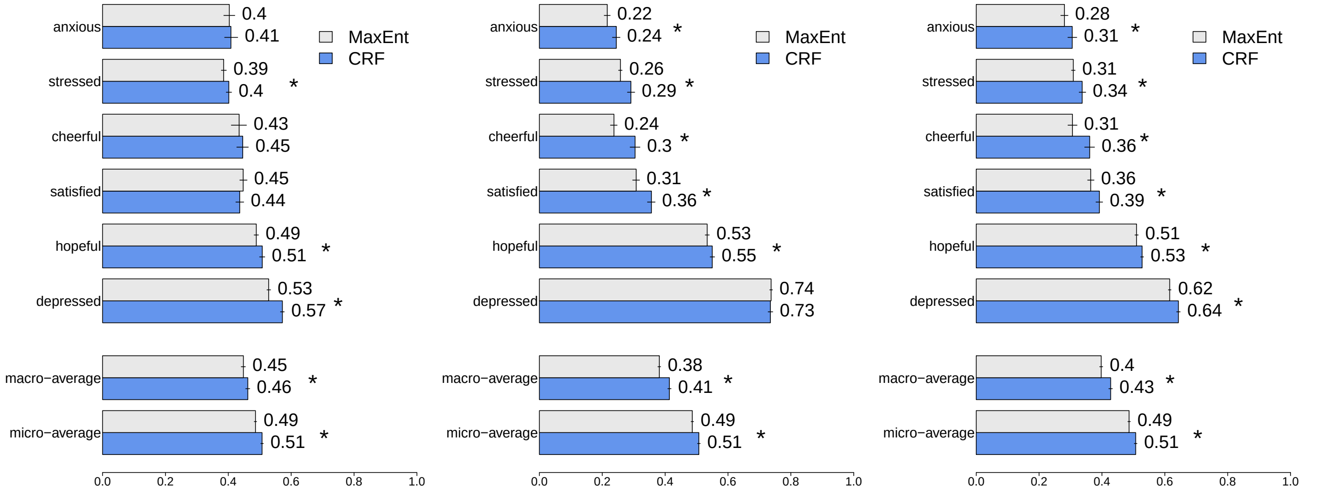


(a) Precision.

(b) Recall.

(c) F1.

Figure 5: Moods polarity performance with bootstrapped 95% confidence intervals (often very small). Stars mark statistically significant differences ($p < 0.001$) according to a Wilcoxon rank-sums test.



(a) Precision.

(b) Recall.

(c) F1.

Figure 6: Multidimensional moods performance with bootstrapped 95% confidence intervals (often very small). Stars mark statistically significant differences ($p < 0.001$) according to a Wilcoxon rank-sums test.

dataset consists of approximately 20,000 sequences containing 60,000 posts overall.

3.5 Results

The polarity and multidimensional CRF models achieve absolute micro-average F1 gains of 2% over the MaxEnt baselines, and performance on individual labels improves across the board. All gains are statistically significant ($p < 0.001$) according to a Wilcoxon rank-sums test. Figure 5 summarizes polarity results, and Figure 6 summarizes multidimensional moods results. The results are presented with bootstrapped 95% confidence intervals [9], often very small because of the large amount of data and consistent model performance.

Since the multidimensional moods model must navigate a decision space of six labels, its overall performance is lower than that of the two-label polarity model, and inter-label performance disparities are greater. The advantage of the multidimensional moods model, however, is its ability to identify granular emotions that are richer indications of a person's current and future affective states than the non-specific classes 'positive' and 'negative'. For example, 'anxious' and 'depressed' are both negative valence mood states, but they have different implications for a user's future state: 'anxious' is a transitory mood state that tends towards resolution, either positive or negative, whereas 'depressed' is a more enduring affective state with more limited possible future states. In many applications, these distinctions are particularly relevant. In the customer experience industry, for example, it is often most valuable not to characterize a user's valence but rather to identify moments of emotional inflection, which mark the points of volatility when a user is in the process of changing opinions or attitudes.

One measure of a mood's volatility is the entropy of the probability distributions over previous and subsequent mood states, referred to as $H(\mathit{prev})$ and $H(\mathit{next})$, respectively. Moods with high $H(\mathit{prev})$ and $H(\mathit{next})$ are crossroads, in that they arise from and transition to diverse sets of other mood states. Because such moods are transitory, they are relatively infrequent and therefore systematically harder to model. For instance, a mood's $H(\mathit{prev})$ is inversely related to its volume (linear regression; $R^2 = 0.80$, $p < 0.05$), and a mood's volume is inversely related to its F1 score (linear regression; $R^2 = 0.86$, $p < 0.01$). Nonetheless, the CRF models are notably better at identifying them than the MaxEnt models are, especially in terms of recall (linear regression; $R^2 = 0.71$, $p < 0.05$).

For example, 'anxious' is the most high-entropy mood state of the six we model, and it is the second most infrequent. Confusion matrices show that the MaxEnt baseline frequently misclassifies 'anxious' posts: given sparse or ambiguous 'anxious' documents, the contextually-unaware MaxEnt baseline tends toward the more frequent classes 'stressed' and 'depressed'. Our CRF, however, is less susceptible to this mistake because it incorporates knowledge of the particular transition characteristics of 'anxious'. CRF gains in recall for 'anxious' are above average, and, conversely, precision gains for 'stressed' and 'depressed' are also higher than average.

Overall, these experiments validate that a CRF model can incorporate the affective transition structure described in Section 3.2 to improve performance on both the more conventional polarity prediction task and the more complicated multidimensional prediction task. The strong performance gains on infrequent labels, which are often of particular interest, is reflected by the macro-average F1 score, which shows greater CRF gains over the baseline than the micro-average F1 score.

4. SOCIAL FORCES

We now extend the above model of affective transitions to a social setting in which the evidence suggests that current sentiment evaluations are influenced by previous ones.

4.1 RateBeer Data

We develop our basic model of social influences on sentiment evaluations using the RateBeer corpus, a collection of 2.9 million user-supplied beer reviews.[2] With respect to its structure, this corpus is like most corpora of user-supplied product reviews. The review texts are typically just a few sentences long and have associated with them a number of different kinds of rating, including aspects of the beer (e.g., aroma, palate, taste) and an overall rating. (For discussion of these multi-aspect ratings, see [19].) The overall ratings are on the scale 1–20, which we rescale into a more familiar space of 5 star ratings (with fractions of a star possible).

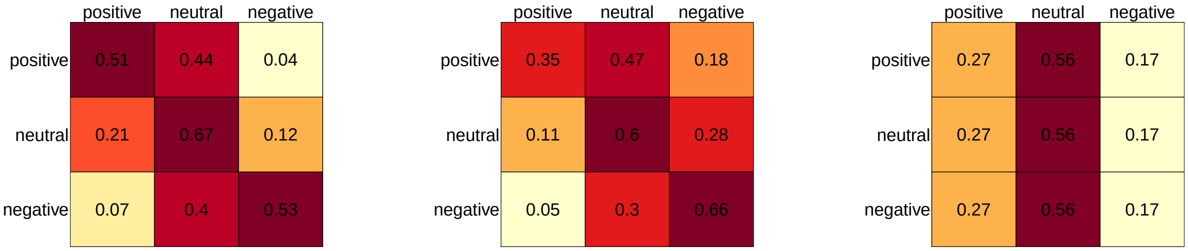
4.2 Social forces

The RateBeer corpus reflects complex social phenomena that make it especially useful for our purposes. First, individual community members often write many reviews (4,798 wrote more than 50; [8]). Second, we know from [18] that reviewers themselves vary greatly in their level of expertise about beers. Third, we know from [8] that the community dynamics of the site are complex, with specific kinds of users influencing the evaluative and communicative norms in complex ways that find many counterparts in off-line communities. These factors lead us to expect reviewers to be influenced by each other when making rating choices. For instance, if the current reviewer is new to the site, she might shift her judgments towards those of more expert members who already posted reviews. Conversely, experts might feel the need to push back against overly positive or negative reviews by newcomers. We do not at present have a deep understanding of the social forces at work, but we require only that forces like these be active.

Figure 8 provides additional information about the rating distribution and how we construe it. For the purposes of our classification experiments, 'positive' reviews are those with ratings 4 or above, 'negative' reviews are those with ratings 2.5 or below, and 'neutral' ones fall in the middle, as indicated in the figure. These boundaries were chosen with the underlying rating distribution in mind. The distribution is noteworthy for being much more skewed to the middle of the rating scale than is typical for corpora of user-supplied reviews, which are usually dominated by positive reviews ([26], p. 74). We think that this too reflects the underlying social dynamics of the site: extreme evaluations (at either end of the scale) are socially riskier; for example, one doesn't want to appear too glowing about a beer that is perceived as mundane, nor too critical of one that is perceived as requiring expertise.

Figure 7 provides a high-level picture of the extent of these influences, using the label categories indicated in Figure 8.

[2] `http://snap.stanford.edu/data/web-RateBeer.html`



(a) All actual sequences.

(b) High-variance products.

(c) Randomized sequences (averaged over 20 randomizations).

Figure 7: By-product transition probabilities, from the row state to the column state.

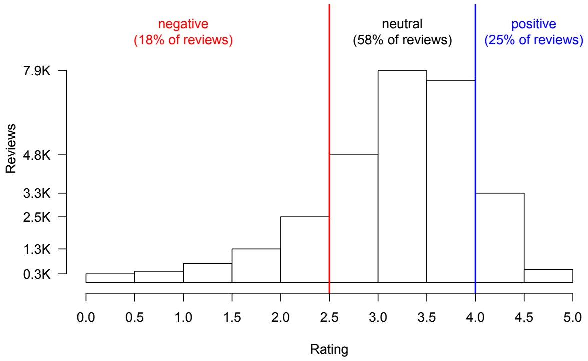


Figure 8: RateBeer ratings and label categories.

Each square depicts the probability of seeing a sequence consisting of the row evaluation and then the column evaluation, using the probability model from our moods study (Section 3.2). The overall picture is easily characterized: self-transitions are most likely for all three categories, polarity reversing sequences are least-likely, and 'neutral'-to-'positive' sequences are more likely than 'neutral'-to-'negative'. This pattern makes sense if we assume that there are generally social costs to disagreeing with others. Importantly, Figure 7(b) shows that the general pattern holds even when we restrict attention to products with extremely highly variable ratings, defined as those with rating standard deviation in the third quartile for those values. This suggests that the patterns are not merely a consequence of overall community agreement. Indeed, the higher-variance reviews seem to reflect the ways in which disagreements are negotiated. Figure 7(c) shows that the pattern disappears when we randomize the sequences; the only remaining pattern is fully explained by the overall label distribution.

4.3 Models

The CRF model is the same as it was for the moods data (see Section 3.3), except label sequence features are now product-level rating sequences. Once again, we compare the CRFs to MaxEnt classifiers with the same textual features. For both, regularization parameters are optimized independently using cross-validation on development sets.

4.4 Experimental Set-up

To study the gains achieved by incorporating social influences into our model, we run 20 classification trials. In each trial, we randomly sample sequences such that the data set consists of roughly 500,000 reviews. We train on 80% of the sampled sequences and test on the remaining 20%.

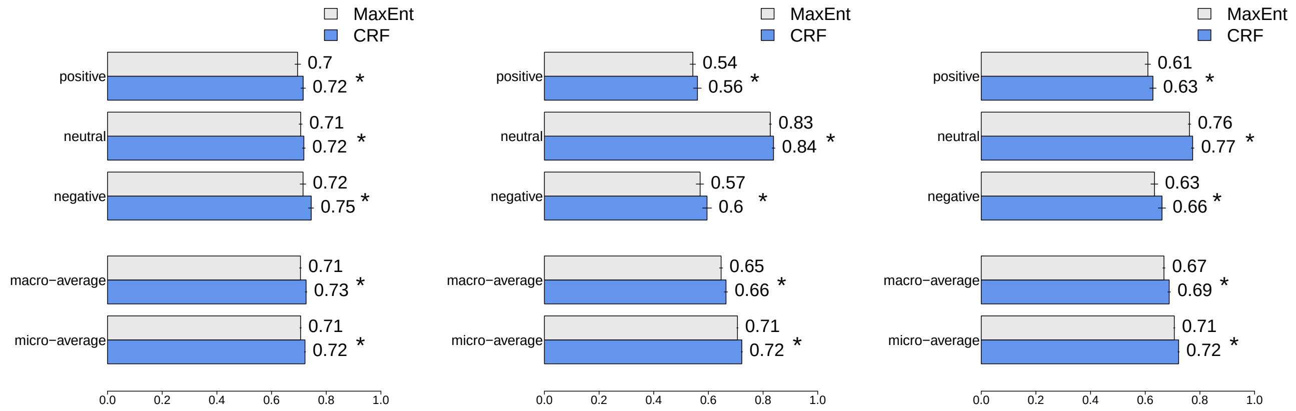
4.5 Results

The CRF model achieves an absolute micro-average F1 gain of 1% over the MaxEnt baseline, which translates to roughly 1,000 additional correct classifications over the baseline. Figure 9 summarizes by-label performance for both the MaxEnt and CRF models.

Although performance improves for all labels, we observe stronger performance gains for 'positive' and 'negative' documents than 'neutral' ones. This trend is reflective of the sequence structures described in Figure 7: the probability distributions of states that transition to and arise from 'positive' and 'negative' states are more discriminating than those for 'neutral' states. 'Negative' reviews are highly unlikely to follow 'positive' ones and vice versa, and 'neutral' is often the intermediate step in sequences that do transition from one polarity to the other. As a result, given an ambiguous or sparse review that the MaxEnt baseline may, for example, predict to be 'neutral' or 'negative' with equal likelihood, the CRF model can use the previous review's predicted class to help discriminate. In this example, if the previous review is predicted to be 'positive', the CRF will be particularly unlikely to decide that the ambiguous review in question is 'negative'.

5. CONCLUSION

Many sentiment models and technologies make the simplifying assumption that each sentiment expression (emotion, evaluation, perspective) was produced independently of all others. Our central observation is that this simplifying assumption ignores information that is psychologically and socially important, particularly for inherently transitional emotions like hope and anxiety. Using linear-chain CRFs, we showed furthermore that this information has significant predictive value, not only for multidimensional sentiment data, but also in more traditional polarity tasks. To highlight the independent value of internal, cognitive factors and external, social ones, we kept these two kinds of influence apart in our models, but the CRF approach is flexible enough to accommodate the complex graphical structure needed to bring these two kinds of influence together into a single predictive model. We look forward to future experi-



(a) Precision. (b) Recall. (c) F1.

Figure 9: RateBeer results with bootstrapped 95% confidence intervals (often very small). Stars mark statistically significant differences ($p < 0.001$) according to a Wilcoxon rank-sums test.

ments that combine personal affective transitions and social forces to further improve sentiment prediction.

Acknowledgements

We thank Experience Project for providing us with the sample of moods data and for being encouraging about our research. This work was supported in part by ONR grant N00014-13-1-0287 and NSF grant IIS-1016909.

References

[1] A. Aji and E. Agichtein. The “nays” have it: Exploring effects of sentiment in collaborative knowledge sharing. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 1–2, Los Angeles, California, USA, June 2010. Association for Computational Linguistics.

[2] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[3] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era (NLPIX)*, Beijing, China, 2008.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, June 2007. Association for Computational Linguistics.

[5] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J. A. Hołyst. Collective emotions online and their influence on community life. *PLoS ONE*, 6(7):e22207, July 2011.

[6] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, October 2005. Association for Computational Linguistics.

[7] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *Conference on Email and Spam*, 2004.

[8] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International World Wide Web Conference*, pages 307–317, New York, 2013. ACM.

[9] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–57, February 1986.

[10] L. Feldman Barrett and J. A. Russell. Independence and bipolarity in the structure of affect. *Journal of Personality and Social Psychology*, 74(4):967–984, 1998.

[11] A. B. Goldberg and J. Zhu. Seeing stars when there aren’t many stars: Graph-based semi-supervised leaarning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*, 2006.

[12] M. Koppel and J. Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, pages 282–289, 2001.

[14] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of Intelligent User Interfaces (IUI)*, pages 125–132, 2003.

[15] H. Lövheim. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2):341–348, 2012.

[16] A. Maas, A. Ng, and C. Potts. Multi-dimensional sentiment analysis with learned representations. Ms., Stanford University, 2011.

[17] Y. Mao and G. Lebanon. Isotonic conditional random fields and local sentiment flow. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 961–968, Cambridge, MA, 2006. MIT Press.

[18] J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International World Wide Web Conference*, pages 897–907, New York, 2013. ACM.

[19] J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *12th International Conference on Data Mining*, pages 1020–1025, Washington, D.C., 2012. IEEE Computer Society.

[20] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts. Sentiment flow through hyperlink networks. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, July 2011. Association for the Advancement of Artificial Intelligence.

[21] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[22] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 806–814, Beijing, China, August 2010. COLING 2010 Organizing Committee.

[23] N. Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[24] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Barcelona, Spain, July 2004.

[25] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, MI, June 2005. Association for Computational Linguistics.

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

[27] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, July 2002. Association for Computational Linguistics.

[28] D. C. Rubin and J. M. Talerico. A comparison of dimensional models of emotion. *Memory*, 17(8):802–808, 2009.

[29] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[30] B. Snyder and R. Barzilay. Multiple aspect ranking using the Good Grief algorithm. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pages 300–307, 2007.

[31] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK., July 2011. ACL.

[32] C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[33] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1397–1405, San Diego, CA, August 2011. ACM Digital Library.

[34] D. Watson and A. Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235, 1985.

[35] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210, 2005.

[36] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? Finding strong and weak opinion clauses. *Computational Intelligence*, 2(22):73–99, 2006.

[37] Y. Yang, C. Chen, and S. Pang. WLPA: A novel algorithm to detect community structures in social networks. *Journal of Computational Information Systems*, 7(2):515–522, 2011.

[38] R. Zafarani, W. Cole, and H. Liu. Sentiment propagation in social networks: A case study in livejournal. In S.-K. Chai, J. Salerno, and P. Mabry, editors, *Advances in Social Computing*, volume 6007 of *Lecture Notes in Computer Science*, pages 413–420. Springer Berlin / Heidelberg, 2010.