TweetLDA: Supervised Topic Classification and Link Prediction in Twitter

Daniele Quercia
The Computer Laboratory
University of Cambridge
United Kingdom
dq209@cl.cam.ac.uk

Harry Askham
The Computer Laboratory
University of Cambridge
United Kingdom
ha293@cam.ac.uk

Jon Crowcroft
The Computer Laboratory
University of Cambridge
United Kingdom
jac22@cam.ac.uk

ABSTRACT

L-LDA is a new supervised topic model for assigning "topics" to a collection of documents (e.g., Twitter profiles). User studies have shown that *L-LDA* effectively performs a variety of tasks in Twitter that include not only assigning topics to profiles, but also re-ranking feeds, and suggesting new users to follow. Building upon these promising qualitative results, we here run an extensive quantitative evaluation of *L-LDA*. We test the extent to which, compared to the competitive baseline of Support Vector Machines (*SVM*), *L-LDA* is effective at two tasks: 1) assigning the correct topics to profiles; and 2) measuring the similarity of a profile pair. We find that *L-LDA* generally performs as well as *SVM*, and it clearly outperforms *SVM* when training data is limited, making it an ideal classification technique for infrequent topics and for (short) profiles of moderately active users. We have also built a web application that uses *L-LDA* to classify any given profile and graphically map predominant topics in specific geographic regions.

ACM Classification Keywords

H.3.1 Information Systems Applications

General Terms

Algorithms, Experimentation

INTRODUCTION

This work focuses on the task of document classification in Twitter - given a Twitter profile and a set of possible topics, determine which topics best fit the profile's tweets. Traditional supervised machine learning approaches of document classification are usually trained upon standard written English documents [4]. By contrast, tweets are short and noisy and, for this reason, one resorts to classification approaches that require little supervision.

One such approach is an unsupervised machine learning technique known as Latent Dirichlet Allocation (*LDA*) [1]. The main problem of *LDA* here is that it returns topics that are latent - its topics are simply numbered distributions over words, and it is of little use to an end-user to know a profile pertains mostly to one latent topic rather than another. To overcome this, Ramage *et al.* proposed a variation of *LDA* known as Labelled Latent Dirichlet Allocation (*L-LDA*) [5] that associates a document with easily-interpretable topics.

We thus set out to perform a quantitative evaluation of *L-LDA*. Our contributions are not algorithmic (we do not propose any new topic model) but focus on understanding how well a fairly new version of topic modeling (i.e., *L-LDA*) works in the specific context of Twitter, an increasingly useful source of informative textual data.

EVALUATION OF L-LDA

The goal of *L-LDA* is to correctly assign topics to Twitter profiles. To ascertain its effectiveness at meeting this goal, our evaluation makes two assessments. It measures the extent to which *L-LDA* predicts: *(1)* the correct topics for a profile; and *(2)* the similarity of profile pairs. Before performing these two evaluation steps, we need to crawl and prepare our dataset and produce the ground truth for it.

We use Twitter data previously collected [3]: tweets were captured using Twitter's Streaming API between the dates of 27 September and 10 December 2010, collecting at most 200 tweets for any one user (200 is the limit set by the API) and ending up having 31.5M tweets. We then process these tweets as follows by retaining all non-stopword tokens and stripping the corpus of its 29,983 least frequently occuring terms (10% of the vocabulary).

To perform our evaluation, we need to topically classify the dataset and produce a set of ground truths. We do so by pre-classifying the dataset using three text classification APIs. Had only a single API been used, bias might have been introduced into the pre-classification; the performance of one API may differ significantly from that of another for certain topics, for example. For this reason, we have used these three APIs: *Alchemy API*[1]; *OpenCalais API*[2]; and *Textwise SemanticHacker API*[3]. To then build a ground truth from each API, each profile has to be processed by all three APIs. An alternative to using these APIs as ground truths would have been manual dataset classification. However, human labeling ability is variable, and introduces biases[4] - our approach avoids

[1] http://www.alchemyapi.com/api/

[2] http://www.opencalais.com/documentation

[3] http://textwise.com/api/categorization

[4] Ideally, manual and API-based topical codings should be used together as they allow for different types of validation.

WebSci 2012, June 22–24, 2012, Evaston, Illinois, USA.

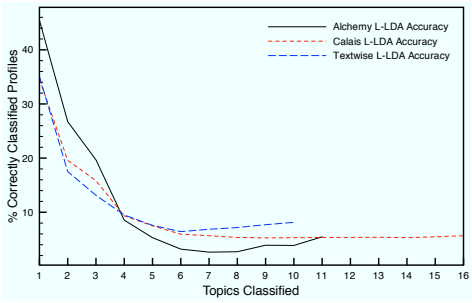


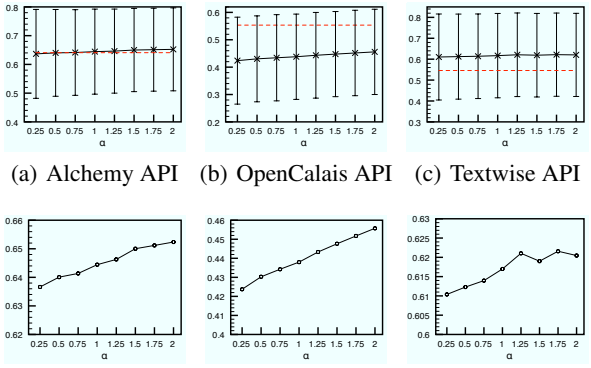
Figure 1. Accuracy of *L-LDA*, trained using each of the three ground truths, as an increasing number k of topics are considered. The concentration parameter α is fixed at 1.0.

this, and has produced consistent results across the three topic spaces evaluated.

Predicting a profile's topics

To evaluate how well *L-LDA* infers the correct topics for previously unseen Twitter profiles, for each of the three ground truths, we run a *10-fold* cross validation. That is, we divide the dataset into 10 segments, we take one segment s at a time, consider it to be the testing set and consider the remaining segments to be test sets. We then compare the inferred topics of each profile to those given by the ground truth. This is repeated for the three ground truths and for each value of the number k of topics per profile to be classified. Figure 1 shows the percentage of the profiles for which *L-LDA* was able to infer the correct top-k topics. Performance quickly tails off, as the task of determining a profile's top 2 or top 3 most likely topics is a more difficult one than that of single-topic classification. A small performance increase is noticed as k approaches the maximum number of topics in each ground truth - this can be attributed to the model successfully determining the least prominent topics in each profile's topic distribution. Values of k equal to the cardinality of each API's global topic set are omitted, as this artificially yields 100% accuracy. Results obtained by performing the above experiment under values of the Dirichlet hyper-parameter α between 0.25 and 2.0 gave similar results, fluctuating less than 1% from those reported in Figure 1. What these results show is that *L-LDA* can only accurately discover the most probable topics in a given profile; the probabilities it assigns to the less prominent topics in a distribution are further from those in the ground truth. However, this does not pose a problem, as one would not use less probable topics to classify a profile. For $k \leq 3$, Alchemy-trained *L-LDA* outperforms *L-LDA* models trained on both OpenCalais and Textwise (although these models perform better for other values of k). This could be an indicator that the Alchemy API has more accurately captured the topics of the profiles in the dataset than the other APIs, with the corresponding *L-LDA* correctly classifying around 9% more profiles than when it is trained on the other APIs, if only the top topic is considered (i.e., $k = 1$) .

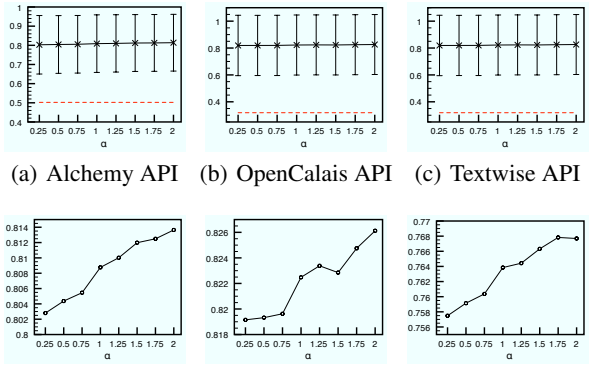
A more comprehensive evaluation would, however, compare *L-LDA*'s performance to that of a competitive baseline. Because of their simplicity and effectiveness, previous works have used Support Vector Machines (*SVM*) as a competitive



(a) Alchemy API (b) OpenCalais API (c) Textwise API

(d) Alchemy API (e) OpenCalais API (f) Textwise API

Figure 2. The average cosine similarity between *L-LDA* (with varying α) and each ground truth and the similarity between *SVM* (the dashed red line) and each ground truth. Error bars indicate ± 1 standard deviation. Graphs (d) to (f) are zoomed versions of graphs (a) to (c).



(a) Alchemy API (b) OpenCalais API (c) Textwise API

(d) Alchemy API (e) OpenCalais API (f) Textwise API

Figure 3. The average cosine 2-similarity between *L-LDA* (with varying α) and each ground truth and the similarity between *SVM* (the dashed red line) and each ground truth. Error bars indicate ± 1 standard deviation. Graphs (d) to (f) are zoomed versions of graphs (a) to (c).

baseline for the evaluation of topic models, including labeled topic models [5]. Therefore, our evaluation also considers *SVM* and, by performing a 10-fold cross validation, it sets the *SVM*'s parameters to the values that yield the best results[5]. To then determine which model performs best at determining the two most probable topics in a profile, we plot the cosine similarity for both *L-LDA* and *SVM*. Figure 2 shows that *L-LDA* manages to outperform *SVM* in some cases, such as when it is trained with *Alchemy* for $\alpha \geq 0.75$, and when it is trained with *Textwise* for all α. *SVM*, however, is more capable of topic distribution inference in the Calais topic space. Since Calais classifies over a larger topic space than the other APIs, this could be an indicator that *L-LDA* performs best when it has fewer topics to deal with.

In reality, a model does not need to correctly infer the ranking of all topics but rather only the ranking of top topics. We therefore define a metric called cosine k-similarity. This is

[5]These parameters are the kernel's parameters and the soft margin parameter C that determines the support vectors used to decide the hyperplane.

the cosine similarity of only the k greatest elements of the two topic distributions. In Figure 3, we compare the 2-similarity scores of both *L-LDA* and *SVM*. *L-LDA* greatly outperforms *SVM* in every case, in stark contrast to the previous results, proving that *L-LDA* can more accurate ascertain a profile's top two topics. More generally, *SVM* sees a performance increase as k increases, whereas *L-LDA*'s performance decreases.

For both cosine similarity and cosine 2-similarity, we have seen slight improvements in performance as *L-LDA*'s concentration parameter α increases. However, since these improvements are negligible compared to the standard errors (reflected in the error bars of Figures 2 and 3), it can be concluded that, within a reasonable range, the chosen value of α has little effect on *L-LDA*'s inference procedure.

In addition to the number k of topics to be assigned to each profile and to the Dirichlet hyper-parameter α, *L-LDA*'s performance might be affected by another parameter - a profile's topics. Some topics might be easier to identify than others. Determining for which topics *L-LDA* exhibits strong performance, and for which it fails to match up to the ground truth, will reveal its topic-specific strengths and weaknesses and will tells us for which applications the model may be best suited, and which areas do not lend themselves readily to latent semantic analysis. Firstly, because the number of profiles varies from topic to topic, we must know the topic density of the corpus, both before and after inference. Figure 4(a) shows this for the Alchemy ground truth and *L-LDA* inferred classifications. As expected, the inferred topic distribution of the corpus closely matches that of the ground truth. This does not necessarily tell us about *L-LDA*'s performance - a random classifier could achieve the same simply by assigning each topic with a probability proportional to that topic's prominence in its training data. Instead, to evaluate *L-LDA*'s performance, we compute topic-wise precision and recall measures (i.e., we compute the recall and precision of a topic over the entire set of users): precision is the number of correctly classified profiles over the number of *classified* profiles, and *recall* is number of correctly classified profiles over the number of *all* profiles. We then combine precision and recall in the composite metric of F-measure (or $F1$ score): $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$. It has to be stressed that whether or not an F_1 score is deemed a "good" performance measure is a domain-specific problem - in our case, it serves as a reasonable measure of *relative* performance. Figure 4(b) shows *L-LDA*'s F_1 for profiles of different subject matter. It can be seen, again, that changing the value of α has little effect on the results. This makes sense, as α governs how concentrated each profile's inferred topic distribution is, and smoothing or concentrating a probability distribution does not alter its most likely outcome (the top topic). Previous work [5] on *L-LDA* showed effective topic classification with F_1 scores of around 0.5, and so our results correlate with these previous results. A score of 0.55, as is the case in Figure 4(b), for topic "Computer_Internet", therefore indicates good performance at the task of inference. The small standard error (± 0.034) for this category indicates that the F_1 score represents its precision and recall scores almost equally. As one might expect, topics that contain a large number of subject-specific terminology (such as "Computer_Internet") are more readily identified by *L-LDA* than those consisting of more abstract concepts (such as"Religion"). *L-LDA* outperforms *SVM* in all cases, with the difference in performance being particularly noticeable when the quantity of training data is small. For instance, only 1.2% of the profiles have a ground truth classification of "Recreation", and only 2.0% "Law_crime". *L-LDA* exhibits reasonably strong performance in these cases, while *SVM* fails to accurately classify profiles with these topics. This shows us not only that *L-LDA* does not require large quantities of training data to achieve strong performance, but also that the model is not biased by the proportions of each topic that exist in its training set (otherwise, we would expect topics such as "Recreation" to have the lowest F-measures).

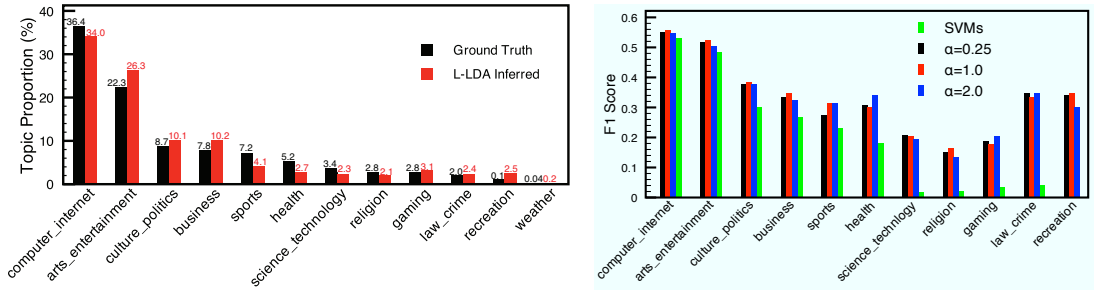
Ground truth	*L-LDA* ρ	*SVM* ρ
Alchemy	.60	.22
OpenCalais	.48	.17
Textwise	.40	.25

Table 1. Values of ρ obtained after comparing inferred pair-wise similarity rankings with those of *LDA*. The α of both *LDA* and *L-LDA* is fixed at 1.0.

Predicting the similarity of a pair of profiles

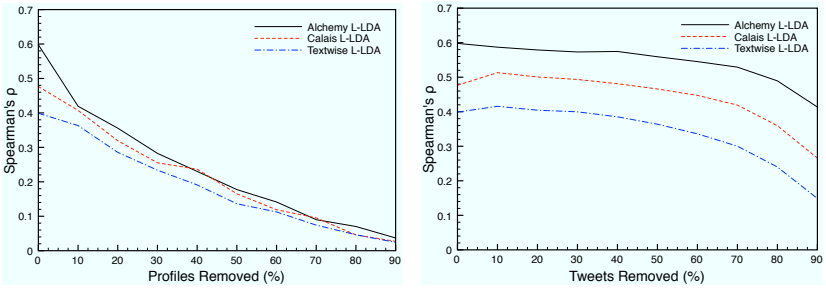
Previous work [2] has shown *LDA* to be very effective at determining how similar two Twitter profiles are (and this has been applied to the problem of "link prediction" in Twitter). We would hope that *L-LDA* can determine profile similarity as well as *LDA* - this would mean that *L-LDA* has the accuracy of *LDA*, with the added advantage of profile labels. This would make *L-LDA* a more desirable topic modeling technique than standard *LDA* in situations where the additional data to support semi-supervised learning is present. To test this hypothesized similarity between the two topic models, we calculate inferred topic distributions for each profile using LDA, calculate the cosine similarity of each profile pair, and accordingly rank profile pairs. We do the same using *L-LDA* (*SVM*) and produce another two rankings of profile pairs. We finally calculate the Spearman's rank correlation coefficient (*SRCC*) of the two pairs of rankings. The value of the *SRCC* (described below) tells us how similar *L-LDA*'s (*SVM*'s) ranking of profile pairs is to *LDA*'s. A higher value (closer to 1.0) shows that *L-LDA* (*SVM*) has performed well, and approximates *LDA*'s performance in this regard, whereas values closer to 0.0 (no correlation) show that *L-LDA* (*SVM*) disagrees with *LDA* about which profile pairs are the most similar. Table 1 contains the results of these evaluations and shows that *L-LDA* significantly outperforms *SVM* in determining profile similarity - *SVM* fails to recognize the cases where *LDA* has deemed two profiles similar, while *L-LDA* shows high correlation between its profile pair rankings and those of *LDA*.

This is not a particularly surprising result - the machinery of *L-LDA* is similar to that of *LDA*, and so we might expect them to perform similarly. It will be therefore more informative to see how the value of ρ changes as properties of the models, such as quantity of training data, are altered. To determine how *L-LDA* performs in the face of a reduced-size



(a) Topic distribution (b) F_1 score

Figure 4. (a) The proportion of topics in the ground truth (i.e., in *Alchemy* classifications) and the proportion of topics in the results produced by *L-LDA* with an α of 1.0. (b) For each topic, the F_1 scores of *L-LDA* and *SVM* inference (the ground truth is generated by *Alchemy*). Topic "weather" is omitted as both its precision and recall are 0.0, leaving F1 undefined. Topics on the x-axis are ordered by their predominance in the ground truth, as per panel (a).



(a) Corpus reduction by profile (b) Corpus reduction by tweet

Figure 5. The effects of removing profiles and tweets from the corpus on the value of Spearman's ρ. The value of α for both *LDA* and *L-LDA* is 1.0.

corpus, two further evaluations were performed. Firstly, a certain percentage of the profiles in the corpus are removed, and secondly, each profile remains present in the corpus, but each has a certain percentage of its tweets removed. In both cases, *L-LDA* cross validation was performed, and pair-wise Spearman rankings were calculated. This experiment was not performed for SVMs, as even with a full corpus, the results they yield show little to no correlation with the ground truth LDA values. This corpus reduction simulates many potential use cases for *L-LDA* - we might, for example, wish to train a model using a very specific subset of Twitter users (those in a certain location, for example), which is simulated by removing full profiles from the corpus. Another example might be that we wish only to consider tweets from a certain period of time, or only a small portion of a user's tweets (because downloading a full profile using the Twitter API can be a slow procedure) - this is simulated by removing a subset of each user's tweets. Figure 5(a) shows that *L-LDA*'s performance is strongly linked to the number of documents in its corpus. As more and more profiles are removed, *L-LDA*'s performance quickly degrades. With 90% of the profiles removed from the corpus, there is effectively no correlation between *L-LDA*'s inferred pair-wise rankings and those of *L-LDA*. Interestingly, however, Figure 5(b) shows that we are able to shrink the corpus by the same percentage while retaining much of *L-LDA*'s performance - removing some of the tweets in each profile does not alter the number of profiles in the corpus. We can remove around 70% of the tweets from each profile without significantly impacting *L-LDA*'s ability to train and learn from these profiles. This indicates that the subject matter of a profile can be represented by only a small sample of the tweets it contains, suggesting that *L-LDA* is a suitable profile classification method when only a small number of tweets exist for each training profile. Its effectiveness, however, may be limited when the number of training profiles is small.

Summary. *L-LDA* has proven effective at the task of Twitter profile classification, outperforming the competitive *SVM*. *L-LDA* can accurately classify a profile with topics for which it has seen only small amounts of training data and greatly outperforms SVMs at determining how similar a pair of profiles is, showing that *L-LDA*'s techniques of inference are preferable to the linear classification of each SVM when dealing with rich, mixed-topic documents such as Twitter profiles.

REFERENCES

1. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal Machine Learning Research* (March 2003).
2. Puniyani, K., Eisenstein, J., Cohen, S., and Xing, E. P. Social links from latent topics in Microblogs. In *Proceedings of the Workshop on Computational Linguistics in a World of Social Media NAACL HLT* (June 2010).
3. Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. In the Mood for Being Influential on Twitter. In *Proceedings of the 3^{rd} IEEE Conference on Social Computing (SocialCom)* (October 2011).
4. Ramage, D., Dumais, S., and Liebling, D. Characterizing Microblogs with Topic Models. In *AAAI Conference on Weblogs and Social Media (ICWSM)* (May 2010).
5. Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)* (August 2009).