A Distant Supervision Approach for Identifying Perspectives in Unstructured User-Generated Text

Attapol Thamrongrattanarit[*]
Brandeis University
Waltham, MA
tet@brandeis.edu

Colin Pollock
Yelp Inc
San Francisco, CA
cpollock@yelp.com

Benjamin Goldenberg
Yelp Inc
San Francisco, CA
benjamin@yelp.com

Jason Fennell
Yelp Inc
San Francisco, CA
jfennell@yelp.com

Abstract

With the overabundance of online user-generated content, the ability to filter based on relevant perspectives is becoming increasingly important. Identifying the perspective of the authors with the review text would enhance the retrieval of pertinent information. This problem can be traditionally formulated as a text classification task and solved by annotating the data and building a supervised learning system. However, rare classes might render annotation even more difficult and expensive. Here, we used a distant supervision approach to identify restaurant reviews that were written from the perspective of a vegetarian, and we achieved a macro-average F1 score of 79.40% with minimal annotation effort.

1 Introduction

The center of the information world has shifted from select few authorities to the global wisdom of the crowd and user-generated content. While useful and large, the volume of the information requires efficient organization, information retrieval, and data mining techniques to select the most relevant contents. For example, restaurant goers looking for vegetarian-friendly restaurants might want to read restaurant reviews that are written by a vegetarian. Authors' perspectives potentially provide a meaningful axis along which the documents can be organized.

Past studies have formulated this problem as a document-level supervised text classification problem (Manning et al., 2008), but the supervised learning paradigm might not be suitable in certain scenarios. Supervised learning algorithms can achieve superior performance compared to unsupervised learning algorithms at the expense of costly annotation efforts in creating labeled datasets for the algorithms to learn from. The perspectives that we would like to identify, however, might be very specific and somewhat rare in the document collection. To continue with the restaurant review example, only an estimated of 3.2% of the U.S. population identify themselves as vegetarian (Haddad and Tanzman, 2003), so the restaurant reviews written from the perspective of a vegetarian might be very rare in the corpus.

This rare class problem necessitates a larger number of annotated documents to collect sufficient positive examples. For instance, approximately 10,000 data instances must be labeled in order to obtain a mere 320 data points for vegetarian reviews. Additionally, the resulting classifier trained on a specific annotated corpus will tend to be biased toward that text domain, and the performance might degenerate when the classifier is applied to text in another domain. Although the rare class problem is well studied in supervised settings (He and Ghodsi, 2010; Joshi et al., 2001), to our knowledge we have not encountered a distantly supervised algorithm applied to a dataset where a rare class is of interest.

Distant supervision approaches address these problems by exploiting prior knowledge or external resources to gather large number of training data or features to train a classifier without manual annotation. A distant supervision algorithm might start by using a set of simple rules or a knowledge base to form distant supervision criteria (Mintz et al., 2009), then create an ini-

[*] The author conducted the work during his internship at Yelp.

International Joint Conference on Natural Language Processing, pages 922–926,
Nagoya, Japan, 14-18 October 2013.

tial training set from such criteria. The labels of these initial training samples are sometimes said to be *weakly labeled* because they are not individually supervised or manually labeled by a human annotator. Instead, the labels are distantly supervised by heuristics or informed by an extensive knowledge base. For instance, a distant supervision approach has been applied to Twitter data, which is massive and hard to annotate (Marchetti-Bowick and Chambers, 2012). Emoticons were used to identify tweets with positive or negative sentiments, which then served as training samples for supervised classifiers (Go et al., 2009). Notably, distant supervision approaches were successfully used in relation extraction. Mintz et al. (2009) employed a knowledge base to extract patterns and features for a relation extraction system, where the supervised training data were relatively small and domain specific.

The supervised learning paradigm might not suit ever-growing user-generated datasets that contain rare classes and suffer from prohibitive annotation cost. Here, we present a distant supervision approach for identifying rare author's perspectives in unstructured user-generated content. This method alleviates the problem of rare classes and reduces the time and cost of annotating data.

2 Corpus and Task Description

We randomly selected ten million user-generated restaurant reviews in English from yelp.com, a consumer review website.[1] The review authors' personal information was removed from the data. Most of the text consists of well formed sentences, due to the greater character limit than, say, Twitter. Although, like most unstructured user-generated corpora, these reviews contain typos and non-standard structure, e.g. ASCII art and use of dashes as bullet points. Each review contains 151.04 tokens on average, so we have a total of approximately 1.5 billion tokens in this dataset.

For this corpus, we focus on identifying the restaurant reviews that are written from the perspective of a vegetarian. We annotated a small number of reviews to use as a test set for final evaluation. Each review was annotated independently by two annotators. The inter-annotator agreement is moderate (Cohen's $\kappa = 0.58$). Only 34 reviews out of 1,021 labeled reviews are labeled as written from the perspective of a vegetarian, which suggests that this perspective occurs rarely in the corpus.

3 Methodology

We employed simple phrase matching to collect weakly labeled data. If a review contains the phrase "I'm a vegetarian," "I'm vegetarian," or "As a vegetarian, I," we regard those reviews as written from the perspective of a vegetarian. On the other hand, if a review mentions beef, pork, or chicken without the word "fake" preceding it, then such review is tagged as not written from a perspective of a vegetarian. These simple phrase matching rules are applied to every review in the corpus. The reviews that are not selected by the rules are discarded from the weakly labeled training data.

This weakly labeled training set is used to train a two-way classification Multinomial Naive Bayes model. The classifier uses all of the words in the documents as features. Weakly labeled data are noisier than manually labeled data, which might cause the classifier to be less generalizable due to overwhelming noisy features. Thus, we perform feature selection by using the Bayesian Information Criterion (BIC) to reduce noise and improve the performance (Schwarz, 1978; Forman, 2003). We noted that the proportion of each perspective in the weakly labeled data does not necessarily match the true proportion. We therefore manually set the prior probabilities of the labels to 0.9 and 0.1 for non-vegetarian and vegetarian respectively.

In general, one can use any arbitrary criteria to cull weakly labeled data from the corpus, as long as the criteria are high in precision. If the corpus is massive, which is usually the case in user-generated content on the web, then we afford to sacrifice recall for less noisy training instances. With regard to training classifiers, one can choose any supervised learning algorithm.

4 Experiment Setup and Results

Our distant supervision criteria identified 12,514 reviews written from the perspective of a vegetarian, and 3,076,256 reviews not written from such perspective. 7,193,878 reviews were left unannotated and discarded because they don't contain any of the phrases in our criteria. These reviews constitute weakly labeled training data and account for roughly 30% of the original unlabeled data of ten million reviews.

[1] The corpus is available upon request on the website `www.yelp.com/academic_dataset`

	Precision	Recall	F_1
Vegetarian	80.85	80.07	80.46
Non-vegetarian	97.50	97.62	97.56
Macro-average	89.17	88.84	89.01

Table 1: 7-fold cross-validation results based on the weakly labeled data. The classifier achieved the macro-averaged F_1 measure of 89.01.

4.1 Experiment 1

To evaluate the distant supervision method as it is applied to this task, we ran a 7-fold cross-validation on the weakly labeled training data and computed precision, recall, and F_1 measure for each class. The words used to collect weakly labeled data build the nearly perfect classifier features, therefore we excluded those words from the feature set before training the classifier. The results are shown in Table 1. We achieve the macro-average F_1 score of 89.01% and the accuracy rate of 95.67%.

4.2 Experiment 2

The evaluation based on the 7-fold cross-validation over the weakly labeled data might not accurately reflect how well the resulting classification will perform when applied to the unlabeled dataset. We evaluated the classifier on the manually annotated test set detailed in the earlier section. Like the first experiment, all of the words and phrases involved in gathering weakly labeled data were excluded from the feature set for training a classifier. The test set has a total of 1,021 labeled reviews, none of which overlaps with the original unlabeled dataset.

The classifier was evaluated on four different subsets of the test set to see the performance of the system in different scenarios:

1. All reviews in the test set. The reviews are for restaurants which also include bars and coffee shops, where the perspectives of vegetarians are even more rare or not applicable.

2. All reviews in the test set that are longer than 250 characters. Some reviews are too short to contain useful information for the vegetarian perspectives. This subset contains 623 reviews.

3. Food reviews only. In this scenario, we exclude reviews for bars and coffee shops. We were left with 316 reviews.

4. Food reviews that are longer than 250 characters. This subset contains 270 reviews.

The classifier attained the best performance when tested on food reviews longer than 250 characters. In this scenario, it achieved the macro-averaged F_1 score of 79.40% and an accuracy rate of 92.22%. The baseline accuracy by guessing non-vegetarian for all reviews is 88.88%. The performance report for all scenarios is summarized in Table 2.

The F_1 scores for vegetarian perspectives are lower across all experimental conditions possibly because of the highly imbalanced label distribution or insufficient positive training samples. Since any supervised algorithms can be integrated with our distant supervision approach, these problems can be remedied by downsampling or models that are robust to imbalanced data (Japkowicz, 2000).

It is important to note that the set of rules alone do not make any prediction on the label of our test set, because the phrases that constitute the rules do not match any of the reviews in the test set. Therefore, the distantly supervised training is necessary to build a classifier.

5 Discussion

In our proposed method for identifying vegetarian-written reviews, we exploited the fact that some of the reviews are already weakly labeled, motivating the words and phrases used for the distant supervision criteria. The key step for this approach is writing rules or criteria for collecting the weakly labeled data. As we focus on massive unstructured user-generated text, the criteria we use must be highly precise to prevent mislabeled data from being introduced into the training process. Although high-precision rules by definition will only recognize a small percentage of the positive samples, the problem is remedied by the fact that user-generated data are massive and constantly grow larger. In this study, our restrictive rules yield 12,514 reviews written from the perspective of a vegetarian, which is approximately 0.001% of the original dataset but suffices to build a classifier, as shown by the evaluation result.

Our approach can be thought of as similar to the bootstrapping technique, which has been explored extensively in the context of relation extraction (Gabbard et al., 2011) and text classifica-

		Vegetarian			Non-vegetarian			Overall	
	size	P	R	F1	P	R	F1	Acc.	F1
Long food reviews	270	66.67	60.00	63.15	95.06	96.25	95.65	92.22	79.40
Long reviews only	623	71.42	50.00	58.82	93.97	97.50	95.70	92.22	77.26
Food reviews only	316	64.51	58.82	61.53	95.08	96.09	95.59	92.08	78.56
All reviews	1,021	32.22	58.52	41.66	98.54	95.74	97.12	94.51	69.39

Table 2: Evaluation result based on the manually annotated test set. The reviews that contain more than 250 characters are considered long. The system performs the best when tested with long food reviews only.

tion (Mccallum, 1999). A bootstrapping algorithm starts with a small set of annotated seed training instances. Classifier training or pattern extraction is done based on the seed instances and then used to reap more training instances from the unlabeled data. This cycle continues for multiple iterations, and the performance is monitored at each iteration to ensure the improvement. A downside of this approach is that one bad iteration might introduce many mislabeled instances, degenerating the algorithm. In practice, extra human supervision must then check if the new training instances in each iteration are acceptable (Freedman et al., 2011). Our distant supervision approach requires human supervision only when writing criteria for initial training instances.

Unlike bootstrapping, our approach trains the classifier only once. Therefore, classifiers that take long to train such as Support Vector Machine can be trained within reasonable amount of time. When paired with automatic feature selection like the one used in this study, building a distantly supervised classifier is a matter of coming up with high-precision criteria to initiate the algorithm. If the criteria for distant supervision are precise enough, very little noise will be introduced into the training instances. These characteristics of our approach are attractive for massive data from user-generated content that might render computational cost of bootstrapping too costly.

6 Conclusion and Future Directions

We presented a distant supervision approach to identify authors' perspectives in an unstructured, user-generated, and possibly massive corpus. Our experiment shows that high-precision restrictive rules can potentially gather weakly labeled data to train a classifier robust enough to perform well on the rest of the corpus. This method demonstrates the potential to enhance user experience on a user-generated business review website like www.yelp.com, by allowing an information-retrieval system that can fetch documents based on authors' perspectives.

As a future direction, this similar method can be applied to massive user-generated microblogs like Twitter data to identify authors' perspectives. For example, if one could identify each tweet as written by a Republican or a Democrat, one might be able to mine opinions from each political party separately. One could also endeavor to identify documents written by the same authors. For instance, a vegetarian is more likely to write from the perspective of a vegetarian, so we can restrict distant supervision rules to require consistent labels for the same authors in order for the reviews to enter the training set.

References

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

Marjorie Freedman, Lance Ramshaw, Elizabeth Boschee, Ryan Gabbard, Gary Kratkiewicz, Nicolas Ward, and Ralph Weischedel. 2011. Extreme extraction: machine reading in a week. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1437–1446. Association for Computational Linguistics.

Ryan Gabbard, Marjorie Freedman, and Ralph Weischedel. 2011. Coreference for learning to extract relations: yes, virginia, coreference matters. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2, pages 288–293.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Ella H Haddad and Jay S Tanzman. 2003. What do vegetarians in the united states eat? *The American journal of clinical nutrition*, 78(3):626S–632S.

He He and Ali Ghodsi. 2010. Rare class classification by support vector machine. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 548–551. IEEE.

Nathalie Japkowicz. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68.

Mahesh V Joshi, Ramesh C Agarwal, and Vipin Kumar. 2001. Mining needle in a haystack: classifying rare classes via two-phase rule induction. In *ACM SIGMOD Record*, volume 30, pages 91–102. ACM.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with twitter. *EACL 2012*, page 603.

Andrew Mccallum. 1999. Text classification by bootstrapping with keywords, em and shrinkage. In *In ACL99-Workshop for Unsupervised Learning in Natural Language Processing*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.