Topic Classification of Blog Posts Using Distant Supervision

Stephanie D. Husby
University of Alberta
shusby@ualberta.ca

Denilson Barbosa
University of Alberta
denilson@ualberta.ca

Abstract

Classifying blog posts by topics is useful for applications such as search and marketing. However, topic classification is time consuming and error prone, especially in an open domain such as the blogosphere. The state-of-the-art relies on supervised methods, requiring considerable training effort, that use the whole corpus vocabulary as features, demanding considerable memory to process. We show an effective alternative whereby *distant* supervision is used to obtain training data: we use Wikipedia articles labelled with Freebase *domains*. We address the memory requirements by using only named entities as features. We test our classifier on a sample of blog posts, and report up to 0.69 accuracy for multi-class labelling and 0.9 for binary classification.

1 Introduction

With the ever increasing popularity of blogging grows the need of finding ways for better organizing the *blogosphere*. Besides identifying SPAM from legitimate blogs, one promising idea is to *classify* blog posts into *topics* such as travel, sports, religion, and so on, which could lead to better ways of exploring the blogosphere. Besides navigation, blog classification can be useful as a data preprocessing step before other forms of analysis can be done: for example companies can view the perception and reception of products, movies, books and more based on opinions in blogs of different segments.

We approach the problem by using machine learning. In particular, in the development of a learning-based classifier, two crucial tasks are the choice of the features and the building of training data. We adopt a novel approach when selecting features: we use an off-the-shelf Named Entity Recognition (NER) tool to identify entities in the text. Our hypothesis is that one can detect the topic of a post by focusing on the entities discussed in the post. Previous text classification tools use the entire vocabulary as potential features, which is a superset of our feature set. Our results show that despite using a smaller feature set, our method can achieve very high accuracy.

Obtaining training data is a challenge for most learning tools, as it often involves manual inspection of hundreds or thousands of examples. We address this by using *distant supervision*, where a separate dataset is used to obtain training data for the classifier. The distant dataset used here is Freebase[1], which is an open online database, along with related Wikipedia articles. The classes in our tests are *domains* in Freebase, which are defined by their curators.

Summary of Results. For our evaluation, we use a large sample of blog posts from a public snapshot of the blogosphere, collected around 2008. These posts are manually labeled by volunteers (undergraduate students in Computing Science), and used as the ground-truth test data.

Our results indicate that training a classifier relying on named entities using Freebase and Wikipedia, can achieve high accuracy levels on manually annotated data. We also identify some potential problems related to selecting the categories to be used in the classification. Overall, our results indicate that robust classifiers are possible using off-the-shelf tools and freely available

[1] http://www.freebase.com/.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 28–36,
Avignon, France, April 23 - 27 2012.

training data.

2 Related Work

Our work is related to topic identification techniques such as Latent Dirichlet Analysis (LDA), Latent Semantic Analysis (LSA) and Latent Semantic Indexing (LSI) (Steyvers and Griffiths, 2007). These techniques infer possible topic classes, and use unsupervised learning (clustering) approaches. In contrast, our technique allows the specification of classes (topics) *of interest* and attempts to classify text within those classes only. Next we discuss two other lines of work more closely related to ours.

Blog classification. There have been few attempts at classifying blog posts by topic. Most previous methods focus on classification of the authors and the sentiment of the posts.

Ikeda *et al.* (2008) discussed the classification of blog authors by gender and age. They use a semi-supervised technique and look at the blogs in groups of two or more. These groupings are based on which are relatively similar and relatively different. They assume that multiple entries from the same author are more similar to each other than to posts from other blogs, and use this to train the classifier. The classifier they use is support vector machines, and the bag-of-words feature representation. Thus, they consider all unique words in their classification. They find their methods to be 70-95% accurate on age classification, depending on the particular age class (*i.e.* the 20s vs the 30s class is more difficult to distinguish than the 20s vs the 50s) and up to 91% accurate on gender identification. This is quite different than the approach presented here, as we are examining topic classification.

Yang *et al.* (2007) consider the sentiment (positive or negative) analysis of blog posts. Their classifier is trained at the sentence level and applied to the entire document. They use emoticons to first create training data and then use support vector machines and conditional random fields in the actual classification. They use individual words as features and find that conditional random fields outperform support vector machines. This paper works both with blog posts and distance learning based on the emoticons, however this type of distant supervision is slightly different than our approach. It may also be referred to as using weakly labeled data.

Elgersma and de Rijke (2008) classify blogs as personal vs non-personal. The authors define personal blogs as diary or journal, presenting personal accounts of daily life and intimate thoughts and feelings. They use the frequency of words more often used in personal blogs versus those more frequently used in general blogs, pronouns, in-links, out-links and hosts as the features for the blogs. They then perform supervised training on the data using a set of 152 manually labeled blogs to train their classifier. The results show that the decision tree method produced the highest accuracy at about 90% (Elgersma and de Rijke, 2008).

A work which looks at true topic classification of blogs, as is being done here, is that of Hashimoto and Kurohashi (2008), who use a *domain dictionary* to classify blog posts without machine learning (i.e., using a rule-based system). They use keywords for each domain, or category as the basis for classification. They then create a score of a blog post based on the number of keywords from each domain; the domain with the highest count becomes the category for that post. They also expand the keywords in their domain by adding new words on the fly. This is done by taking an unknown word (one that does not currently exist in a domain) and attempting to categorize it using its online search results and/or Wikipedia article. They attempt to classify the results or article and then, in turn, classify the word. They find their classification method to be up to 99% accurate. This idea can be related to the use of Freebase as the domain dictionary in the current problem, but will be expanded to include machine learning techniques, which these authors avoid.

Distant supervision. Distant supervision is a relatively new idea in the field of machine learning. The term was first introduced by Mintz *et al.* (2009) in 2009 in their paper on *relation extraction*. The idea is to use facts in Freebase to obtain training data (i.e., provide distant supervision), based on the premise that if a pair of entities that have a relation in Freebase, it will likely be expressed in some way in a new context. They found their approach to be about 66-69% accurate on large amounts of data. Although the goal of their work (namely, extracting relations from the text) was different from ours, the use of Freebase and entities is directly related to the work

presented here.

Go *et al.* (Go et al., 2009) use distant supervision to label the sentiment associated with Twitter posts. They use tweets containing emoticons to label the training data, as follows. If a tweet contains a :) or an : (then it is considered to have a positive or a negative sentiment. Those tweets with multiple emoticons were discarded. Then emoticons themselves are *removed* from all data (to avoid them being used as features), and the labeled data is used to train the classifier. They found their approach to be around 78-83% accurate using several different machine learning techniques (Go et al., 2009). The authors do not discuss their feature representations in detail, but make use of both unigrams and bigrams.

Phan *et al.* (Phan et al., 2008) consider using a *universal data set* to train a classifier for web data similar to blogs . This idea is very similar to the concept of distant supervision. They consider Wikipedia and MEDLINE, as universal data sets, and they use the maximum entropy as their classifier. They apply their methods to two problems, topic clustering of web search results and disease classification for medical abstracts; they report accuracy levels around 80%.

3 Method

Our hypothesis is that one can predict the topic of a blog post based on "what" that post is about. More precisely, we focus on the recognizable named entities that appear in the blog post. Our intuition is that if a blog post mentions "Barack Obama" and the "White House" prominently, it is probably a post about politics. On the other hand, a post mentioning "Edmonton Oilers" and "Boston Bruins" is most likely about hockey. Naturally, there will be posts mentioning entities from different topics, say for example, a comment about the president attending a hockey game. In such cases, our hypothesis is that the other entities in the same post would help break the tie as to which class the post belongs to.

Our method consists of using a classifier trained with all topics of interest. We obtain training data using distant supervision, as follows. The topics come from Freebase, an open, online database compiled by volunteers. At the time of writing, it contains approximately 22 million objects which belong to one or more of a total of 86 domains. Each object in Freebase is a unique person, place, thing or concept that exists in the world. An example of an entity would be "Barack Obama" or "republican". A major data source for Freebase is Wikipedia; indeed, there is even a one-to-one mapping between articles in Wikipedia and the corresponding objects in Freebase.

Category	Articles	Distinct Entities
government	2,000	265,974
celebrities	1,605	85,491
food & drink	2,000	70,000
religion	2,000	175,948
sports	2,000	189,748
travel	2,000	125,802
other	2,000	384,139

Table 1: Topic categories chosen from Freebase domains

Discussion. Our motivation to use Freebase and Wikipedia comes from their large size and free availability, besides the fact these are fairly high quality resources–given the dedication of their contributors. It should be noted that this is a perfect example where distant supervision comes as an ideal approach, in the sense that the classification of objects into domains (i.e., topics) is done manually, and with great care, leading to high quality training data. Moreover, the nature of both datasets, which allow any web user to update and contribute to them, leads us to believe they will remain up-to-date, and will likely contain mentions to recent events which the bloggers would be discussing. Thus, one should expect a high overlap between the named entities in these resources and the blog posts.

3.1 Classifying Blog Posts

The classification of blog posts by topic is done by using the named entity recognition tool to extract all named entities (features) for the blog post, and feeding those to the topic classifier. We consider two classification tasks:

- **Multi-class**: In this case, we are given a blog post and the task is to determine which of the 7 topics (as in Table 1) it belongs to.
- **Binary classification:** In this case, we are given a blog post and a specific topic (i.e.,

	Blog (Test) Data		Wikipedia (Training) Data	
	words/post	**entities/post**	**words/article**	**entities/article**
celebrities	420	49	2,411	311
food & drink	256	28	1,782	144
government	20,176	2,363	6,013	803
other	395	50	10,930	1,245
religion	516	52	3,496	402
sports	498	73	4,716	741
travel	359	41	2,101	239

Table 2: Average word count and entity count per blog post and per Wikipedia article.

class), and the task is to determine whether or not the post belongs in that topic.

The multi-class task is more relevant in an exploratory scenario, where the user would browse through a collection of posts and use the classifier as a means to organize such exploration. The binary classification, on the other hand, is more relevant in a scenario where the user has a specific need. For example, a journalist interested in politics would rather use a classifier that filtered out posts which are not relevant. By their nature, the binary classification task demands higher accuracy.

Features The only features that make sense to use in our classification are those named entities that appear both in the training data (Wikipedia) and the test data (the blog posts). That is, we use only those entities which exist in at least one blog post **and** in at least one Wikipedia article. It is worth mentioning that this reduces drastically the memory needed for classification, compared to previous methods that use the entire vocabulary as features.

Each data point (blog or Wikipedia article) is represented by a vector, where each column of the vector is an entity. Two feature representations were created:

- **In-out:** in this representation we record the presence (1) or absences (0) of the named entity in the data point; and
- **Count:** in this representation we record the number of times the named entity appears in the data point.

	In-Out		Count	
	10-Fold	*Test*	*10-Fold*	*Test*
NB	0.59	0.37	0.51	0.29
SVM	0.26	0.18	0.49	0.22
NBM	0.71	0.57	0.68	**0.60**

Table 3: Summary of Accuracy on Multi-Class Data

4 Experimental Design

We collected the training data as follows. First, we discarded generic Freebase domains such as *Common* and *Metaweb System Types*, which do not correspond to meaningful topics. We also discarded other domains which were too narrow, comprising only a few objects. We then concentrated on domains for which we could find many objects and for which we could perform a reasonable evaluation. For the purposes of this paper, the 7 domains shown in Table 1 were used as topics. For each topic, we find all Freebase objects and their corresponding Wikipedia articles, and we collect the 2,000 longest articles (as those are most likely to contain the most named entities). The exception was the celebrities topic, for which only 1,605 articles were used. From these articles, we extract the named entities (i.e., the features), thus obtaining our training data. In the end, we used 4,000 articles for each binary classification experiment and 13,605 for the multi-class one.

As for test data, we used the ICWSM 2009 Spinn3r Blog Dataset (Burton et al., 2009), which was collected during the summer of 2008, coinciding with the build-up for the 2008 Presidential Elections in the US. In total, the collections has approximately 25M blog posts in English. For

a	*b*	*c*	*d*	*e*	*f*	*g*	← *classified as*	
0	0	0	0	0	0	50	*a*	*celebrities*
0	0	0	0	0	0	50	*b*	*food & drink*
0	0	15	27	0	0	8	*c*	*government*
0	0	0	0	0	0	50	*d*	*other*
0	0	0	0	0	0	50	*e*	*religion*
0	0	0	0	0	0	50	*f*	*sports*
0	0	0	0	0	0	50	*g*	*travel*

Table 4: Confusion Matrix of SVM on Test Set with In-Out Rep.

our evaluations, we relied on volunteers[2] who labeled hundreds of blogs, chosen among the most popular ones (this information is provided in the dataset), until we collected 50 blogs for each category. For the binary classifications, we used 50 blogs as positive examples and 200 blogs randomly chosen from the other topics as negative examples. For the multi-class experiment, we use the 350 blogs corresponding to the 7 categories.

Both the blogs and the Wikipedia articles were tagged using the Stanford Named Entity Recognizer (Finkel et al., 2005), which labels the entities according to these types: *Time, Location, Organization, Person, Money, Percent, Date,* and *Miscellaneous*. After several tests, we found that *Location, Organization, Person* and *Miscellaneous* were the most useful for topic classification, and we thus ignored the rest for the results presented here. As mentioned above, we use only the named entities in both the training and test data, which, in our experiments, consisted of 14,995 unique entities.

Classifiers. We performed all our tests using the Weka suite (Hall et al., 2009), and we tested the following classifiers. The first was the Naive Bayes (John and Langley, 1995) (NB for short), which has been successfully applied to text classification problems (Manning et al., 2008). It assumes attribute independence, which makes learning simpler when the number of attributes is large. A variation of the NB classifier, called Naive Bayes Multinomial (NBM) (McCallum and Nigam, 1998), was also tested, as it was shown to perform better for text classification tasks in which the vocabulary is large (as in our case). Finally, we also used the LibSVM classifier (Chang and Lin, 2001) (SVM), which is an implementation of support vector machines, a binary linear classifier. The results reported in this paper were obtained with LibSVM's default tuning parameters. SVMs are often used successfully in text classification problems (Ikeda et al., 2008; Yang et al., 2007; Go et al., 2009). These classifiers were chosen specifically due to their success rate with text classification as well as with other applications of distant supervision.

[2]Undergraduate students in our lab.

	In-Out		**Count**	
	10-Fold	*Test*	*10-Fold*	*Test*
NB	0.66	0.59	0.58	0.32
SVM	0.33	0.22	0.53	0.22
NBM	0.76	0.64	0.72	**0.64**

Table 5: Summary of Accuracy on Multi-Class without *Travel*

a	*b*	*c*	*d*	*e*	← *classified as*	
46	0	0	3	1	*a*	*celebrities*
3	25	21	0	1	*b*	*government*
40	2	0	3	5	*c*	*other*
5	1	1	43	0	*d*	*religion*
13	0	0	0	37	*e*	*sports*

Table 6: Confusion Matrix of NB on Test Set with In-Out Rep

5 Experimental Results

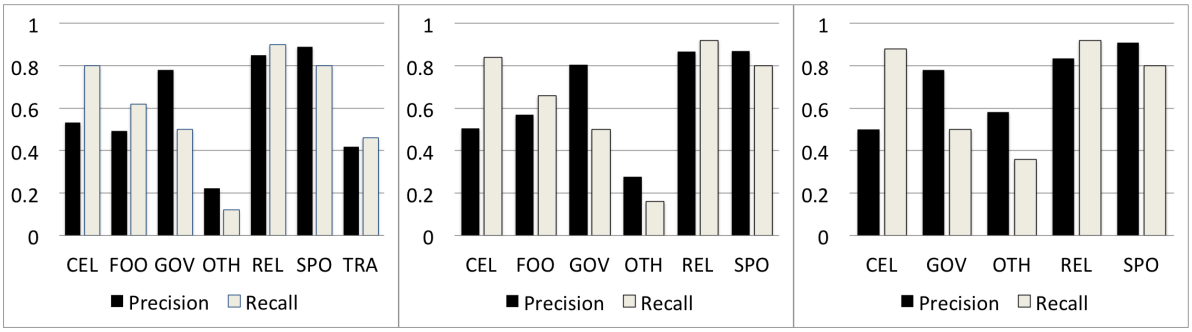
We now present our experimental results, starting with the multi-class task, in which the goal is to classify each post into one of 7 possible classes (as in Figure 1).

Accuracy in the Multi-class Task We report accuracy numbers both for 10-fold cross validation (on the training data) as well as on the manually labelled blog posts (test data). The summary of results is given in Table 3. Accuracy as high as 60% was obtained using the NBM classifier. The standard NB technique performed quite poorly in this case; as expected, NBM outperformed NB by a factor of almost two, using the count representation. Overall, the count representation produced better results than in-out on the test data, while losing on the cross-validation tests. Surprisingly, SVM performed very poorly in our tests.

These results were not as high as expected, so

	In-Out		Count	
	10-Fold	*Test*	*10-Fold*	*Test*
NB	0.70	0.60	0.62	0.40
SVM	0.47	0.38	0.67	0.40
NBM	0.79	0.67	0.76	**0.69**

Table 7: Summary of Accuracy on Multi-Class sans Travel, Food



(a) (b) (c)

Figure 1: Precision and Recall for Multi-Class Results Using Count Representation. Legend: CEL (Celebrities), FOO (food & drink), GOV (government), OTH (other), REL (religion), SPO (sport), TRA (travel).

we inspected why that was the case. What we found was that the classifiers were strongly biased towards the travel topic: NB, for instance, classified 211/350=60% of the samples that way, instead of the expected 14% (50/350). In the case of SVM, this effect was more pronounced: 88% of the posts were classified as *travel*. Table 4 shows the confusion matrix for the worst results in our tests (SVM with in-out feature representation), and fully illustrates the point.

We then repeated the tests after removing the *travel* topic, resulting in an increase in accuracy of about 5%, as shown in Table 5. However, another inspection at the confusion matrices in this case revealed that the *food & drink* class received a disproportionate number of classifications.

The highest accuracy numbers we obtained for the multi-class setting were when we further removed the *food & drink* class (Table 7). Consistent with previous results, our highest accuracy was achieved with NBM using the count feature representation: 69%. Table 6. gives the confusion matrix for this task, using NB. We can see that the posts are much better distributed now than in the previous cases, approximating the ideal confusion matrix which would have only non-zero entries in the diagonal, signifying all instances were correctly classified.

Recall in Multi-Class experiment. Accuracy (or precision, as used in information retrieval) measures the fraction of correct answers among those provided by the classifier. A complementary performance metric is recall, which indicates the fraction of correctly classified instances out of the total instances of the class. Figure 1 shows the breakdown of precision and recall for each class using the NBM classifier, using the Count feature representation for the tests with all 7 classes (a), as well as after removing *travel* (b) and both *travel* and *food&drink* (c).

As one can see, the overall accuracy by class does change (and improves) as we remove *travel* and then *food&drink*. However, the most significant change is for the class *other*. On the other hand, both the accuracy and recall for *celebrities*, *religion* and *sports* remain virtually unchanged with the removal of these classes.

Discussion of Multi-class results. One clear conclusion from our tests is the superiority of NBM using Count features for this task. The margin of this superiority comes somewhat as a surprise in some cases, especially when one compares against **SVM**, but does not leave much room

for argument.

As expected, some classes are much easier to handle than others. Classes such as *celebrities* are expected to be hard as documents in this topic deal with everything about the celebrities, including their preferences in politics, sports, the food they like and the places they travel. Looking at Figure 1, one possible factor for the relatively lower performance for *travel* and *food & drink* could be that the training data in these categories have the lowest average word count and entity count (recall Table 2). Another category with relatively less counts is *celebrities*, which can also be explained by the lower document count (1,605 available articles relating to this topic in Freebase).

Another plausible explanation is that articles in some classes can often be classified in either topic. Articles in the *travel* topic can include information about many things that can be done and seen around the world, such as the culinary traits of the places being discussed and the celebrities that visited them, or the religious figures that represent them. Thus, one would expect some overlap among the named entities relating to these less well-defined classes. These concepts tie easily into the various other topic categories we have considered and help to explain why misclassification was higher for these cases.

We also observed that with the NBM results, in all three variations of the multi-class experiments, there was a fairly consistent trade-off between recall and precision for the *celebrities* class. The erroneous classification of posts into celebrities could be explained in a similar way to those in *food&travel*. The fact that celebrities can exist in sports, politics, and religion means that many of the posts may fit into two or more classes and explains the errors. The best way to explore this further would be to do multiple class labels per post rather than just choosing a single label.

One interesting point that Figure 1 supports is the following. Recall that the need for the class *other* is mostly to test whether the classifier can handle "noise" (blogs which are too general to be classified). With this in mind, the trend in Figure 1 (increasing classification performance as classes are removed) is encouraging, as it indicates that more focused classes (e.g., *religion* and *sports*) can actually be separated well by a classifier using distant supervision, even in the presence of less well-defined classes. Indeed, taken to the extreme, this argument would suggest that the performance in the binary classification scenario for such classes would be the highest (which is indeed the case as we discuss next).

5.1 Binary Classification

We now consider a different scenario, in which the task is to perform a *binary* classification. The goal is to identify posts of a specific class amongst posts of all other classes. The percentage of correctly classified posts (i.e. test data) in this task, based on each feature representation can be seen in Table 8.

Overall, all classifiers performed much better in this setting, although NBM still produced consistently better results, with accuracy in the mid-90% level for the count feature representation. It is worth noting that **SVM** performed much better for binary classifications compared to the multi-class experiments, in some cases tying or ever so slightly surpassing other methods.

Also, note that the classifiers do a much better job on the more focused classes (e.g., *religion*, *sports*), just as was the case with the multi-class scenario. In fact, the accuracy for such classes is near-perfect (92% for *religion* and 93% for *sports*).

6 Conclusion

This paper makes two observations. First, our novel approach of using a standard named entity tagger to extract features for classification does not compromise classification accuracy. Reducing the feature contributes to increasing the scalability of topic classification, compared to the state of the art which is to process the entire vocabulary. The second observation is that distant supervision is effective in obtaining training data: By using Freebase and Wikipedia to obtain training data for standard machine learning classifiers, accuracy as high as mid-90% were achieved on our binary classification task, and around 70% for the multi-class task.

Our tests confirmed the superiority of NBM for text classification tasks, which had been observed before. Moreover, our test also showed that this superior performance is very robust across a variety of settings. Our results also show that it is important to consider topics carefully, as there can be considerable overlap in many general classes

	In-Out			Count		
Class	**NB**	**NBM**	**SVM**	**NB**	**NBM**	**SVM**
religion	0.63	0.90	0.80	0.43	0.92	0.81
government	0.96	0.85	0.80	0.88	0.82	0.87
sports	0.62	0.79	0.79	0.90	0.93	0.79
celebrities	0.60	0.68	0.80	0.40	0.76	0.80
average	0.71	**0.81**	0.79	0.65	**0.86**	0.82

Table 8: Accuracy of Binary Classification.

and this can cause misclassification. Obviously, such overlap is inevitable–and indeed expecting that a single topic can be found for each post can be viewed as a restriction. The most straightforward way to overcome this is by allowing multiple class labels per sample, rather than forcing a single classification.

Given the difficulty of the task, we believe our results are a clear indication that distant supervision is a very promising option for topic classification of social media content.

Future Work. One immediate avenue for future work is understanding whether there are techniques that can separate the classes with high overlap, such as *celebrities*, *food&drinks* and *travel*. However, it is very hard even for humans to separate these classes, so it is not clear what level of accuracy can be achieved. Another option is to examine additional features which could improve the accuracy of the classifier without drastically increasing the costs. Features of the blog posts such as link structure and post length, which we disregarded, may improve classification.

Moreover, one could use unsupervised methods to find relations between the named entities and exploit those, e.g., for bootstrapping. A similar idea would be to exploit dependencies among relational terms involving entities, which could easily be done on blogs and the Wikipedia articles. Topic selection is another area for future work. Our selection of topics was very general and based on Freebase domains, but a more detailed study of how to select more specific topics would be worthwhile. For instance, one might want to further classify *government* into political parties, or issues (e.g., environment, energy, immigration, etc.).

Acknowledgements

This was supported in part by NSERC—Natural Sciences and Engineering Research Council, and the NSERC Business Intelligence Network (project Discerning Intelligence from Text).

References

K. Burton, A. Java, and I. Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

E. Elgersma and M. de Rijke. 2008. Personal vs non-personal blogs. *SIGIR*, July.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1301–1306. AAAI Press.

A. Go, R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reuteman, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11.

C. Hashimoto and S. Kurohashi. 2008. Blog categorization exploiting domain dictionary and dynamically estimated domains of unknown words. *Proceedings of ACL-08, HLT Short Papers (Companion Volume)*, pages 69–72, June.

D. Ikeda, H. Takamura, and M. Okumura. 2008. Semi-supervised learning for blog classification.

Association for the Advancement of Artificial Intelligence.

George H. John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

A. McCallum and K. Nigam. 1998. A comparison of event models for naive bayes text classification. *In AAAI-98 Workshop on Learning for Text Categorization*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *ACL-IJCNLP '09: 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*, 2:1003–1011.

X. Phan, L. Nguyen, and S. Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *International World Wide Web Conference Committee*, April.

M. Steyvers and T. Griffiths, 2007. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum.

C. Yang, K. Lin, and H. Chen. 2007. Emotion classification using web blog corpora.