Open Domain Targeted Sentiment

Margaret Mitchell **Jacqueline Aguilar** **Theresa Wilson** **Benjamin Van Durme**
Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD 21218, USA
{m.mitchell,jacqui.aguilar}@jhu.edu, Theresa.Wilson@oberlin.edu, vandurme@cs.jhu.edu

Abstract

We propose a novel approach to sentiment analysis for a low resource setting. The intuition behind this work is that sentiment expressed towards an entity, *targeted sentiment*, may be viewed as a span of sentiment expressed across the entity. This representation allows us to model sentiment detection as a sequence tagging problem, jointly discovering people and organizations along with whether there is sentiment directed towards them. We compare performance in both Spanish and English on microblog data, using only a sentiment lexicon as an external resource. By leveraging linguistically-informed features within conditional random fields (CRFs) trained to minimize empirical risk, our best models in Spanish significantly outperform a strong baseline, and reach around 90% accuracy on the combined task of named entity recognition and sentiment prediction. Our models in English, trained on a much smaller dataset, are not yet statistically significant against their baselines.

1 Introduction

Sentiment analysis is a multi-faceted problem. Determining when a positive or negative sentiment is being expressed is a large part of the challenge, but identifying other attributes, such as the target of the sentiment, is also crucial if the ultimate goal is to pinpoint and extract opinions. Consider the examples below, all of which contain a positive sentiment:

(1) So happy that Kentucky lost to Tennessee!

(2) Kentucky versus Kansas I can hardly wait...

(3) Kentucky is the best alley-oop throwing team since Sherman Douglas' Syracuse squads!!

The entities in these examples are college basketball teams, and the events referred to are games. In (1), although there is a positive sentiment, the target of the sentiment is an event (Kentucky losing to Tennessee). However, from the positive sentiment toward this event, we can infer that the speaker has a negative sentiment toward Kentucky and a positive sentiment toward Tennessee. In (2), the positive sentiment is toward a future event, but we are not given enough information to infer a sentiment toward the mentioned entities. In (3), Kentucky is the direct target of the positive sentiment. We can also infer a positive sentiment toward Douglas's Syracuse teams, and even toward Douglas himself.

These examples illustrate the importance of the target when interpreting sentiment in context. If we are looking for sentiments toward Kentucky, for example, we would want to identify (1) as negative, (2) as neutral (no sentiment) and (3) as positive. However, if we are looking for sentiment toward Tennessee, we would want to identify (1) as positive, and (2) and (3) as neutral.

The expression of these and other kinds of sentiment can be understood as involving three items:

(1) An experiencer

(2) An attitude

(3) A target (optionally)

Research in sentiment analysis often focuses on (2), predicting overall sentiment polarity (Agarwal et al., 2011; Bora, 2012). Recent work has begun to combine (2) with (3), examining how to automatically predict the sentiment polarity expressed towards a target entity (Jiang et al., 2011; Chen et al., 2012) for a fixed set of targets. This topic-dependent sentiment classification requires that the target entity be

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1643–1654,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

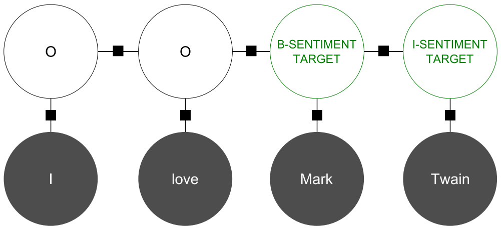


Figure 1: Sentiment expressed across an entity.

given, and returns statements expressing sentiment towards the given entity.

In this paper, we take a step towards open-domain, *targeted sentiment analysis* by investigating how to detect both the named entity and the sentiment expressed toward it. We observe that sentiment expressed towards a target entity may be possible to learn in a graphical model along the span of the entity itself: Similar to how named entity recognition (NER) learns labels along the span of each word in an entity name, sentiment may be expressed along the entity as well. A small example is shown in Figure 1. We focus on people and organizations (*volitional named entities*), which are the primary targets of sentiment in our microblog data (see Table 1).

Both NER and opinion expression extraction have achieved impressive results using conditional random fields (CRFs) (Lafferty et al., 2001) to define the conditional probability of entity categories (McCallum and Li, 2003; Choi et al., 2006; Yang and Cardie, 2013). We develop such models to jointly predict the NE and the sentiment expressed towards it using minimum risk training (Stoyanov and Eisner, 2012). We learn our models on informal Spanish and English language taken from the social network Twitter,[1] where the language variety makes NLP particularly challenging (see Figure 2).

Our ultimate goal is to develop models that will be useful for low resource languages, where a sentiment lexicon may be known or bootstrapped, but more sophisticated linguistic tools may not be readily available. We therefore do not rely on an external part-of-speech tagger or parser, which are often used for features in fine-grained sentiment analysis; such tools are not available in many languages, and if they are, are not usually adapted for noisy social media.

@[user] le dijo erralo muy por lo bajo jaja un grande juancito grandes amigos mios
@[user] he told him it was very on the dl haha a great juancito great friends of mine

---

@[user] buenos días Profe!! Nos quedamos accidentados otra vez en la carretera vieja guarenas echando gasoil, estamos a la interperie
@[user] good morning, Prof!! We were wrecked again on the old guarenas highway while getting diesel, we're out in the open

---

Sin ánimo de ofender a los Militares, que realmente se merecen ese aumento y más. Pero, dónde queda la misma recompensa para Médicos.
I do not intend to offend the military in the slightest, they truly deserve the raise and more. However, I'm wondering whether doctors will ever receive a similar compensation.

Figure 2: Messages on Twitter use a wide range of formality, style, and errors, which makes extracting information particularly difficult. Examples from Spanish (screen names anonymized), with approximate translations in English.

Instead, we use information from sentiment lexicons and some simple hand-written features, and otherwise use only features of the word that can be extracted without supervision. These include features based on unsupervised word tags (Brown clusters) and a method that automatically syllabifies a word based on the orthography of the language. All tools and code used for this research are released with this paper.[2]

2 Related Work

As the scale of social media has grown, using sources such as Twitter to mine public sentiment has become increasingly promising. Commercial systems include Sentiment140[3] (products and brands) and tweetfeel[4] (suggests searching for popular movies, celebrities and companies).

The majority of academic research has focused on supervised classification of message sentiment irrespective of target (Barbosa and Feng, 2010; Pak and Paroubek, 2010; Bifet and Frank, 2010; Davidov et al., 2010; Kouloumpis et al., 2011; Agarwal et al., 2011). Large datasets are collected for this work by leveraging the sentiment inherent in emoticons (e.g., smilies and frownies) and/or select Twitter hashtags (e.g., #bestdayever, #fail), resulting in noisy collec-

[1] www.twitter.com

[2] www.m-mitchell.com/code

[3] www.sentiment140.com

[4] www.tweetfeel.com

tions appropriate for initial exploration. Prior work includes: the use of a social network (Speriosu et al., 2011; Tan et al., 2011; Calais Guerra et al., 2011; Jiang et al., 2011; Li et al., 2012; Hu et al., 2013); user-adapted models based on collaborative online-learning (Li et al., 2010b); unsupervised, joint sentiment-topic modeling (Saif et al., 2012); tracking changing sentiment during debates (Diakopoulos and Shamma, 2010); and how orthographic conventions such as word-lengthening can be used to adapt a Twitter-specific sentiment lexicon (Brody and Diakopoulos, 2011).

Efforts in targeted sentiment (Bermingham and Smeaton, 2010; Jin and Ho, 2009; Li et al., 2010a; Jiang et al., 2011; Tan et al., 2011; Wang et al., 2011; Li et al., 2012; Chen et al., 2012), have mostly focused on topic-dependent analysis. In these approaches, messages are collected on a fixed set of topics/targets, such as products or sports teams, and sentiment is learned for the given set. In contrast, we aim to predict sentiment in tweets for any named person or organization. We refer to this task as *open domain targeted sentiment analysis*.

Within topic-dependent sentiment analysis, several approaches have explored applying CRFs or HMMs to extract sentiment and target words from text (Jin and Ho, 2009; Li et al., 2010a). In these approaches, opinion expressions are extracted, and polarity is annotated across the opinion expression. However, as noted by many researchers in sentiment, opinion orientation towards a specific target is often not equal to the orientation of a neighboring opinion expression; and opinion expressions in one context may not be opinion expressions in another (Kim and Hovy, 2006), making open domain approaches particularly challenging.

The above work by Jiang et al. (2011) is most similar to our own. They do not use joint learning, but they do incorporate a number of parse-based features designed to capture relationships between sentiment terms and topic references. In our work these relationships are captured by the CRF model, and we compare against their approach in Section 6.

Recent work by Yang and Cardie (2013) is similar in spirit to our own, where the identification of opinion holders, opinion targets, and opinion expressions is modeled as a sequence tagging problem using a CRF. However, similar to previous work applying CRFs to extract sentiment, Yang and Cardie use syntactic relations to connect an opinion target to an opinion expression. In contrast, we model the expression of sentiment polarity across the sentiment target itself, extracting both the sentiment target and the sentiment expressed towards it within the same span of words. This allows us to use surrounding context to determine sentiment polarity without identifying explicit opinion expressions or relying on a parser to help link expression to target.

Most work in targeted sentiment outside the microblogging domain has been in relation to product review mining (e.g., Yi et al. (2003), Hu and Liu (2004), Popescu and Etzioni (2005), Qiu et al. (2011)). Rather than identify named entities (NEs), this work seeks to identify products and their features mentioned in reviews, and classify these for sentiment. Recent work by Qui et al. jointly learns targets and opinion words, and Jakob and Gurevych (2010) use CRFs to extract the targets of opinions, but do not attempt to classify the sentiment toward these targets. To the best of our knowledge, this is the first work to approach targeted sentiment in a low resource setting and to jointly predict NEs and targeted sentiment.

3 Data

Twitter Collection We use the Spanish/English Twitter dataset of Etter et al. (2013) to train and test our models. Approximately 30,000 Spanish tweets and 10,000 English were labeled for named entities in BIO encoding: The start of an NE is labeled B-{NE} and the rest of the NE is labeled I-{NE}. The

NE	COUNT	NEUTRAL	POS	NEG
PERSON	5462	80%	20%	0%
ORGANIZATION	4408	80%	20%	0%
LOCATION	1405	100%	0%	0%
URL	1030	100%	0%	0%
TIME	535	70%	10%	20%
DATE	222	100%	0%	0%
MONEY	95	90%	0%	10%
PERCENT	81	80%	20%	0%
TELEPHONE	23	100%	0%	0%
EMAIL	8	100%	0%	0%

Table 1: Distribution of named entities in our Spanish Twitter corpus. Targeted sentiment percentages are based on expert annotations from a random sample of 10 (or all) of of each entity. Most entities are not sentiment targets (NEUTRAL). PERSON and ORGANIZATION are most frequent, and among the top recipients of sentiment.

full set of NE categories are shown in Table 1. For example, the sequence "Mark Twain" would be labeled B-PERSON, I-PERSON. We are interested in both PERSON and ORGANIZATION entities, which make up the majority of named entities in this data, and we evaluate these using the more general entity category VOLITIONAL. Removing retweets, 7,105 Spanish tweets contained a total of 9,870 volitional entities and 2,350 English tweets contained a total of 3,577 volitional entities.

Sentiment Lexicons We use two sentiment lexicon sources in each language. For English, we use the MPQA lexicon (Wilson et al., 2005), which identifies 12,296 manually and semi-automatically produced subjective terms along with their polarity. For the second lexicon, we use SentiWordNet 3.0 (Baccianella et al., 2010), which assigns positive and negative polarity scores to WordNet synsets. We use the majority polarity of all words with a subjectivity score above 0.5.

For Spanish, the first lexicon is obtained from Volkova et al. (2013), who automatically translated strongly subjective terms from the MPQA lexicon (Wilson et al., 2005) into Spanish. The resulting Spanish lexicon contains about 65K words. The second lexicon is available from Perez-Rosas et al. (2012). This contains approximately 1000 sentiment-bearing words collected leveraging manual resources and 2000 collected leveraging automatic resources.

Annotation To collect sentiment labels, we use crowdsourcing through Amazon's Mechanical Turk.[5] Annotators ("Turkers") were shown six tweets at a time, each with a single highlighted named entity. Turkers were instructed to (1) select the sentiment being expressed towards the entity (*positive*, *negative*, or *no sentiment*); and (2) rate their level of confidence in their selection. Following best practices on collecting language data with Mechanical Turk (Callison-Burch and Dredze, 2010), two controls were placed among each set of six tweets to screen out unreliable judgments. An example prompt is shown in Figure 3.

Each ⟨tweet, NE⟩ pair was shown to three Turkers, and those with majority consensus on sentiment polarity were extracted. Tweets without sentiment consensus on all NEs were removed. In Spanish, this yielded 6,658 unique ⟨tweet, NE⟩ pairs. In English, which is a smaller data set, this yielded 3,288 unique pairs. We split the data into folds for 10-fold cross-validation, developing on the data from one fold and reporting results for the remaining nine.

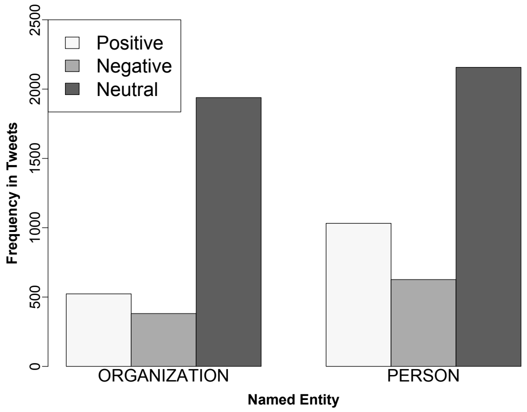


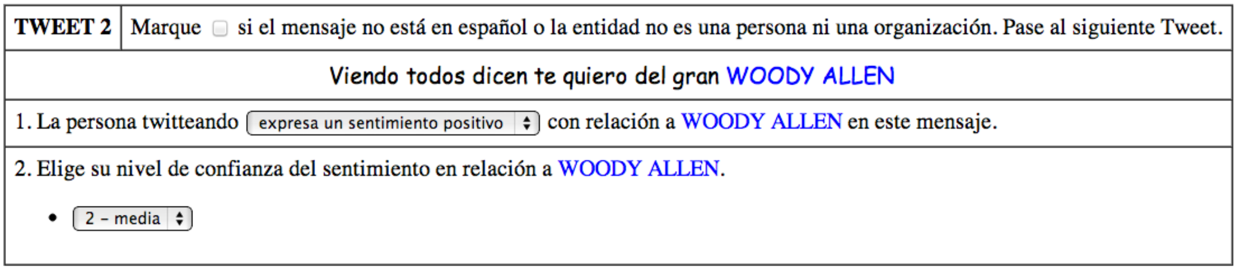
Figure 4: Targeted sentiment annotated for Spanish.

		Majority		
		POS	NEUTRAL	NEG
Minority	POS	757	1249	130
	NEUTRAL	707	2151	473
	NEG	129	726	452

Table 2: Number of targeted sentiment instances where at least two of the three annotators (Majority) agreed. Common disagreements with a third annotator (Minority) were over whether no sentiment or positive sentiment was expressed, and whether no sentiment or negative sentiment was expressed.

The distribution of sentiment for the named entities annotated by Turkers is shown in Figure 4. Neutral (no targeted sentiment) dominates, followed by positive sentiment for both organizations and people. As shown in Table 2, common disagreements were over whether or not there was targeted positive sentiment, and whether or not there was targeted negative sentiment. This is in line with previous research showing that distinguishing positive sentiment from no sentiment (and distinguishing negative sentiment from no sentiment) is often more challenging than distinguishing between positive and negative sentiment (Wilson et al., 2009). Indeed, we see that it was more common for annotators to disagree than to agree on targeted sentiment, particularly for negative targeted sentiment, where more instances had NEUTRAL/NEGATIVE disagreement than NEGATIVE three-way agreement.

[5] www.mturk.com/mturk



TWEET 2 Marque □ si el mensaje no está en español o la entidad no es una persona ni una organización. Pase al siguiente Tweet.

Viendo todos dicen te quiero del gran WOODY ALLEN

1. La persona twitteando expresa un sentimiento positivo con relación a WOODY ALLEN en este mensaje.

2. Elige su nivel de confianza del sentimiento en relación a WOODY ALLEN.

- 2 - media

Figure 3: Example Tweet shown to Turkers.

Variable	Possible values
Sentiment (s) (PIPE & JOINT models)	NOT-TARG, SENT-TARG
Named Entity (l) (PIPE & JOINT models)	O, B-VOLITIONAL, I-VOLITIONAL
Combined Sent/NE (y) (COLL models)	O, B+NOT-TARG, I+NOT-TARG B+SENT-TARG, I+SENT-TARG

Table 3: Possible values for random variables, targeted subjectivity (is/is not sentiment target). COLL models collapse targeted subjectivity and NE label into one node.

Variable	Possible values
Sentiment (s) (PIPE & JOINT models)	NOT-TARG, POS, NEG
Named Entity (l) (PIPE & JOINT models)	O, B-VOLITIONAL, I-VOLITIONAL
Combined Sent/NE (y) (COLL models)	O, B+NOT-TARG, I+NOT-TARG B+POS, I+POS B+NEG, I+NEG

Table 4: Possible values for random variables, targeted sentiment. The COLL models collapse both targeted sentiment and NE label into one node.

4 Targeted Subjectivity and Sentiment

Formally, we define the problem as follows: Given an observed message $\mathbf{w} = (w_1 \dots w_n)$, where n is the number of words in the message and $w_j (1 \leq j \leq n)$ is a word, we learn the probability of a label sequence $\mathbf{l} = (l_1 \dots l_n)$, where $l_i \in$ the set of named entity values; and a sentiment sequence $\mathbf{s} = (s_1 \dots s_n)$, where $s_i \in$ the set of sentiment values. We additionally explore simpler linear-chain models that learn the probability of a single label sequence $\mathbf{y} = (y_1 \dots y_n)$, where $y_i \in$ the set of conjoined entity+sentiment values (Tables 3 and 4).

Our basic model is a linear conditional random field, an undirected graph that represents the conditional distribution $p(\mathbf{l}, \mathbf{s}|\mathbf{w})$.[6] Sentiment towards a named entity may be modeled in a CRF as a sequence of random variables for sentiment $\mathbf{s}$ connected to named entities $\mathbf{l}$. In all models, entity variables are connected by a factor to their neighbors in sequence, and we include skip-chains (Finkel and Manning, 2010) connecting identical words where at least one is capitalized. Our model strategies include: a pipeline that first learns volitional entities then sentiment directed towards them (PIPE); one that jointly learns volitional entities along with sentiment directed towards them (JOINT); and one that learns volitional entities and targeted sentiment with combined labels (COLL) (Figure 5).

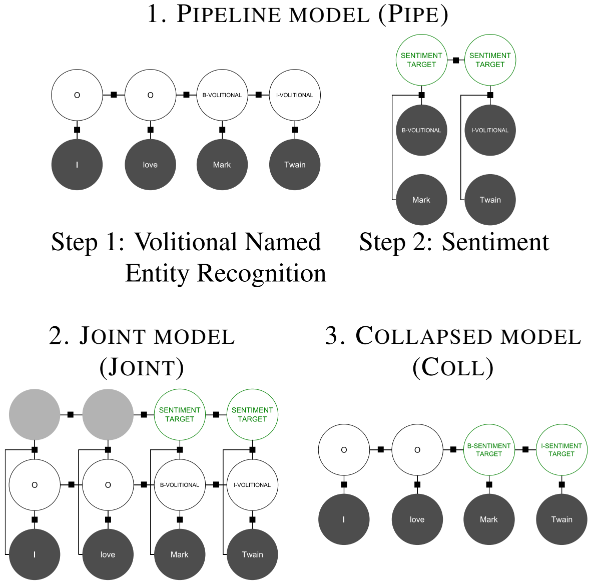
Using these models, we explore two primary tasks: (1) the task of detecting whether sentiment is targeted at an entity, which we refer to as *targeted subjectivity*; and (2) the task of detecting whether positive, negative, or neutral sentiment (no sentiment) is targeted at an entity, which we refer to as *targeted sentiment*. Moving from targeted subjectivity prediction to targeted sentiment prediction is possible by changing the sentiment target (SENT-TARG) variable into two variables, one for positive targeted sentiment (POS) and one for negative (NEG). Possible values for targeted subjectivity are shown in Table 3, and possible values for targeted sentiment are shown in Table 4.

In the pipeline models (PIPE), we first build a CRF where each word is connected by a factor to an entity label $l_i \in \mathbf{l}$. In a second model, every observed volitional entity node is connected by a factor to a sentiment label $s_i \in \mathbf{s}$. An example is shown in Figure 5 (1).

In the joint models (JOINT), each $s_i \in \mathbf{s}$ is connected by a factor to the corresponding entity label in the sequence, $l_i \in \mathbf{l}$. Sentiment in this model is partially observed: All sentiment variables are treated as latent except for the sentiment connected to the volitional entity. An example is shown in Figure 5 (2).

[6] For the COLL models, this is instead the conditional distribution $p(\mathbf{y}|\mathbf{w})$, where entity and sentiment labels are conjoined in one sequence assignment $\mathbf{y}$.

In the collapsed models (COLL), we combine sentiment and named entity into one label sequence (e.g., O, B+SENT-TARG, I+SENT-TARG). An example is shown in Figure 5 (3). The JOINT and PIPE models therefore predict named entity sequences, their category labels, and the sentiment expressed towards volitional named entities.[7] The collapsed models predict volitional labels and targeted sentiment as combined categories. The COLL and PIPE models are considerably faster than JOINT models, where exact inference is intractable.


1. PIPELINE MODEL (PIPE)

Step 1: Volitional Named Entity Recognition

Step 2: Sentiment

2. JOINT MODEL (JOINT)

3. COLLAPSED MODEL (COLL)

Figure 5: Example CRFs for targeted subjectivity with observed variables (dark nodes), predicted variables (white nodes) and hidden variables (light grey nodes).

5 Training

Minimum-Risk CRF Training We use the ERMA system (Stoyanov et al., 2011) to learn our models.[8] ERMA (Empirical Risk Minimization under Approximations) learns parameters to minimize loss on the training data. Predicting NE labels using a linear-chain CRF trained with empirical risk minimization has been shown to result in a statistically significant improvement over the common approach of maximum likelihood estimation (Stoyanov and Eisner, 2012). All models are trained to optimize log likelihood using 20 iterations of stochastic gradient descent, and a maximum of 100 iterations of belief propagation to compute the marginals for each example.

Features Features of the models are shown in Table 5. For an observed word, features are extracted for the word itself as well as within a context window of three words in either direction. Words seen only once are treated as out-of-vocabulary. Surface features and linguistic features are concatenated in groups of two and three to create further features. All algorithms and code that we have developed for feature extraction are available online.[9]

Because we aim to develop models that do not heavily rely on language-specific resources, we are interested in exploring unsupervised and lightly supervised methods for learning relevant features. Rather than use part-of-speech tags, we therefore use Brown cluster labels as unsupervised word tags (Brown et al., 1992; Koo et al., 2008). Brown clustering is a distributional similarity method that merges pairs of word clusters in the training data[10] to create the smallest decrease in corpus likelihood, using a bigram language model on the clusters. For our task, we cut clusters at length 3 and length 5, and these serve as rough part-of-speech tags without the need to train additional models. For example, the word *hello* is tagged as belonging to cluster 011 (length 3) and 01111 (length 5).

During development, we found that being able to syllabify the word (break the word into syllables) was a positive indicator of people names, but a negative indicator of organization names. This observation can be approximated automatically using constraints from the sonority sequencing principle (Hooper, 1976; Clements, 1990; Blevins, 1996; Morelli, 2003) on a language's orthography. This is a phonotactic principle that states that syllables will tend to have a sonority peak, usually a vowel, in the center of the syllable, followed on either side by consonants with decreasing sonority. Although languages may violate this principle, the core idea that a vowel forms the nucleus of a syllable with op-

[7] We found that learning the VOLITIONAL categories during training rather than maintaining beliefs about separate named entities during inference (ORGANIZATION, PERSON) and then post-processing to VOLITIONAL leads to slightly better accuracy.

[8] sites.google.com/site/ermasoftware

[9] www.m-mitchell.com/code

[10] For Spanish, we train on a sample of ~7 million Spanish tweets. For English, we train on the essays (Pennebaker et al., 2007) and Facebook data (Kosinskia et al., 2013) available from ICWSM 2013.

tional consonants before (the onset) and after (the coda) can be used to begin to automatically learn syllable structure.[11] We learn this in an unsupervised way, using the most frequent (seen more than 1,000 times) word-initial non-vowel sequences from the Brown cluster data as allowable syllable onset consonants. Similarly, the most frequent word-final non-vowel sequences are learned as possible syllable codas. For each word, we then attempt to segment syllables using the learned onsets and codas around each vowel. If a word cannot be syllabified, it is often an initialism (e.g., *CND*, *lsat*).

We follow the approach from the out-of-vocabulary assignment in the Berkeley parser (Petrov et al., 2006) to encode common surface patterns such as capitalization and lexical patterns such as verb endings as a single feature for words we have seen once or less. We also use the Jerboa toolkit (Van Durme, 2012) to extract further language-independent features from the data, such as features for emoticons and binning for repeated characters (like !!!). In addition, we include features for whether the word is three or four letters, which is often used for acronyms and initialisms in several languages (including Spanish and English); whether the word is neighbored by a punctuation mark; word identity; word length; message length; and position in the sentence.

We utilize a speaker of each language to simply list word forms for sentiment features that may be indicative of sentiment, totaling less than two hours of annotation time. This set includes intensifiers (e.g., *hella*, *freakin'* in English; e.g., *muy*, *sumamente* in Spanish), positive/negative abbreviations (*WTF*, *pso*), positive/negative slang words, and positive/negative prefix and suffixes (e.g., *anti-* in English and Spanish, *-ito* in Spanish).

6 Experiments

We are interested in both PERSON and ORGANIZATION entities, and evaluate these in the collapsed category VOLITIONAL. This suggests that the data may be pre-processed to label all volitional entities as VOLITIONAL NEs, or the models may be learned with the traditional named entities in place, and post-processed to identify those that are VOLITIONAL. We explored results using both methods, and found that training models on VOLITIONAL tags yielded the best performance overall; we report numbers for this approach below.

SURFACE FEATURES
binned word length, message length, and sentence position; Jerboa features; word identity; word lengthening; punctuation characters, has digit; has dash; is lower case; is 3 or 4 letters; first letter capitalized; more than one letter capitalized, etc.
LINGUISTIC FEATURES
function words; can syllabify; curse words; laugh words; words for *good*, *bad*, *no*, *my*; slang words; abbreviations; intensifiers; subjective suffixes and prefixes (such as diminutive forms); common verb endings; common noun endings
BROWN CLUSTERING FEATURES
cluster at length 3; cluster at length 5
SENTIMENT FEATURES
is sentiment-bearing word; prior sentiment polarity

Table 5: Features used in model.

We compare against a baseline (BASE-NS) where we use our volitional entity labels and assign no sentiment directed towards the entity (the majority case). This is a strong baseline to isolate how our methods perform specifically for the task of identifying sentiment targeted at an entity.

We report on precision, recall, and sensitivity for the tasks of NER and targeted subjectivity/sentiment prediction in isolation; and we report on accuracy for the targeted subjectivity and targeted sentiment models. For sentiment, a true positive is an instance where the label has sentiment, and a true negative is an instance where the label has no sentiment (neutral). For NER, a true positive is an instance where the label is a B- or I- label; a true negative is an instance where the label is O. The three systems are evaluated against one another for NER, subjectivity (entity has/does not have sentiment expressed towards it), and sentiment (positive/negative/no sentiment) using paired t-tests across folds, with a Bonferroni correction to set α to 0.02.

NER We include results for the isolated task of volitional named entity recognition in Table 6. In both Spanish and English, all three models are roughly comparable for precision, recall, and specificity. The task of finding O tags – spans that are *not* named entities – works especially well (**NE spec**). Common

[11] Further development is necessary to extend a similar idea to languages that do not ordinarily mark all vowels in their orthography, such as Hebrew and Arabic.

	Spanish			English		
Model	**Joint**	**Pipe**	**Coll**	**Joint**	**Pipe**	**Coll**
NE prec	**65.2**	64.3	65.1	59.8	**62.3**	60.5
NE rec	**65.8**	64.7	61.2	**60.2**	57.2	56.5
NE spec	95.4	95.2	**95.6**	94.3	**95.1**	94.7

Table 6: Average precision, recall, and specificity for volitional entity NER (in %).

	Spanish			English		
Model	**Joint**	**Pipe**	**Coll**	**Joint**	**Pipe**	**Coll**
Subj prec	58.3	58.8	**58.9**	46.6	**52.2**	45.9
Subj rec	40.1	**50.9**	19.1	44.5	**48.5**	16.4
Subj spec	**79.6**	77.5	77.8	77.6	**80.8**	74.0

Table 7: Average precision, recall, and specificity (in %) for subjectivity prediction (has/does not have sentiment) along the target entity.

	Spanish			English		
Model	**Joint**	**Pipe**	**Coll**	**Joint**	**Pipe**	**Coll**
Sent prec	36.6	**45.8**	42.5	31.6	**42.9**	38.5
Sent rec	38.0	**40.6**	15.5	**36.6**	34.8	9.7
Sent spec	67.1	**75.2**	73.3	72.3	**82.0**	78.1

Table 8: Average precision, recall, and specificity (in %) for sentiment prediction (positive/negative/no sentiment) along the target entity.

mistakes include confusing B- labels with I- labels.

Subjectivity and Sentiment Table 7 shows results for the isolated task of predicting the presence of sentiment about a volitional entity. In Spanish, the pipeline models (PIPE) perform optimally for subjectivity recall (**Subj rec**), and significantly above the COLL models (p<.001). Precision and specificity are comparable across models. In English as in Spanish, the collapsed model is particularly poor at subjectivity recall.

As discussed in Section 2, the subtask of predicting whether subjectivity is expressed towards an entity is comparable to the main task of Jiang et al. (2011), and so we compare our approach here. The Jiang et al. study is similar to the current study in that they aim to detect targeted sentiment, but it differs from the current study in that they focus exclusively on subjectivity towards five manually selected entities: {*Obama*, *Google*, *iPad*, *Lakers*, *Lady Gaga*}. They also evaluate on artificially balanced evaluation data, and evaluate sentiment polarity (positive/negative) separately from subjectivity (has/does not have sentiment).

Our dataset includes any entity labeled as PERSON or ORGANIZATION, and is not balanced (most targets have no sentiment expressed towards them; see Table 1), thus we can only roughly compare against their approach. *Lakers* and *Lady Gaga* are rare in our collection (appearing less than 3 times), and so we updated the comparison set prior to evaluation to: {*Obama*, *Google*, *iPad*, *BBC*, *Tebow*}. On this set, a baseline that always guesses no sentiment reaches an accuracy of 66.9%, compared to Jiang et al.'s 65.5% accuracy on a balanced set (not strictly comparable, but provided for reference). The JOINT models reach an accuracy of 71.04% on this set, demonstrating this approach as potentially useful for topic-dependent targeted sentiment.

Table 8 shows results for the task of predicting the polarity of the sentiment expressed about an entity. In Spanish, the PIPE models significantly outperform the COLL models on sentiment recall, and the JOINT models on sentiment precision (p<.01). In English, PIPE significantly outperforms JOINT on precision (p<.001).

Targeted Subjectivity and Targeted Sentiment The JOINT and PIPE models work reasonably well for the isolated tasks of NER and subjectivity/sentiment prediction. We now examine results for *targeted subjectivity* – labeling an entity and predicting whether there is sentiment directed towards it – in Table 9; and *targeted sentiment* – labeling an entity and predicting what the sentiment directed towards it is – in Table 10.

We evaluate using two accuracy metrics: **Acc-all**, which measures the accuracy of the entire named entity span along with the sentiment span; and **Acc-Bsent**, which measures the accuracy of identifying the start of a named entity (B- labels) along with the sentiment expressed towards it. Acc-all primarily measures the correctness of O labels, while Acc-Bsent focuses on the beginning of named entities.

For the targeted subjectivity task, our JOINT models perform optimally in Spanish, and significantly above their baselines. For the Acc-Bsent task, JOINT models perform best, significantly outperforming their baseline for subjectivity prediction. In English, where our data is half the size, we do not see a statistically significant difference between the predictive models and the no sentiment baselines.

For the targeted sentiment task, the JOINT models again perform relatively well in Spanish (Table 10), labeling volitional entities, predicting whether or not there is sentiment targeted towards them, and

	Model	Joint	Joint Base	Pipe	Pipe Base	Coll	Coll Base
Spa	Acc-all	89.5*	89.3	89.3**	89.1	89.5*	89.3
	Acc-Bsent	32.1***	29.5	30.9***	28.3	30.1**	28.1
Eng	Acc-all	88.0	88.1	88.6	88.6	87.9	88.1
	Acc-Bsent	30.4	30.8	30.7	30.3	28.1	29.2

***p<.001 **p<.01 *p<.05

Table 9: Average accuracy on Targeted Subjectivity Prediction: Identifying volitional entities and whether they are a sentiment target. In the core task, **Acc-Bsent**, the best model in Spanish is Joint, significantly outperforming the baseline. In English, the best model (Pipe) does not significantly improve over its baseline.

	Model	Joint	Joint Base	Pipe	Pipe Base	Coll	Coll Base
Spa	Acc-all	89.4	89.4	89.0	89.0	89.2	89.3
	Acc-Bsent	29.7*	29.0	30.0	29.2	28.9	29.0
Eng	Acc-all	88.0	88.1	88.2	88.4	87.7	88.1
	Acc-Bsent	30.4	30.6	30.5	30.8	27.9	29.8

*p<.05

Table 10: Average accuracy on Targeted Sentiment Prediction: Identifying volitional entities and the polarity of the sentiment expressed towards them. The Spanish Joint models significantly improve over their baseline for the core task. In English, no models outperform their baseline.

the sentiment polarity above their no sentiment baselines. We find this to be the most difficult task: It may be clear that sentiment is being expressed towards an entity, but it is not always clear what the polarity of that sentiment is. Error analysis is given below in this section. In the smaller English set, the models do not outperform the no sentiment baseline.

7 Discussion

Feature Analysis Examples of some of the top-weighted features in the Spanish models are shown in Table 11. In addition to lexical identity and Brown cluster, we find that positive indicators include positive suffixes such as diminutive forms, whether the word can be syllabized (Section 5), and whether it is three or four letters.

Error Analysis Because it is relatively common for there *not* to be sentiment targeted at a named entity, it is difficult to tease out the polarity in instances where there is targeted sentiment. Similarly, our predictions are most reliable for detecting the absence of a named entity (O labels).

B-VOLITIONAL FEATURES	
Negative	is a function word; jerboa tags; followed by a word with 3 or 4 letters that cannot be syllabified
Positive	ends in -a, -o, or -s; is capitalized; has one non-initial capital letter; is 3 or 4 letters
B-VOLITIONAL, POS FEATURES	
Negative	preceded by a curse word; followed by a word with a positive suffix; immediately preceded by a word with a negative prefix
Positive	not in a sentiment lexicon; preceded by a happy emoticon; followed by an exclamation or a 'my' word; immediately preceded by a laugh; has two or more sentiment-bearing words in the sentence
B-VOLITIONAL, NEG FEATURES	
Negative	is immediately followed by a question mark or positive abbreviation word
Positive	preceded by a 'bad' word or curse word; has four or more sentiment lexicon items
B-VOLITIONAL, NOT-TARG FEATURES	
Negative	immediately followed by a 'no' word or word with a negative prefix; is preceded by a question mark; is immediately preceded by a curse word or laugh; is followed by an exclamation mark
Positive	not followed by sentiment lexicon word

Table 11: Example strongly weighted features for a Spanish joint sentiment model. In addition to lexical identity, we find that curse words and positive and negative prefixes are used to detect volitional entities and the sentiment directed towards them.

Label confusions are shown in Table 12. Mistakes are often made by confusing B- labels (the start of an entity) with I- labels (inside an entity); and by predicting sentiment polarity when the gold annotations say there is not sentiment targeted at the entity. Some example errors are shown in Figure 13. In (1), "CANSADO" (*"TIRED"*) was predicted to be volitional, while "Matthew" was not. In (2), "Matias del río" was not predicted to be an entity, likely due to the fact that the capitalization patterns we see in this sentence are indicative of the start of a sentence rather than a proper name (similar to 1). In (3),

a.

		Observed		
		B	I	O
Predicted	B	423	21	186
	I	36	236	135
	O	197	90	7168

b.

	Observed		
	POS	NEG	NEUT
POS	68	24	42
NEG	58	65	102
NEUT	115	61	468

Table 12: Predicted vs. observed values for a joint model. (**a**) For named entities, most common confusions were between B-VOLITIONAL and O labels. (**b**) For sentiment, most common mistakes were to predict that a positive sentiment was neutral (no sentiment), and that a neutral sentiment was negative.

NE prediction errors

1.	**Spanish:**	Cuando	estoy	CANSADO	,	él	es	mi	DESCANSO	.	Mateo	.	11	:	29	.
	Predicted:	O	O	B-VOLITIONAL	O	O	O	O	O	O	O	O	O	O	O	O
	Gold:	O	O	O	O	O	O	O	O	O	B-VOLITIONAL	O	O	O	O	O
	English:	*When*	*I'm*	*TIRED*	*,*	*he*	*is*	*my*	*REST*	*.*	*Matthew*	*.*	*11*	*:*	*29*	*.*

2.	**Spanish:**	Matias	del	río	fue	una	lata	. . .
	Predicted:	O	O	O	O	O	O	. . .
	Gold:	B-VOLITIONAL	I-VOLITIONAL	I-VOLITIONAL	O	O	O	. . .
	English:	*Matias*	*del*	*río*	*was*	*a*	*drag*	*. . .*

Sentiment prediction errors

3.	**Spanish:**	Mario	que	dio	este	contigo
	Predicted:	NOT-TARG	-	-	-	-
	Gold:	POSITIVE	-	-	-	-
	English:	*Mario*	*may*	*God*	*be*	*with you*

4.	**Spanish:**	. . .	si	de	verdad	estas	en	cielo	,	ayudame	Superman	!!!
	Predicted:		-	-	-	-	-	-	-	-	POSITIVE	-
	Gold:		-	-	-	-	-	-	-	-	NOT-TARG	-
	English:	*. . .*	*if*	*you*	*really*	*are*	*in*	*the skies*	*,*	*help me*	*Superman*	*!!!*

Sentiment and NE prediction errors

5.	**Spanish:**	Salen	del	gobierno	de	Humala	dos	connotados	izquierdistas	,	Giesecke	y	Eiguiguren
	Predicted:	O -	O -	O -	O -	B-VOLITIONAL NOT-TARG	I-VOLITIONAL NOT-TARG	O -	O -	O -	B-VOLITIONAL NOT-TARG	O -	B-VOLITIONAL NOT-TARG
	Gold:	O -	O -	O -	O -	B-VOLITIONAL NOT-TARG	O -	O -	O -	O -	B-VOLITIONAL NEGATIVE	O -	B-VOLITIONAL NOT-TARG
	English:	*Leaving*	*the*	*Humala*		*government*	*are two*	*notorious*	*leftists*	*,*	*Giesecke*	*and*	*Eiguiguren*

Table 13: Example errors made by joint models.

sentiment may not be clear without spelling correction: "dio" should be "dios", meaning "God"; otherwise, "dio" is the word for "gave". Humans can easily fix the spelling error, which changes the overall reading of the expression. In (4), the positive polarity item "verdad" (*"believe"*) and the exclamation marks (!!!) were likely used as indicators of positive sentiment; however, in this case the annotators marked the targeted sentiment as neutral. In (5), the "Humala" entity was predicted to be longer than it is ("Hamala dos" or *"Hamala two"*). It was also predicted that both "Giesecke" and "Eiguiguren" had no sentiment expressed towards them; annotators disagreed, with the majority of those who annotated "Giesecke" marking negative sentiment, and the majority of those who annotated "Eiguiguren" marking no sentiment. This highlights some of the difficulty in predicting sentiment discussed in Section 3, where annotators will often disagree as to whether there is no sentiment or positive/negative sentiment.

During development, we found that the collapsed model (COLL) performed best on small amounts of data. However, as we scaled up the amount of data we trained on, the PIPE and JOINT models significantly improved, while the COLL models did not have significant performance gains.

8 Conclusion

We have introduced the task of open domain targeted sentiment: predicting sentiment directed towards an entity along with discovering the entity itself. Our approach is developed to find targeted sentiment towards both person and organization named entities by modeling sentiment as a span along the entity.

We find that by modeling targeted sentiment in this way, we can reliably detect entities and whether or not they are sentiment targets above a no sentiment baseline. How best to determine the *polarity* of the sentiment expressed towards the entity, however, is still an open issue. Our data suggests that it is usually not clear-cut whether sentiment is being expressed or not; the strong disagreement between annotators suggests that detecting sentiment polarity in microblogs is difficult even for humans.

In future work, we hope to explore further methods for teasing apart sentiment polarity expressed towards a target. This research has achieved promising results for detecting sentiment targets without relying on external supervised models, and we hope that the features and approaches developed here can aid in sentiment analysis in noisy text and languages without rich linguistic resources.

References

A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of Coling: Posters*.

Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of CIKM-2010*.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in Twitter streaming data. In *Proceedings of the International Conference on Discovery Science (DS-2010)*.

Juliette Blevins. 1996. The syllable in phonological theory. In John A. Goldswmith, editor, *The Handbook of Phonological Theory*. Blackwell Publishing, Blackwell Reference Online.

N. N. Bora. 2012. Summarizing public opinions in tweets. In *Proceedings of CICLing-2012*.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of EMNLP-2011*.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J.C. Lai, and R.L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the KDD-2011*.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL:HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of ICWSM-2012*.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. *Proceedings of EMNLP 2006*.

G. N. Clements. 1990. The role of the sonority cycle in core syllabification. In J. Kingston and M. Beckman, editors, *Papers in Laboratory Phonology*, pages 283–333. CUP, Cambridge.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of Coling: Posters*.

Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of CHI-2010*.

David Etter, Francis Ferraro, Ryan Cotterell, Olivia Buzek, and Benjamin Van Durme. 2013. Nerit: Named entity recognition for informal text. Technical Report 11, Human Language Technology Center of Excellence, Johns Hopkins University, July.

Jenny Rose Finkel and Christopher D. Manning. 2010. Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of ACL-2010*.

Joan B. Hooper. 1976. The syllable in phonological theory. *Language*, 48(3):525–540.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*.

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM-2013)*.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of EMNLP*.

Long Jiang, Mo Yu, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of ACL-2011*.

Wei Jin and Hung Hay Ho. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. *Proceedings of ICML 2009*.

Soo-Min Kim and Eduard Hovy. 2006. Identifying and analyzing judgment opinions. *Proceedings of NAACL 2006*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL/HLT*.

Michal Kosinskia, David Stillwell, and Thore Graepel. 2013. Private trains and attributes are predictable from digital records of human behavior. *Proc. of the National Academy of Sciences of the USA*, 110(5).

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of ICWSM-2011*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-2001*.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010a. Structure-aware review mining and summarization. *Proceedings of Coling 2010*.

Guangxia Li, Steven CH Hoi, Kuiyu Chang, and Ramesh Jain. 2010b. Micro-blogging sentiment detection by collaborative online learning. In *Proceedings of ICDM-2010*.

Hao Li, Yu Chen, Heng Ji, Smaranda Muresan, and Dequan Zheng. 2012. Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC-2012)*.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction, and web-enhanced lexicons. In *Proceedings of CoNLL-2003*.

Frida Morelli. 2003. The relative harmony of /s+stop/ onsets: Obstruent clusters and the sonority sequencing principle. In C. Fery and R. van de Vijver, editors, *The syllable in optimality theory*, pages 356–371. CUP, New York.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC-2010*.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic inquiry and word count: Liwc2007, operator's manual.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. *Proceedings of the Conference on Language Resources and Evaluations (LREC 2012)*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of Coling:ACL-2006*.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT:EMNLP-2005*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1).

Hassan Saif, Yulan He, and Harith Alani. 2012. Alleviating data sparsity for twitter sentiment analysis. *Proceedings of the WWW Workshop on Making Sense of Microposts (# MSM2012)*.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the EMNLP-2011 Workshop on Unsupervised Learning in NLP*.

Veselin Stoyanov and Jason Eisner. 2012. Minimum-risk training of approximate crf-based nlp systems. In *Proceedings of NAACL:HLT-2012*.

Veselin Stoyanov, Alexander Ropson, and Jason Eisner. 2011. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *AIStats*.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the KDD-2011*.

Benjamin Van Durme. 2012. Jerboa: A toolkit for randomized and streaming algorithms. Technical report, Human Language Technology Center of Excellence, Johns Hopkins University.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Association for Computational Linguistics (ACL)*.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of CIKM-2011*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3).

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. *Proceedings of ACL 2013*.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of ICDM-2003*.