Sentiment Identification by Incorporating Syntax, Semantics and Context Information

Kunpeng Zhang, Yusheng Xie, Yu Cheng,
Daniel Honbo, Doug Downey, Ankit Agrawal, Wei-keng Liao, Alok Choudhary
Department of Electric Engineering and Computer Science,
Northwestern University, Evanston, IL 60208, USA

NORTHWESTERN UNIVERSITY

McCormick Northwestern Engineering

Introduction

Problem: Understanding the sentiment of sentences allows us to summarize opinions which could help people make informed decisions. All of the state-of-the-art algorithms perform well on individual sentences without considering any context in- formation, but their accuracy is dramatically lower on the document level because they fail to consider context and the syntactic structure of sentences at the same time.

Challenges: There are many difficulties owing to the special characteristics and diversity in sentence structure in the way people express their opinions, including mixed sentiments in one sentence, sarcastic sentences, and opinions expressed indirectly through comparison, etc. In addition, complicated sentence structure and Internet slang make sentiment analysis even more challenging.

Goal: In this work, we not only consider syntax that may influence the sentiment, including newly emerged Internet language, emoticons, positive words, negative words, and negation words, but also incorporate information about sentence structure, like conjunction words and comparisons. The context around a sentence also plays an important role in determining the sentiment. Therefore, we employ a conditional random field (CRF) [2] model to capture syntactic, structural, and contextual features of sentences.

Results: Our experiment results on customer reviews and Facebook comments show better accuracy compared to supervised and rule-based methods. Furthermore, we also employ active learning to help collect more labeled data. We propose two different strategies to select data with high uncertainty for human beings to label, and our experimental results on customer reviews show faster convergence compared to baselines.

Problem Definition

Different subjectivity can generate different or even reversed sentiments for sentences. Therefore, the input is a set of m documents: { d_1, d_2, . . . , d_m } along with the specified subject: {sub_1, sub_2, . . . , sub_m}. Each d_i contains n_i sentences S^i : {s^i_1, s^i_2, . . . , s^i_{ni} }. The output for all documents is that for the j^{th} sentence in the i^{th} document s^i_j , it will assign a sentiment $o^i_j \in$ { *P* : positive, *N* : negative, *O* : objective }.

Conditional Random Fields (CRF)

CRF provides a probabilistic framework for calculating the probability of label sequences Y globally conditioned on sequence data X to be labeled. Parameters $\Theta = \{\lambda_k , \mu_l\}$ are estimated by maximizing the conditional log-likelihood function $L(\Theta)$ of the training data.

$$\mathcal{P}(Y \mid X) = \frac{1}{Z_X} exp(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l} \mu_l g_l(y_i, X))$$

$$\mathcal{L}(\Theta) = \sum_{j=1 \dots M} log(\mathcal{P}(Y^{(j)} \mid X^{(j)}; \Theta)) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2} - \sum_l \frac{\mu_l^2}{2\sigma_l^2}$$

Data Collection

Table 2 shows the data collected from Amazon Mechanical Turk. For each of these reviews, we asked 10 different workers from AMT to label the sentences as positive, negative, or objective. We used majority vote to determine the final label for each sentence. We also randomly selected 500 sentences from each of the camera and TV reviews and checked the labeling accuracy. The average response accuracy for all workers for the camera and TV reviews was 0.66 and 0.62 respectively. We also manually labeled 500 Facebook comments. We did some preprocessing tasks on the original data, including word correction (e.g., changing "*luv*" to "*love*") and part-of-speech (POS) tagging.

Table 2: Data distribution. nrc|ns|nps|nns|nos: # of reviews/comments | sentences | positive sentences | negative sentences | objective sentences

Data	nrc	ns	nps	nns	nos
Camera	300	5156	2524	1185	1447
TV	300	5036	2364	1252	1420
Facebook	500	723	313	157	253

Table 1: Features used for this sequence labeling problem.

Semantic Features	
n_pos_words	Number of positive words (a positive word list: 1948 words)
n_neg_words	Number of negative words (a negative word list: 4550 words)
if_pos_emo	Existence of positive emoticons (a positive emoticon list: 52 emoticons)
if_neg_emo	Existence of negative emoticons (a negative emoticon list: 35 emoticons)
if_comp_sent	A sentence is comparative if it contains comparative parts-of-speech (JJR, JJS,RBR, RBS), or comparative phrases ("compare to", "in contrast", etc.)
type_conjunction_words	Type of conjunction words: subordinating, coordinating, and correlative
Syntactic Features	
sent_post	Sentence position. If the sentence is within first 20% of the sentences, it's a beginning sentence; an end sentence if within the last 20%, and middle for all others
post_pos_words	Position of positive words occurring. 0: no positive words occur; 1: only exist in the first part of a sentence; 2: only exist in the second part; -1: exist in both parts (mixed).
post_neg_words	Position of negative words occurring. Same as above.
post_negation_words	Position of negation words. Same as above.
comp_sub	Comparison subject: If the subjectivity is the same as the input subjectivity.
cos_sim_neigh_sent	cosine similarity score to neighboring sentences (previous sentence and next sentence).
LSI_sim_neigh_sent	LSI similarity score to neighboring sentences (previous sentence and next sentence).

Experimental Results

We compare our proposed method against the following rule-based algorithms and supervised methods: compositional semantic rules (CSR) [1], support vector machine (SVM), logistic regression (LR), and hidden Markov models (HMM). Table 3 shows that CRFs outperform the other four methods in all cases on the Amazon review dataset. Using our CRF-based method with semantic and syntactic features is 5-15% more accurate than the other methods tested. However, CSR performs the best on the Facebook comments dataset, while all other methods generated similar results. We believe that this result is due to the length of the Facebook comments, which provide little to no context for our CRF-based method, as well as the use of emoticons, which convey sentiments directly.

Table 3: Accuracy results of CRF model comparing to other methods (CSR, SVM, LR, and HMM) with semantic features only (SO) and with semantic and syntactic features (SS).

Data+Feature	CSR	SVM	LR	HMM	CRF
Camera (SO)	0.57	0.633	0.615	0.631	0.654
Camera (SS)	0.57	0.640	0.648	0.651	**0.72**
TV (SO)	0.54	0.612	0.60	0.629	0.630
TV (SS)	0.54	0.622	0.619	0.633	**0.665**
Overall (SO)	0.55	0.622	0.610	0.627	0.634
Overall (SS)	0.55	0.632	0.637	0.640	**0.693**
Facebook (SO)	**0.72**	0.60	0.610	0.607	0.612
Facebook (SS)	**0.72**	0.60	0.612	0.61	0.614

Active Learning

Since collecting labeled data is expensive, we use active learning to collect the most valuable labeled examples. The fundamental step of active learning procedure is to choose what data to present to the oracle. When we apply our trained model on inferring unlabeled data, we get a sequence of label probabilities for a document which has *m* sentences : {p_1, p_2, . . . , p_m}. Each pi is the probability for the most probable label. In Strategy 1 (S1), we rank documents based on the average probability: $\frac{1}{m}\sum_{i=1}^{m} p_i$ and select the document with the smallest value to present to oracle. In Strategy 2 (S2), we rank sentences based on the probability in an ascending order and calculate the average of the probabilities in the smaller half P. We then rank the document based on P and present the document with the smallest P to oracle. We start from a training size of 10 documents and add one document at a time. We compare these strategies against two baselines, (B1) selecting a document at random and (B2) selecting a document based on the minimum probability of its sentences. In this paper, we use customer reviews to test the convergence speed. Figure 1 shows that S2 achieves the same accuracy faster than S1. Because documents with the smallest average probability may have some sentences with high probability, which do not need to be disambiguated.

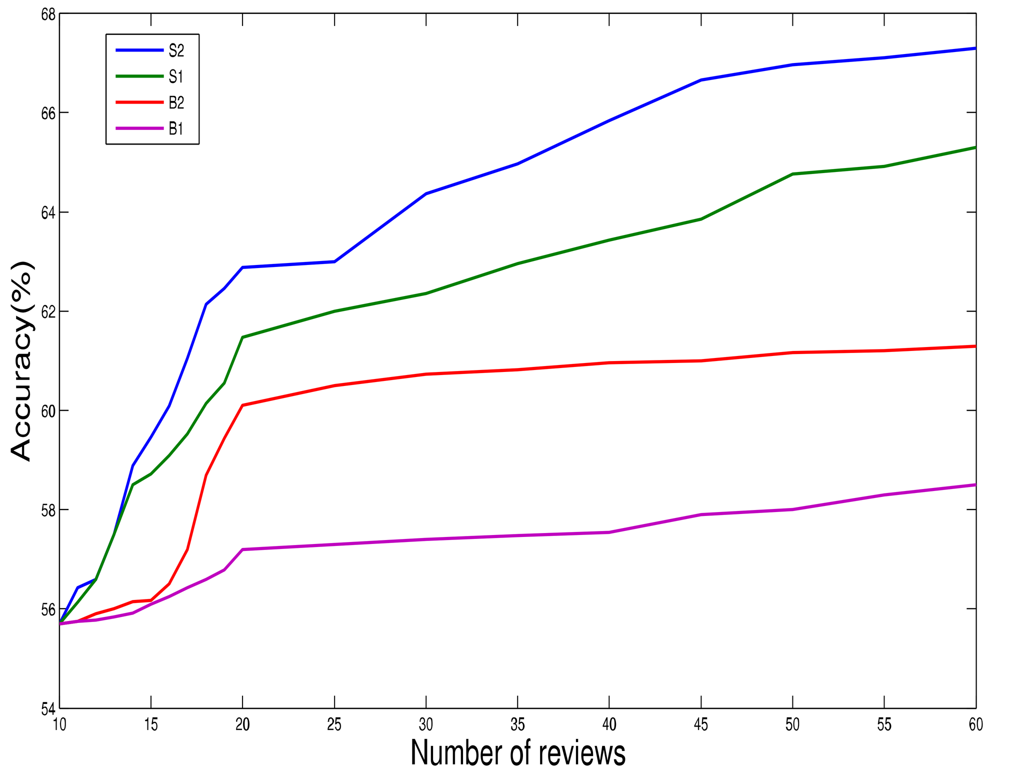


Figure 1: The convergence speed of classification accuracy (10-fold cross validation).

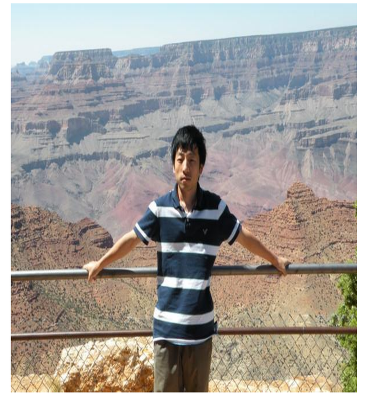
Acknowledgements

This work is supported in part by NSF award numbers: CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456.

References

➢Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. EMNLP '08, pages 793–801, 2008.
➢J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01, pages 282–289, 2001.

Contact



kzh980@eecs.northwestern.edu choudhar@eecs.northwestern.edu