Recognizing Rare Social Phenomena in Conversation: Empowerment Detection in Support Group Chatrooms

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
{emayfiel, dadamson, cprose}@cs.cmu.edu

Abstract

Automated annotation of social behavior in conversation is necessary for large-scale analysis of real-world conversational data. Important behavioral categories, though, are often sparse and often appear only in specific subsections of a conversation. This makes supervised machine learning difficult, through a combination of noisy features and unbalanced class distributions. We propose *within-instance content selection*, using cue features to selectively suppress sections of text and biasing the remaining representation towards minority classes. We show the effectiveness of this technique in automated annotation of empowerment language in online support group chatrooms. Our technique is significantly more accurate than multiple baselines, especially when prioritizing high precision.

1 Introduction

Quantitative social science research has experienced a recent expansion, out of controlled settings and into natural environments. With this influx of interest comes new methodology, and the inevitable question arises of how to move towards testable hypotheses, using these uncontrolled sources of data as scientific lenses into the real world.

The study of conversational transcripts is a key domain in this new frontier. There are certain social and behavioral phenomena in conversation that cannot be easily identified through questionnaire data, self-reported surveys, or easily extracted user metadata. Examples of these social phenomena in conversation include overt displays of power (Prabhakaran et al., 2012) or indicators of rapport and relationship building (Wang et al., 2012). Manually annotating these social phenomena cannot scale to large data, so researchers turn to automated annotation of transcripts (Rosé et al., 2008). While machine learning is highly effective for annotation tasks with relatively balanced labels, such as sentiment analysis (Pang and Lee, 2004), more complex social functions are often rarer. This leads to unbalanced class label distributions and a much more difficult machine learning task. Moreover, features indicative of rare social annotations tend to be drowned out in favor of features biased towards the majority class. The net effect is that classification algorithms tend to bias towards the majority class, giving low accuracy for rare class detection.

Automated annotation of social phenomena also brings opportunities for real-world applications. For example, real-time annotation of conversation can power adaptive intervention in collaborative learning settings (Rummel et al., 2008; Adamson and Rosé, 2012). However, with the considerable power of automation comes great responsibility. It is critical to avoid intervening in the case of erroneous annotations, as providing unnecessary or inappropriate support in such a setting has been shown to be harmful to group performance and social cohesion (Dillenbourg, 2002; Stahl, 2012).

We propose adaptations to existing machine learning algorithms which improve recognition of rare annotations in conversational text data. Our primary contribution comes in the form of *within-instance content selection*. We develop a novel algorithm based on textual cues, suppressing information which is likely to be irrelevant to an instance's class label. This allows features which predict minority classes to gain prominence, helping to sidestep the frequency of common features pointing to a majority class label.

Additionally, we propose modifications to existing algorithms. First, we identify a new application of logistic model trees to text data. Next,

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 104–113,
Sofia, Bulgaria, August 4-9 2013.

we define a modification of confidence-based ensemble voting which encourages minority class labeling. Using these techniques, we demonstrate a significant improvement in classifier performance when recognizing the language of *empowerment* in support group chatrooms, a critical application area for researchers studying conversational interactions in healthcare (Uden-Kraan et al., 2009).

The remainder of this paper is structured as follows. We introduce the domain of empowerment in support contexts, along with previous studies on the challenges that these annotations (and similar others) bring to machine learning. We introduce our new technique for improving the ability to automate this annotation, along with other optimizations to the machine learning workflow which are tailored to this skewed class balance. We present experimental results showing that our method is effective, and provide a detailed analysis of the behavior of our model and the features it uses most. We conclude with a discussion of particularly useful applications of this work.

2 Background

We ground this paper's discussion of machine learning with a real problem, turning to the annotation of empowerment language in chat[1]. The concept of empowerment, while a prolific area of research, lacks a broad definition across professionals, but broadly relates to "the power to act efficaciously to bring about desired results" (Boehm and Staples, 2002) and "experiencing personal growth as a result of developing skills and abilities along with a more positive self-definition" (Staples, 1990). Participants in online support groups feel increased empowerment (Uden-Kraan et al., 2009; Barak et al., 2008). Quantitative studies have shown the effect of empowerment through statistical methods such as structural equation modeling (Vauth et al., 2007), as have qualitative methods such as deductive transcript analysis (Owen et al., 2008) and interview studies (Wahlin et al., 2006).

The transition between these styles of research has been gradual. Pioneering work has demonstrated the ability to distinguish empowerment language in written texts, including prompted writing samples (Pennebaker and Seagal, 1999), narratives in online forums (Hoybye et al., 2005), and some preliminary analysis of synchronous discussion (Ogura et al., 2008; Mayfield et al., 2012b). These transitional works have used limited analysis methodology; in the absence of sophisticated natural language processing, their conclusions often rely on coarse measures, such as word counts and proportions of annotations in a text.

Users, of course, do not express empowerment in every thread in which they participate, which leads to a challenge for machine learning. Threads often focus on a single user's experiences, in which most participants in a chat are merely commentators, if they participate at all, matching previous research on shifts in speaker salience over time (Hassan et al., 2008). This leads to many user threads which are annotated as not applicable (N/A). We move to our proposed approach with these skewed distributions in mind.

Table 1: Empowerment label distribution in our corpus.

Annotation	Label	#	%
Self-Empowerment	NA	1522	79.3
	POS	202	10.5
	NEG	196	10.2
Other-Empowerment	NA	1560	81.3
	POS	217	11.3
	NEG	143	7.4

3 Data

Our data consists of a set of chatroom conversation transcripts from the Cancer Support Community[2]. Each 90-minute conversation took place in the context of a weekly meeting in a real-time chat, with up to 6 participants in addition to a professional therapist facilitating the discussion. In total, 2,206 conversations were collected from 2007-2011. This data offers potentially rich insight into coping and social support; however, annotating such a dataset by hand would be prohibitively expensive, even when it is already transcribed.

Twenty-one of these conversations have been annotated, as originally described and analyzed in (Mayfield et al., 2012b)[3]. This data was disentangled into *threads* based on common themes or topics, as in prior work (Elsner and Charniak,

[1] Definitions of empowerment are closely related to the notion of *self-efficacy* (Bandura, 1997). For simplicity, we use the former term exclusively in this paper.

[2] www.cancersupportcommunity.org

[3] All annotations were found to be adequately reliable between humans, with thread disentanglement $f = 0.75$ and empowerment annotation $\kappa > 0.7$.

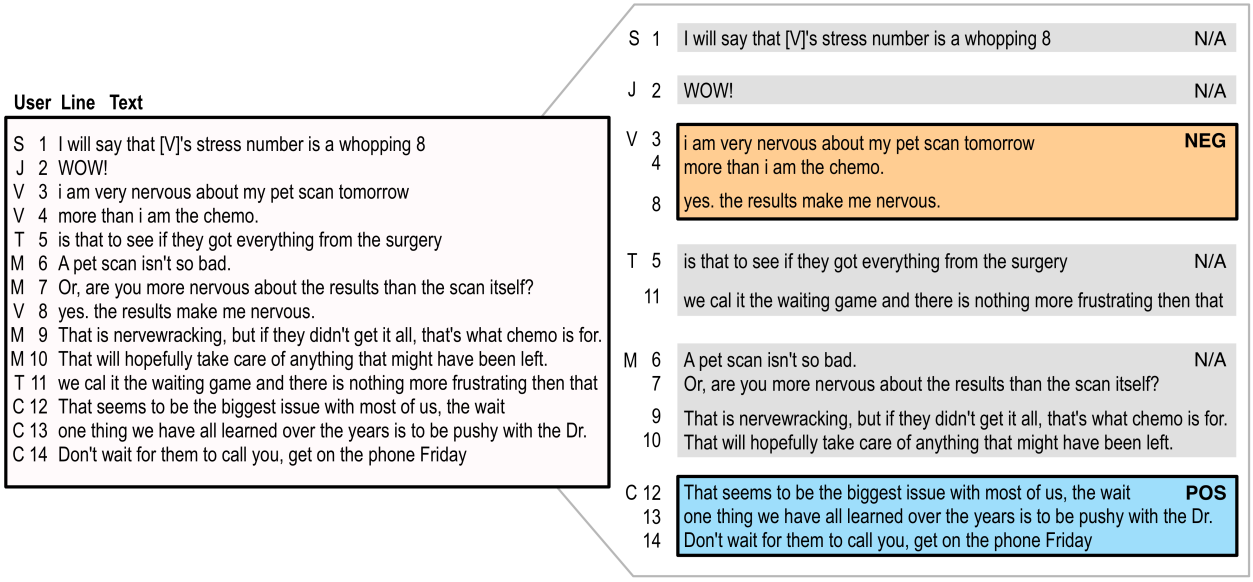

Figure 1: An example mapping from a single thread's chat lines (left) to the per-user, per-thread instances used for classification in this paper (right), with example annotations for self-empowerment indicated.

2010; Adams and Martel, 2010). A novel *per-user*, *per-thread* annotation was then employed for empowerment annotation, following a coding manual based on definitions like those in Section 2. Each user was assigned a label of positive or negative empowerment if they exhibited such emotions, or was left blank if they did not do so within the context of that thread. This annotation was performed both for their *self-empowerment* as well as their attitude towards others' situations (*other-empowerment*). An example of this annotation for self-empowerment is presented in Figure 1 and the distribution of labels is given in Table 1.

Most previous annotation tasks attempt to annotate on a per-utterance basis, such as dialogue act tagging (Popescu-Belis, 2008), or on arbitrary spans of text, such as in the MPQA subjectivity corpus (Wiebe et al., 2005). However, for our task, a per-user, per-thread annotation is more appropriate, because empowerment is often indicated best through narrative (Hoybye et al., 2005). Human annotators are instructed to take this context into account when annotating (Mayfield et al., 2012b). It would therefore be nonsensical to annotate individual lines as "embodying" empowerment. Similar arguments have been made for sentiment, especially as the field moves towards aspect-oriented sentiment (Breck et al., 2007). Assigning labels based on thread boundaries allows for context to be meaningfully taken into account, without crossing topic boundaries.

However, this granularity comes with a price: the distribution of class values in these instances is highly skewed. In our data, the vast majority of users' threads are marked as not applicable to empowerment. Perhaps more inconveniently, while taking context into account is important for reliable annotation, it leads to extraneous information in many cases. Many threads can have multiple lines of contributions that are topically related to an expression of empowerment (and thus belong in the same thread), but which do not indicate any empowerment themselves. This exacerbates the likelihood of instances being classified as N/A.

We choose to take advantage of these attributes of threads. We know from research in discourse analysis that many sections of conversations are formulaic and rote, like introductions and greetings (Schegloff, 1968). We additionally know that polarity often shifts in dialogue through the use of discourse connectives such as conjunctions and transitional phrases. These issues have been addressed in work in the language technologies community, most notably through the Penn Discourse Treebank (Prasad et al., 2008); however, their applications to noisier synchronous conversation has beenrare in computational linguistics.

With these linguistic insights in mind, we examine how we can make best use of them for machine learning performance. While techniques for predicting rare events (Weiss and Hirsh, 1998) and compensating for class imbalance (Frank and

Bouckaert, 2006), these approaches generally focus on statistical properties of large class sets without taking the nature of their datasets into account. In the next section, we propose a new algorithm which takes advantage specifically of the linguistic phenomena in the conversation-based data that we study for empowerment detection. As such, our algorithm is highly suited to this data and task, with the necessary tradeoff in uncertain generality to new domains with unrelated data.

4 Cue Discovery for Content Selection

Our algorithm performs content selection by learning a set of cue features. Each of these features indicates some linguistic function within the discourse which should downplay the importance of features either before or after that discourse marker. Our algorithm allows us to evaluate the impact of rules against a baseline, and to iteratively judge each rule atop the changes made by previous rules.

This algorithm fits into existing language technologies research which has attempted to partition documents into sections which are more or less relevant for classification. Many researchers have attempted to make use of cue phrases (Hirschberg and Litman, 1993), especially for segmentation both in prose (Hearst, 1997) and conversation (Galley et al., 2003). The approach of content selection, meanwhile, has been explored for sentiment analysis (Pang and Lee, 2004), where individual sentences may be less subjective and therefore less relevant to the sentiment classification task. It is also similar conceptually to content selection algorithms that have been used for text summarization (Teufel and Moens, 2002) and text generation (Sauper and Barzilay, 2009), both of which rely on finding highly-relevant passages within source texts.

Our work is distinct from these approaches. While we have coarse-grained annotations of empowerment, there is no direct annotation of what makes a good cue for content selection. With our cues, we hope to take advantage of shallow discourse structure in conversation, such as contrastive markers, making use of implicit structure in the conversational domain.

4.1 Notation

Before describing extensions to the baseline logistic regression model, we define notation. Our data is arranged hierarchically. We assume that we have a collection of d training documents $\mathbf{Tr} = \{D_1 \dots D_d\}$, each of which contains many training instances (in our task, an instance consists of all lines of chat from one user in one thread). Our total set of n instances $\mathbf{I}$ thus consists of instances $\{I_1, I_2, \dots I_n\}$. Each document contains lines of chat $\mathbf{L}$ and each instance I_i is comprised of some subset of those lines, $L_i \subseteq \mathbf{L}$.

Our feature space $\mathbf{X} = \{x_1, x_2, \dots x_m\}$ consists of m unigram features representing the observed vocabulary used in our corpus. Each instance is associated with a feature vector $\bar{x}$ containing values for each $x \in \mathbf{X}$, and each feature x that is present in the i-th instance maintains a "memory" of the lines in which it appeared in that instance, L_{ix}, where $L_{ix} \subseteq L_i$. Our potential output labels consist of $\mathbf{Y} = \{NA, NEG, POS\}$, though this generalizes to any nominal classification task. Each instance I is associated with exactly one $y \in \mathbf{Y}$ for self-empowerment and one for other-empowerment; these two labels do not interact and our tasks are treated as independent in this paper[4]. We define classifiers as functions $f(\bar{x} \rightarrow y \in \mathbf{Y})$; in practice, we use logistic regression via LibLINEAR (Fan et al., 2008).

We define a content selection rule as a pairing $r = \langle c, t \rangle$ between a cue feature $c \in \mathbf{X}$ and a selection function $t \in T$. We created a list of possible selection functions, given a cue c, maximizing for generality while being expressive. These are illustrated in Figure 2 and described below:

- **Ignore Local Future (A)**: Ignore all features from the two lines after each occurrence of c.
- **Ignore All Future (B)**: Ignore all features occurring after the first occurrence of c.
- **Ignore Local History (C)**: Ignore all features in the two lines preceding each occurrence of c.
- **Ignore All History (D)**: Ignore all features occurring only before the last occurrence of c.

We define an ensemble member $E = \langle R, f_R \rangle$ - the ordered list of learned content selection rules $R = [r_1, r_2, \dots]$ and a classifier f_C trained on instances transformed by those rules. Our final out-

[4]Future work may examine the interaction of jointly annotating multiple sparse social phenomena.

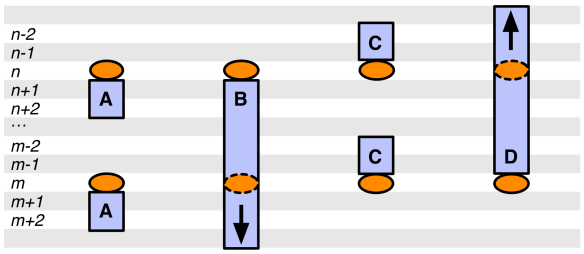

Figure 2: Effects of content selection rules, based on a cue feature (ovals) observed at lines m and n.

put of a trained model is a set of ensemble members $\{E_1, \ldots, E_k\}$.

4.2 Algorithm

Our ensemble learning follows the paradigm of cross-validated committees (Parmanto et al., 1996), where k ensemble members are trained by subdividing our training data into k subfolds. For each ensemble classifier, cue rules R are generated on $k-1$ subfolds ($\mathbf{Tr}_k$) and evaluated on the remaining subfold ($\mathbf{Te}_k$). In practice, with 21 training documents, 7-fold cross-validation, and $k = 3$ ensemble members, each generation set consists of 12 documents' instances, while each evaluation set contains instances from 6 documents.

Our full algorithm is presented in Algorithm 1, and is broken into component parts for clarity. Algorithm 2 begins by measuring the baseline classifier's ability to recognize minority-class labels. After training on $\mathbf{Tr}_k$, we measure the average probability assigned to the correct label of instances in $\mathbf{Te}_k$, but only for instances whose correct labels are minority classes (remember, because both $\mathbf{Tr}_k$ and $\mathbf{Te}_k$ are drawn from the overall $\mathbf{Tr}$, we have access to true class labels). We choose this subset of only minority instances, as we are not interested in optimizing to the majority class.

We next enumerate all rules that we wish to judge. To keep this problem tractable, we ignore features which do not occur in at least 5% of training instances. For the remaining features, we create a candidate rule for each possible pairing of features and selection functions. For each of these candidates, we test its utility by selecting content as if it were an actual rule, then building a new classifier (trained on the generation set) using instances that have been altered in that way. In the evaluation set, we measure the difference in probability of minority class labels being assigned correctly between the baseline and this altered space. This measure of an individual rule's impact is described in Algorithm 3.

Once we have evaluated every possible rule once, we select the top-ranked rule and apply it to the feature set. We then iteratively progress through our now-ranked list of candidates, each time treating the newly filtered dataset as our new baseline. We search only top candidates for efficiency, following the fixed-width search methodology for feature selection in very high-dimensionality feature spaces (Gütlein et al., 2009). Each ensemble classifier is finally retrained on all training data, after applying the corresponding content selection rules to that data.

5 Prediction

Our prediction algorithm begins with a standard implementation of cross-validated committees (Parmanto et al., 1996), whose results are aggregated with a confidence voting method intended to favor rare labels (Erp et al., 2002). Cross-validated committees are an ensemble technique used to subsample training data to produce multiple hypotheses for classification. Each classifier produced by our cue-based transformation is trained on a subset of our training data. Each makes predictions on all test set instances, producing a distribution of confidence across possible labels. These values serve as inputs to a voting method to produce a final label for each instance.

Compared to other ensemble methods, cross-validated committees as described above are a good fit for our task, because of its unique unit of analysis. As thread-level analysis is the set of individual participants' turns in a conversation, we risk overfitting if we sample from the same conversations for the training and testing sets. In contrast to standard bagging, hard sampling boundaries never train and test on instances drawn from the same conversation.

To aggregate the votes from members of this ensemble into a final prediction, we employ a variant on Selfridge's Pandemonium (Selfridge, 1958). If a minority label is selected as the highest-confidence value in any classifier in our ensemble, it is selected. The majority label, by contrast, is only selected if it is the most likely prediction by all classifiers in our ensemble. Thus consensus is required to elect the majority class, and the strongest minority candidate is elected otherwise.

In : generation set $\mathbf{Tr}_k$, evaluation set $\mathbf{Te}_k$
Out: ensemble committee $\{E_1 \dots E_k\}$
for $i = 1$ **to** k **do**
  $R_{final} \leftarrow [\,]$;
  $\mathbf{X}_{freq} \leftarrow \{x \in \mathbf{X} \mid freq(x) \in \mathbf{Tr}_k > 5\%\}$;
  $R \leftarrow \mathbf{X}_{freq} \times T$;
  $R^* \leftarrow R$;
  repeat
    $P_{base} \leftarrow$ EvaluateClassifier($\mathbf{Tr}_k, \mathbf{Te}_k$);
    EvaluateRules($P_{base}, \mathbf{Tr}_k, \mathbf{Te}_k, R^*$);
    $\mathbf{Tr}_k, \mathbf{Te}_k \leftarrow$ ApplyRule($R^*[0]$);
    $R \leftarrow R - R^*[0]$;
    $\Delta \leftarrow score(R^*[0])$;
    $R_{final} \leftarrow R_{final} + R^*[0]$;
    $R^* \leftarrow R[0 \dots 50]$;
  until $\Delta <$ *threshold*;
  $\mathbf{Tr}_{final} \leftarrow \mathbf{Tr}_k \cup \mathbf{Te}_k$;
  foreach $r \in R_{final}$ **do**
    $\mathbf{Tr}_{final} \leftarrow$ ApplyRule($\mathbf{Tr}_{final}, r$);
  end
  Train $f(\bar{x} \rightarrow y)$ on $\mathbf{Tr}_{final}$;
end

Algorithm 1: LearnSelectionCues()

In : generation set $\mathbf{Tr}_k$, evaluation set $\mathbf{Te}_k$
Out: minority class probability average P_{base}
Train $f(\bar{x} \rightarrow y)$ on $\mathbf{Tr}_k$;
$\mathbf{Te}_k^{min} \leftarrow \{Instance\ I \in \mathbf{Te}_k \mid y_I \neq \text{“}NA\text{”}\}$;
$P_{base} \leftarrow 0$;
foreach $Instance\ I \in \mathbf{Te}_k^{min}$ **do**
  $P_{base} \leftarrow P_{base} + P(f(\bar{x}_I) = y_I)$
end
$P_{base} = P_{base}/size(\mathbf{Te}_k^{min})$

Algorithm 2: EvaluateClassifier()

In : $\mathbf{Tr}_k$, $\mathbf{Te}_k$, rules R, base probability P_{base}
Out: R sorted on each rule's improvement score
foreach *Rule* $r \in R$ **do**
  $\mathbf{Tr}'_k, \mathbf{Te}'_k \leftarrow$ ApplyRule($\mathbf{Tr}_k, \mathbf{Te}_k, r$);
  $P_{alter} \leftarrow$ EvaluateClassifier($\mathbf{Tr}'_k, \mathbf{Te}'_k$);
  $score(r) \leftarrow P_{alter} - P_{base}$;
end
Sort R on $score(r)$ from high to low;

Algorithm 3: EvaluateRules()

This approach is designed to bias the prediction of our machine learning algorithms in favor of minority classes in a coherent manner. If there is a plausible model that has been trained which recognizes the possibility of a rare label, it is used; the prediction only reverts to the majority class when no plausible minority label could be chosen. As validation of this technique, we compare our "minority pandemonium" approach against both typical pandemonium and standard sum-rule confidence voting (Erp et al., 2002).

5.1 Logistic Model Stumps

One characteristic of highly skewed data is that, while minority labels may be expressed in a number of different surface forms, there are many obvious cases in which they do not apply. These cases can actually be harmful to classification of borderline cases. Features that could be given high weight in marginal cases may be undervalued in "low-hanging fruit" easy cases. To remove those obvious instances, a very simple screening heuristic is often enough to eliminate frequent phenotypes of instances where the rare annotation is not present. Prior work has sometimes screened training data through obvious heuristic rules, especially in speech recognition; for instance, training speech recognition for words followed by a pause separately from words followed by another word (Franco et al., 2010), or training separate models based on gender (Jiang et al., 1999).

We achieve this instance screening by learning logistic model tree stumps (Landwehr et al., 2005), which allow us to quickly partition data if there is a particularly easy heuristic that can be learned to eliminate a large number of majority-class labels. One challenge of this approach is our underlying unigram feature space - tree-based algorithms are generally poor classifiers for the high-dimensionality, low-information features in a lexical feature space (Han et al., 2001). To compensate, we employ a smaller, denser set of binary features for tree stump screening: instance length thresholds and LIWC category membership.

First, we define a set of features that split based on the number of lines an instance contains, from 1 to 10 (only a tiny fraction of instances are more than 10 lines long). For example, a feature splitting on instances with lines ≤ 2 would be true for one- and two-line instances, and false for all others. Second, we define a feature for each category in the Linguistic Inquiry and Word Count dictionary (Tausczik and Pennebaker, 2010) - these broad classes of words allow for more balanced

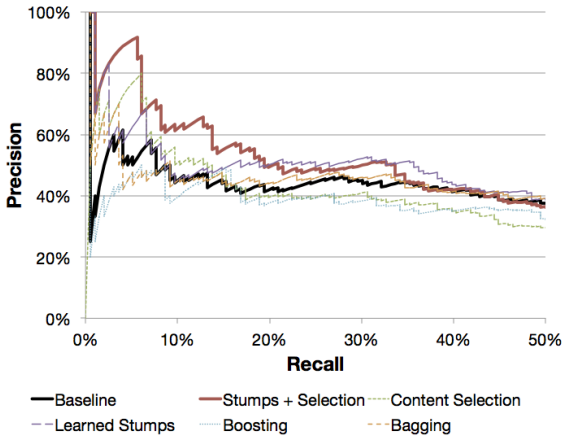

Figure 3: Precision/recall curves for algorithms. After 50% recall all models converge and there are no significant differences in performance.

splits than would unigrams alone. Each category's feature is true if any word in that category was used at least once in that instance.

We exhaustively sweep this feature space, and report the most successful stump rules for each annotation task. In our other experiments, we report results with and without the best rule for this preprocessing step; we also measure its impact alone.

6 Experimental Results

All experiments were performed using LightSIDE (Mayfield and Rosé, 2013). We use a binary unigram feature space, and we perform 7-fold cross-validation. Instances from the same chat transcript never occur in both train and testing folds. Furthermore, we assume that threads have been disentangled already, and our experiments use gold standard thread structure. While this is not a trivial assumption, prior work has shown thread disentanglement to be manageable (Mayfield et al., 2012a); we consider it an acceptable simplifying assumption for our experiments. We compare our methods against baselines including a majority baseline, a baseline logistic regression classifier with L2 regularized features, and two common ensemble methods, AdaBoost (Freund and Schapire, 1996) and bagging (Breiman, 1996) with logistic regression base classifiers[5].

Table 2 presents the best-performing result from each classification method. For self-empowerment recognition, all methods that we introduce are significant improvements in κ, the measurement of agreement over chance, compared to all baselines. While accuracy remains stable, this is due to predictions shifting away from the majority class and towards minority classes. Our combined model using both logistic model tree stumps and content selection is significantly better than either alone ($p < .01$). To compare the minority pandemonium voting method against baselines of simple pandemonium and summed confidence voting, Table 3 presents the results of content selection wrappers with each voting method. Minority voting is more effective compared to standard confidence voting, improving κ while modestly reducing accuracy; this is typical of a shift towards minority class predictions.

[5]These methods usually use weak, unstable base classifiers; however, in our experiments, those performed poorly.

Table 2: Performance for baselines, common ensemble algorithms, and proposed methods. Statistically significant improvements over baseline are marked ($p < .01$, †; $p < .05$, *; $p < 0.1$, +).

	Self		Other	
Method	%	κ	%	κ
Majority	79.3	.000	81.3	.000
LR Baseline	81.0	.367	81.0	.270
LR + Boosting	78.1	.325	78.5	.275
LR + Bagging	81.2	.352	**81.9**	.265
LR + Committee	81.0	.367	81.0	.270
Learned Stumps	**81.8***	.385†	81.7	**.293+**
Content Selection	80.9	.389†	80.7	.282
Stumps+Selection	81.3	**.406**†	79.4	.254

Table 3: Performance of content-selection wrapped learners, for minority voting and two baseline voting methods.

	Self		Other	
Method	%	κ	%	κ
Pandemonium	80.3	.283	81.4	.239
Averaged	80.6	.304	**81.6**	.251
Minority Voting	**80.9**†	**.389**†	80.7	**.282**

7 Discussion

These results show promise for our techniques, which are able to distinguish features of rare labels, previously awash in a sea of irrelevance. Figure 3 shows the impact of our rules as we tune to different levels of recall, with a large boost in precision when recall is not important; our model converges with the baseline for high-recall, low-precision tuning. This suggests that our method is particularly suitable for tasks where confident la-

Table 4: Cue rules commonly selected by the algorithm. Average improvement over the LR baseline is also shown.

Self-Empowerment

Cue	Transformation	Δ%
and,but	Ignore Local Future	+5.0
have	Ignore All History	+4.3
!	Ignore All History	+4.2
me,my	Ignore All History	+3.4

Other-Empowerment

Cue	Transformation	Δ%
and,but	Ignore Local Future	+5.5
you	Ignore Local History	+5.2
's	Ignore Local History	+4.1
that	Ignore Local History	+3.9

Table 5: Best decision rules for logistic model stumps. Significant improvement ($p < 0.05$) indicated with *.

Self-Empowerment

Split Rule		κ	$\Delta\kappa$	%	Δ%
Split ≤ 1	*	0.385	+.018	81.8	+0.8
LIWC-Article		0.379	+.012	81.6	+0.6
LIWC-Swear	*	0.376	+.009	81.4	+0.4
LIWC-Self	*	0.376	+.009	81.5	+0.5

Other-Empowerment

Split Rule		κ	$\Delta\kappa$	%	Δ%
LIWC-You		0.293	+.023	81.7	+0.7
LIWC-Eating	*	0.283	+.013	81.6	+0.6
LIWC-Negate	*	0.282	+.012	82.3	+1.3
LIWC-Present		0.281	+.011	81.6	+0.6

beling of a few instances is more important than labeling as many instances as possible. This is common when tasks have a high cost or carry high risk (for instance, providing real-time conversational supports with an agent, where inappropriate intervention could be disruptive). Other low-recall applications include exploration large corpora for exemplar instances, where the most confident predictions for a given label should be presented first for analyst use. In the rest of this section, we examine notable within-instance and per-instance rules selected by our methods. These rules are summarized in Tables 4 and 5.

For both self- and other-empowerment, we find pronoun rules that match the task (first-person and second-person pronouns for self-Empowerment and other-Empowerment respectively). In both tasks, we find cue rules that suppress the context preceding personal pronouns. These, as well as the possessive suffix *'s*, echo the per-instance effect of the *Self* and *You* splits, anticipating that what follows such a personal reference is likely to bear an evaluation of empowerment. Exclamation marks may indicate strong emotion - we find many instances where what precedes a line with an exclamation is more objective, and what follows includes an assessment. Conjunctions *but* and *and* are selected as cue rules suppressing the two lines that follow the occurrence - suggesting, as suspected, that connective discourse markers play a role in indicating empowerment (Fraser, 1999).

The best-performing stump splits for the Self-Empowerment annotation are *Line Length* ≤ 1 and the LIWC word-categories *Article*, *Swear*, and *Self*. The split on line length corresponds to the observation that longer instances provide greater opportunity for personal narrative self-assessment to occur (95% of single-line instances are labeled NA). The *Article* category may serve as a proxy for content length - article-less instances in our corpus include one-line social greetings and exchanges of contact information. *Swear* words may be a cue for awareness of self-empowerment - a recent study of women coping with illness reported that swearing in the presence of others, but not alone, was related to potentially harmful outcomes (Robbins et al., 2011). Among *other*- oriented split rules, *Eating* stands out as non-obvious, although medical literature has suggested a link between dietary behavior and empowerment attitudes in a study of women with cancer (Pinto et al., 2002).

8 Conclusion

We have demonstrated an algorithm for improving automated classification accuracy on highly skewed tasks for conversational data. This algorithm, particularly its focus on content selection, is rooted in the structural format of our data, which can generalize to many tasks involving conversational data. Our experiments show that this model significantly improves machine learning performance. Our algorithm is taking advantage of structural facets of discourse markers, lending basic sociolinguistic validity to its behavior. Though we have treated each of these rarely-occurring labels as independent thus far, in practice we know that this is not the case. Joint prediction of labels through structured modeling is an obvious next

step for improving classification accuracy.

This is an important step towards large-scale analysis of the impact of support groups on patients and caregivers. Our method can be used to confidently highlight occurrences of rare labels in large data sets. This has real-world implications for professional intervention in social conversational domains, especially in scenarios where such an intervention is likely to be associated with a high cost or high risk. With the construction of more accurate classifiers, we open the possibility of automating annotation on large conversational datasets, enabling new directions for researchers with domain expertise.

Acknowledgments

The research reported here was supported by National Science Foundation grant IIS-0968485.

References

Paige Adams and Craig Martel. 2010. Conversational thread extraction and topic detection in text-based chat. In *Semantic Computing*.

David Adamson and Carolyn Penstein Rosé. 2012. Coordinating multi-dimensional support in collaborative conversational agents. In *Proceedings of Intelligent Tutoring Systems*.

Albert Bandura. 1997. *Self-Efficacy: The Exercise of Control*.

Azy Barak, Meyran Boniel-Nissim, and John Suler. 2008. Fostering empowerment in online support groups. *Computers in Human Behavior*.

A Boehm and L H Staples. 2002. The functions of the social worker in empowering: The voices of consumers and professionals. *Social Work*.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of IJCAI*.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*.

Pierre Dillenbourg. 2002. Over-scripting cscl: The risks of blending collaborative learning with instructional design. *Three worlds of CSCL. Can we support CSCL?*

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*.

Merijn Van Erp, Louis Vuurpijl, and Lambert Schomaker. 2002. An overview and comparison of voting methods for pattern recognition. In *Frontiers in Handwriting Recognition*. IEEE.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification.

Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda. 2010. Eduspeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*.

Eibe Frank and Remco R Bouckaert. 2006. Naive bayes for text classification with unbalanced classes. *Knowledge Discovery in Databases*.

Bruce Fraser. 1999. What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Yoav Freund and Robert E Schapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of ICML*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*.

Martin Gütlein, Eibe Frank, Mark Hall, and Andreas Karwath. 2009. Large-scale attribute selection using wrappers. In *Proceedings of IEEE CIDM*.

Eui-Hong Han, George Karypis, and Vipin Kumar. 2001. Text categorization using weight adjusted k-nearest neighbor classification. *Lecture Notes in Computer Science: Advances in Knowledge Discovery and Data Mining*.

Ahmed Hassan, Anthony Fader, Michael H Crespin, Kevin M Quinn, Burt L Monroe, Michael Colaresi, and Dragomir R Radev. 2008. Tracking the dynamic evolution of participant salience in a discussion. In *Proceedings of Coling*.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*.

Mette Terp Hoybye, Christoffer Johansen, and Tine Tjornhoj-Thomsen. 2005. Online interaction effects of storytelling in an internet breast cancer support group. *Psycho-oncology*.

Hui Jiang, Keikichi Hirose, and Qiang Huo. 1999. Robust speech recognition based on a bayesian prediction approach. In *IEEE Transactions on Speech and Audio Processing*.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning*.

Elijah Mayfield and Carolyn Penstein Rosé. 2013. Lightside: Open source machine learning for text. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions*.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012a. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of SIGDIAL Meeting on Discourse and Dialogue*.

Elijah Mayfield, Miaomiao Wen, Mitch Golant, and Carolyn Penstein Rosé. 2012b. Discovering habits of effective online support group chatrooms. In *ACM Conference on Supporting Group Work*.

Kanayo Ogura, Takashi Kusumi, and Asako Miura. 2008. Analysis of community development using chat logs: A virtual support group of cancer patients. In *Proceedings of the IEEE Symposium on Universal Communication*.

Jason E. Owen, Erin O'Carroll Bantum, and Mitch Golant. 2008. Benefits and challenges experienced by professional facilitators of online support groups for cancer survivors. In *Psycho-Oncology*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics*.

Bambang Parmanto, Paul Munro, and Howard R Doyle. 1996. Improving committee diagnosis with resampling techniques. In *Proceedings of NIPS*.

James W Pennebaker and J D Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of Clinical Psychology*.

Bernardine M Pinto, Nancy C Maruyama, Matthew M Clark, Dean G Cruess, Elyse Park, and Mary Roberts. 2002. Motivation to modify lifestyle risk behaviors in women treated for breast cancer. In *Mayo Clinic Proceedings*.

Andrei Popescu-Belis. 2008. Dimensionality of dialogue act tagsets: An empirical analysis of large corpora. In *Language Resources and Evaluation*.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of NAACL*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*.

Megan L Robbins, Elizabeth S Focella, Shelley Kasle, Ana María López, Karen L Weihs, and Matthias R Mehl. 2011. Naturalistically observed swearing, emotional support, and depressive symptoms in women coping with illness. *Health Psychology*, 30:789.

Carolyn Penstein Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. In *International Journal of Computer Supported Collaborative Learning*.

Nikol Rummel, Armin Weinberger, Christof Wecker, Frank Fischer, Anne Meier, Eleni Voyiatzaki, George Kahrimanis, Hans Spada, Nikolaos Avouris, and Erin Walker. 2008. New challenges in cscl: Towards adaptive script support. In *Proceedings of ICLS*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of ACL*.

Emanuel A Schegloff. 1968. Sequencing in conversational openings. *American Anthropologist*.

Oliver G Selfridge. 1958. Pandemonium: a paradigm for learning. In *Proceedings of Symposium on Mechanisation of Thought Processes, National Physical Laboratory*.

Gerry Stahl. 2012. Interaction analysis of a biology chat. *Productive multivocality*.

Lee H Staples. 1990. Powerful ideas about empowerment. *Administration in Social Work*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*.

Simone Teufel and Marc Moens. 2002. Summarizing scientic articles: Experiments with relevance and rhetorical status. *Computational Linguistics*.

C F Van Uden-Kraan, C H C Drossaert, E Taal, E R Seydel, and M A F J Van de Laar. 2009. Participation in online patient support groups endorses patients empowerment. *Patient Education and Counseling*.

R Vauth, B Kleim, M Wirtz, and P W Corrigan. 2007. Self-efficacy and empowerment as outcomes of self-stigmatizing and coping in schizophrenia. *Psychiatry Research*.

Ingrid Wahlin, Anna-Christina Ek, and Ewa Idvali. 2006. Patient empowerment in intensive carean interview study. *Intensive and Critical Care Nursing*.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan Black, and Justine Cassell. 2012. "love ya, jerkface:" using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of SIGDIAL*.

Gary M Weiss and Haym Hirsh. 1998. Learning to predict rare events in event sequences. In *Proceedings of KDD*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*.