Do Topic-Dependent Models Improve Microblog Sentiment Estimation?

Francine Chen
FX Palo Alto Laboratory
3174 Porter Dr
Palo Alto, CA 94304
chen@fxpal.com

Seyed Hamid Mirisaee
Université Joseph Fourier
Grenoble, France
hamid.mirisaee@imag.fr

Abstract

A topic-independent sentiment model is commonly used to estimate sentiment in microblogs. But for movie and product reviews, domain adaptation has been shown to improve sentiment estimation performance. We examined whether topic-dependent models improve polarity estimation of microblogs. We considered both a model trained on Twitter tweets containing a target keyword and a model trained on an enlarged set of tweets containing terms related to a topic. Comparing the performance of the topic-dependent models to a topic-independent model trained on a general sample of tweets, we noted that for some topics, topic-dependent models performed better. We then propose a method for predicting which topics are likely to have better sentiment estimation performance when a topic-dependent sentiment model is used.

Introduction

Estimating the sentiment of short social media texts, such as microblogs posted on Twitter and Tencent Weibo, is of interest to businesses that wish to better understand customer sentiment about their products and services. It is also of interest to other groups, e.g., political parties, that want to learn about individual and overall sentiment on topics.

It has been shown that for multi-paragraph length documents, a sentiment estimation model that is domain and context-aware improves performance, e.g., (Lu et al. 2011). The improved performance is thought to be due to term sentiment often being dependent on topic. For example, a *predictable* train arrival time is positive, but a *predictable* movie plot is usually negative.

A joint sentiment-topic model (JST) developed by (Lin et al. 2012) showed improved performance on movie and product reviews. However, the set of topics selected for positive and negative sentiments are usually different, which may be non-intuitive and undesirable. In addition, the model is based on Latent Dirichlet Allocation (LDA), which performs poorly on short documents such as microblogs, where the representation of each document is sparse (Phan, Nguyen, and Horiguchi 2008; Hong and Davison 2010).

Our focus is on sentiment estimation of microblogs, and tweets in particular, which are limited to at most 140 characters in length. These texts are very different than the longer, review-type documents traditionally analyzed using topic-dependent sentiment models. In addition to containing few words, the text is often non-grammatical and contains many contractions, acronyms, spelling variants and errors, so that the coverage by a sentiment dictionary is poorer for tweets than for regular reviews.

A single general sentiment model is commonly used with microblogs, e.g., (Kouloumpis, Wilson, and Moore 2011). This may be due in part to the brevity of each tweet and the corresponding practice of defining microblog topics by single terms, often hash-tagged.

(Mejova and Srinivasan 2012) examined the performance of sentiment models trained and tested on different combinations of media streams (i.e., blogs, reviews, and Twitter). Although they created topic-*independent* and topic-*dependent* models, they did not directly compare their performance. (Liu et al. 2013) presented adaptive co-training to improve sentiment classification of tweets. Their data set was small: six topics with less than 11,000 tweets total. They noted that the difference in performance over a baseline decreased as the amount of training data increased; when 40% of the data was used for training, the average increase in accuracy was only 0.92%. In this paper, we examine whether topic-dependent models improve sentiment estimation over a general (topic-independent) model for microblogs trained on millions of tweets. If the answer is 'no', then creating one general model is simpler than creating multiple topic-dependent models.

To create a topic-dependent polarity model, we trained on tweets containing a topic term of interest (a *target topic*). We also examined another approach of identifying related topic terms and then adding tweets containing the related topic terms to the training set. Although LDA is a popular model for identifying topics and topic terms in review text, tweets are not well-modeled directly; methods to handle this include aggregation of the tweets, such as by user (Hong and Davison 2010), or using external "knowledge" (Phan, Nguyen, and Horiguchi 2008). In addition, many tweets are chatter, and the identified topics often reflect chatter terms. Labeled LDA (Ramage, Dumais, and Liebling 2010) models tweets "as a mixture of some labeled dimensions as well as the traditional latent ones", and chatter can be modeled by including some common chatter terms as topic labels. We propose an alternative method for more directly and efficiently identifying candidate terms relevant to the target topic. In contrast to labeled LDA, the topic terms identified by our method are independent of the number and identity of other topics to be modeled.

Sentiment estimation typically involves first determining whether a sentence is subjective, i.e., opinionated. Then the polarity of the subjective sentences is estimated. Methods have been developed for subjectivity analysis, including a method specifically for Twitter (Luo, Osborne, and Wang 2012). Here, we focus on polarity estimation, and assume that subjectivity analysis has already been performed.

We evaluated the performance of topic-dependent polarity estimation models on microblogs and observed that for *some* topics, performance improved significantly over a general model. We then propose a method for predicting topics likely to have better polarity estimation performance when a topic-dependent model is used. The application context for this approach is a user following a set of topics in a microblog app such as Twitter, a common usage scenario. For example, a company may be interested in all tweets where its and its competitors' products are mentioned.

Our contributions in this paper are: 1) investigating whether the observation about domain-specific models improving sentiment estimation of reviews also applies to Twitter tweets 2) exploring whether extending Twitter topics to include related terms improves polarity estimation performance and 3) proposing a method to identify topics for which a topic-dependent model can improve microblog polarity estimation performance.

Data Set

For our experiments, we followed the approach used by (Go, Bhayani, and Huang 2009) and collected tweets with positive or negative emoticons using the Twitter streaming API with the keywords ":)" and ":(", which Twitter expands to include other positive sentiment emoticons such as ":D" and negative sentiment emoticons such as ":-(". The emoticons are stripped from each tweet and used as labels for training and testing. Instead of using full tweets as in (Go, Bhayani, and Huang 2009), we use only the text that precedes an emoticon, since an emoticon normally comments on the preceding text. The tweets were collected from Mar 15 to Oct 12, 2013. They were filtered to remove spam (using a whitelist of 263 clients), non-English tweets, retweets, tweets with both positive and negative sentiment, and duplicates within 10,000 tweets. Each of the remaining tweets was then preprocessed to remove stop words and single letter words, normalize URLs to '<<URL>>', remove '@names', and normalize terms with the Porter stemmer. We limited the maximum number of tweets per day to 100k by uniform sampling. After preprocessing, there were a total of 4,325,646 tweet sentences.

For evaluation, a set of 10 topics, as shown in Table 1, were selected as target topics for our polarity estimation tasks. The topics were motivated by popular Twitter 2012 trends (https://2012.twitter.com/en/trends.html[1]) or were popular terms in the data set.

topic	# sentences
bed	22210
car	10075
coffee	5370
google	2216
home	45246
mom	21628
movie	16952
phone	23310
sleep	41358
tv	6222

Table 1: Data Set Target Topics

Polarity Models

We compared tweet polarity prediction models trained on three types of data: 1) all tweets containing a target topic term (*topic*) 2) all tweets containing a target topic term or closely related topic terms (*extended*) 3) a sample of all tweets (*general*). We next describe our method for creating extended topic models.

Extended Topic Models

To more closely follow a traditional topic modeling approach, we extend a topic term with closely associated terms. Specifically, a set of candidate related terms is first identified; then greedy selection used to identify terms that most improve performance when tweets that contain those terms are added to the training set.

Identifying Candidate Terms: For a given topic word, candidate terms are identified that are both related and have enough exemplars to influence a polarity model. Relatedness is measured using Pointwise Mutual Information (PMI), where the PMI between terms x and y is computed as:

$$PMI(x,y) = \frac{p(x,y)}{p(x)p(y)} = \frac{p(x|y)}{p(x)}.$$

Less relevant terms are filtered by requiring that $PMI(x,y)$, or the probability of the target term x in the context of candidate term y, is greater than the probability of the target term x alone:

$$PMI(x,y) > 1.$$

To remove from consideration terms that are too infrequent to influence polarity model performance, we require the term frequency, $f()$, of term y to be greater than 10% of the term frequency of the most frequently co-occurring word, w:

$$f(y) > .1 * f(w)$$

where $f(z)$ is the number of tweets that contains both z and the topic term.

The best N terms are identified by first sorting the term frequency and PMI values in descending order. Then we iteratively consider the top k (initially k is set to 1) elements of the frequency list and PMI list and find the intersection. We continue the process by increasing k until the intersection contains at least N words or the end of either list is reached. For our experiments, we set N to 10.

Table 2 shows the top related terms selected by our method (PMI-Freq) and Labeled LDA with chatter labels from our pre-processed data set for four topics in Table 1.

[1] Retrieved Mar 29, 2013

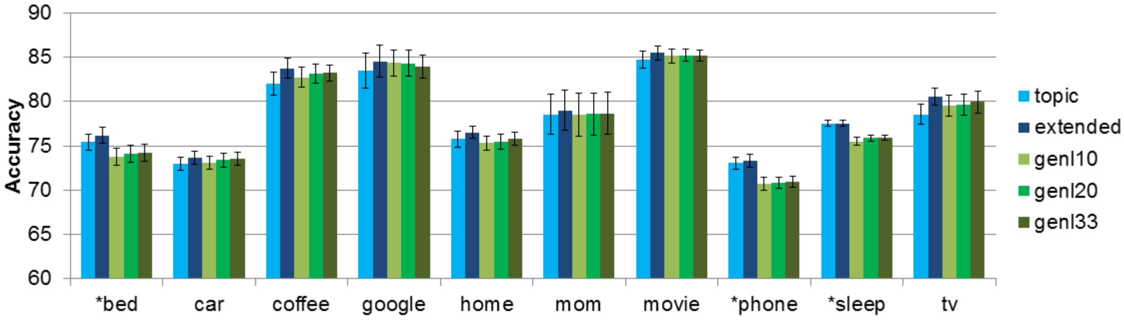

Figure 1: Classification accuracy for *topic*, *extended* topic, and *general* polarity estimation models. * indicates topics with a significant performance difference. Error bars indicate 95% confidence intervals.

topic	PMI-Freq	Labeled LDA
car	wash, driving, accident, drive, ride	care, dont, card, scared, carry
movie	watching, watch, favorite, scary, wanna	watch, watching, night, cant, time
phone	charger, die, dead, broke, charge	iphone, call, text, cant, dont
tv	kris, broken, hehehe, watching, watch	watch, kris, tvd, watching, mtv

Table 2: Top related terms using PMI-Freq and Labeled LDA

Our informal observation is that PMI-Freq identifies related topic terms at least as well as Labeled LDA. While Labeled LDA models topics in the corpus, PMI-Freq is computed only for the target topics of interest and is independent of other labels, such as chatter labels. In the next step, we use the terms identified by PMI-Freq as candidate topic terms.

Topic Term Selection: Candidate terms are selected for addition to the main topic term using a greedy method that improves the polarity classifier accuracy. To illustrate, imagine that the list of related words (obtained by the process explained above) of the *car* topic is: [wash, driving, accident, drive, ride, road, fixed, park, hit, clean]. We first build a classifier for each 2-tuple {car, wash}, {car, driving}, {car, accident}, etc., training on all tweets containing at least one of the tuple terms. We then select the 2-tuple which results in the highest accuracy value using cross-validation. Next, we try all triples containing the best 2-tuple and rebuild the classifiers for each triple. We iteratively add a new topic term until accuracy does not improve.

Polarity Model Comparison

We compared three types of polarity models on the task of classifying subjective sentences as positive or negative. For all models, we used binary term vectors as features and an SVM classifier, SVMlight (Joachims 1999), with a linear kernel, which has shown good performance on classifying tweets (Go, Bhayani, and Huang 2009). For each of our 10 target topics, a model was trained on the tweets containing the topic term (*topic*) and a model was trained on tweets containing the selected related topic terms (*extended*). For use as a baseline, a general polarity classifier was trained on uniformly sampled subsets (10%, 20%, 33%) of the preprocessed tweets (*genl10, genl20, genl33*), sampled to reduce the training time and memory requirements. For each target topic, all models were evaluated on the subset of tweets containing the target topic. We used 10-fold cross-validation (the folds were defined on the full data set) and computed mean classification accuracy and confidence intervals. Results are shown in Figure 1.

We also performed paired t-tests at the 0.05 significance level to assess performance differences between the models. We noted that any performance improvement of the general model as the amount of training increased was slight and non-significant and so used *genl33* as a baseline. Three topic-dependent models, 'bed', 'phone' and 'sleep, had significantly, and noticably, better performance than *genl33*. *Extended* is useful if it performs significantly better than both *topic* and *genl33*; it did so for 'bed', 'home', and 'movie'. However, the improvement for 'home' and 'movie' is somewhat small.

Identifying Useful Topic-Dependent Models

From our observation that a topic-dependent model improves polarity estimation for some topics, we propose a simple method for identifying these topics. It is based on identifying frequent terms for which the normalized distribution of positive and negative polarity differs significantly between: a) tweets containing the test term and b) tweets containing both the test term and the target topic. In addition, the expected polarity of a term in the context of the target topic should differ from that in general usage. The method is outlined in Algorithm 1. Topic terms in *result* are candidates for training a topic-dependent model. A refinement is to filter topics for which the 95% confidence interval, CI, from cross-validation of the polarity classifier is so large that the difference in performance is unlikely to be significant (e.g., 'mom' in Figure 1). We next examine these ideas using our general corpus of preprocessed tweets with emoticon labels.

For the 10 topic terms in Figure 1, we computed coverage, C, that is, the proportion of tweets containing at least one term where polarity changes. The five topic terms with the largest values of C are shown in Table 3, along with CI and sample terms identified as having polarity that differs from a general model. Note that CI is relatively small for all five terms and that the three terms for which a topic-dependent model was significantly better in Figure 1, 'phone', 'sleep'

Algorithm 1 Identification of useful topic models

```
Input: g: a set of polarity-labeled general tweets, W: a set of topic
    terms, M: # of pos/neg terms to use per topic (default=100)
Output: result: topics and scores indicating utility of a topic-
    dependent model
 1: result $\leftarrow$ {}
 2: for each topic t in W do
 3:     s $\leftarrow$ extract sentences containing topic t
 4:     n $\leftarrow$ select the M most frequent positive and M most fre-
        quent negative terms in s
 5:     T $\leftarrow$ {}
 6:     for each term e in n do
 7:         Compute the chi-square statistic comparing the nor-
            malized distribution of polarities over s and g containing e
 8:         if chi-square is significant then
 9:             T $\leftarrow$ T $\cup$ term
10:         end if
11:     end for
12:     Identify terms in T where the polarity differs for topic and
        general tweets
13:     if C, the proportion of tweets containing at least one term
        where the polarity changes, $>$ thresh then
14:         result $\leftarrow$ result $\cup$ (t,C)
15:     end if
16: end for
```

topic	C(%)	CI	terms that change polarity
phone	17.4	0.6977	home, charge, mom, message, week, answer, life
sleep	15.6	0.3437	hour, a.m., home, woke, trying, sorry
bed	14.1	0.9348	hour, comfy, warm, ill, phone
car	11.4	0.7242	hour, money, hit, feel
movie	8.3	0.9462	scary, cuddle, damn, cried

Table 3: Topic-term polarity coverage (C), conf. interval (CI), and sample terms that change polarity, ordered by C.

and 'bed', also have the largest value of C.

The method can also be used to create a topic-dependent, tweet sentiment dictionary by identifying topics and terms for which the polarity differs from the expected polarity (line 12 of Algorithm 1). For example, in Table 3, the term 'charge' is generally positive, but in the context of 'phone' is often negative. Similarly, 'a.m.' changes from generally positive to negative in the context of 'sleep', and 'hit' changes from generally positive to negative in the context of 'car'. Since tweets contain many spelling variants and acronyms, a tweet sentiment dictionary should provide better coverage than sentiment dictionaries constructed without microblogs, e.g., (Dragut et al. 2010; Lu et al. 2011).

Conclusions and Future Directions

We examined whether topic-dependent models improve polarity estimation of microblogs such as Twitter tweets. We observed that for *some* topics, topic-dependent models have significantly better performance than a general model. Our experiments also indicate that an extended-term model often performs better than a single-term model, although the improvement may be minor. Motivated by the idea of training a topic-dependent model only when useful, we presented a method for ranking topics by whether a topic-dependent model is likely to have better performance. We propose that based on C and CI, topics can be roughly grouped into those for which a topic-dependent model is useful (large C and small CI), not useful (small C or large CI), or uncertain and testing is needed. In the future, training and testing additional topic-dependent models would provide more labeled data for creating a classifier that uses C and CI as features.

We followed the sentiment estimation approach described in (Go, Bhayani, and Huang 2009) for our investigations; however, other approaches, such as sentiment estimation of a target term or phrase in a tweet, could also be used. For targeted sentiment models, only the terms related to the target term would be considered in step 4 of Algorithm 1.

The method in Algorithm 1 also identifies terms and topics for which the polarity can differ from the expected polarity. In the future, we would like to explore identifying these terms for more topics and using them to create a context-sensitive polarity dictionary for tweets.

References

Dragut, E. C.; Yu, C.; Sistla, P.; and Meng, W. 2010. Construction of a sentimental word dictionary. In *Proceedings of CIKM*, 1761–1764. ACM.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1–12.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, 80–88. ACM.

Joachims, T. 1999. *Advances in Kernel Methods – Support Vector Learning*. Universität Dortmund. chapter 11 Making large-scale SVM learning practical.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of ICWSM*, 538–541. AAAI.

Lin, C.; He, Y.; Everson, R.; and Ruger, S. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering* 24(6):1134–1145.

Liu, S.; Li, F.; Li, F.; Cheng, X.; and Shen, H. 2013. Adaptive co-training SVM for sentiment classification on tweets. In *Proceedings of CIKM*, 2079–2088. ACM.

Lu, Y.; Castellanos, M.; Dayal, U.; and Zhai, C. 2011. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of WWW*, 347–356. ACM.

Luo, Z.; Osborne, M.; and Wang, T. 2012. Opinion retrieval in Twitter. In *Proceedings of ICWSM*, 507–510. AAAI.

Mejova, Y., and Srinivasan, P. 2012. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and Twitter. In *Proceedings of ICWSM*, 234–241. AAAI.

Phan, X.-H.; Nguyen, L.-M.; and Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proc. of WWW*, 91–100. ACM.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proceedings of ICWSM*, 130–137. AAAI.