IJCLA VOL. 5, NO. 1, JAN-JUN 2014, PP. 43–55
RECEIVED 06/01/14 ACCEPTED 06/02/14 FINAL 17/06/14

Topic Classification using Latent Dirichlet Allocation at Multiple Levels

DIANA INKPEN AND AMIR H. RAZAVI

University of Ottawa, Canada

ABSTRACT

We propose a novel low-dimensional text representation method for topic classification. Several Latent Dirichet Allocation (LDA) models are built on a large amount of unlabelled data, in order to extract potential topic clusters, at different levels of generalization. Each document is represented as a distribution over these topic clusters. We experiment with two datasets. We collected the first dataset from the FriendFeed social network and we manually annotated part of it with 10 general classes [1]. The second dataset is a standard text classification benchmark, Reuters 21578, the R8 subset (annotated with 8 classes). We show that classification based on our multi-level LDA representation leads to improved results for both datasets. Our representation catches topic distributions from generic ones to more specific ones and allows the machine learning algorithm choose the appropriate level of generalization for the task. Another advantage is the dimensionality reduction, which permitting the use of machine learning algorithms that cannot run on high-dimensional feature spaces. Even for the algorithms that can deal with high-dimensional features spaces, it is often useful to speed up the training and testing time by using the lower dimensionality.

1 INTRODUCTION

In order to improve the performance of text classification tasks, we always need informative and expressive methods to represent the texts [2, 3]. If we consider the words as the smallest informative unit of a text, there is a variety of well-known quantitative information measures that can be used to represent a text. Such methods have been used in a variety of information extraction projects, and in many cases have even outperformed some syntax-based approaches. There are a variety of Vector Space Models (VSM) which have been well explained and compared, for example in [4]. However, these kinds of representations disregard valuable knowledge that could be inferred by considering the different types of relations between the words. These major relations are actually the essential components that, at a higher level, could express concepts or explain the main topic of a text. A representation method which could add some kind of relations and dependencies to the raw information items, and illustrate the characteristics of a text at different conceptual levels, could play an important role in knowledge extraction, concept analysis and sentiment analysis tasks.

In this paper, the main focus is on how we represent the topics of the texts. Thus, we select a LDA topic-based representation method, and we extend it to a multi-level representation that can automatically choose the appropriate level of generality. Then, we run machine learning algorithms on each representation (or combinations), in order to explore the most discriminative representation for the task of text classification, for the two datasets that we selected.

2 Related Work

In most text classification tasks, the texts are represented as a set of independent units such as unigrams / bag of words (BOW), bigrams and/or multi-grams which construct the feature space, and the text is normally represented only by the assigned values (binary, frequency or term TF-IDF[1]) [5]. In this case, since most lexical features occur only a few times in each context, if at all, the representation vectors tend to be very sparse. This method has two disadvantages. First, very similar contexts may be represented by different features in the vector space. Second, in short and medium-size texts, we will have too many zero features for machine learning algorithms, including supervised classification methods.

Blei, Ng and Jordan proposed the Latent Dirichlet Allocation (LDA) model and a Variational Expectation-Maximization algorithm for training their model. LDA is a generative probabilistic model of a corpus and the idea behind it is that the documents are represented as weighted relevancy

[1] term frequency / inverse document frequency

vectors over latent topics, where a topic is characterized by a distribution over words. These topic models are a kind of hierarchical Bayesian models of a corpus [6]. The model can unveil the main themes of a corpus which can potentially be used to organize, search, and explore the documents of the corpus. In LDA models, a *topic* is a distribution over the feature space of the corpus and each document can be represented by several topics with different weights. The number of topics (clusters) and the proportion of vocabulary that create each topic (the number of words in a cluster) are considered as two hidden variables of the model. The conditional distribution of words in topics, given these variables, for an observed set of documents, is regarded as the main challenge of the model.

Griffiths and Steyvers in 2004, applied a derivation of the Gibbs sampling algorithm for learning LDA models [7]. They showed that the extracted topics capture a meaningful structure of the data. The captured structure is consistent with the class labels assigned by the authors of the articles that composed the dataset. The paper presents further applications of this analysis, such as identifying *hot topics* by examining temporal dynamics and tagging some abstracts to help exploring the semantic content. Since then, the Gibbs sampling algorithm was shown as more efficient than other LDA training methods, e.g., variational EM and Expectation-Propagation [8]. This efficiency is attributed to a famous attribute of LDA namely, "the conjugacy between the Dirichlet distribution and the multinomial likelihood". This means that the conjugate prior is useful, since the posterior distribution is the same as the prior, and it makes inference feasible; therefore, when we are doing sampling, the posterior sampling become easier. Hence, the Gibbs sampling algorithms was applied for inference in a variety of models that extend LDA [9–13].

Recently, Mimno et al. presented a hybrid algorithm for Bayesian topic modeling in which the main effort is to combine the efficiency of sparse Gibbs sampling with the scalability of online stochastic inference [14]. They used their algorithm to analyze a corpus that included 1.2 million books (33 billion words) with thousands of topics. They showed that their approach reduces the bias of variational inference and can be generalized by many Bayesian hidden-variable models.

LDA topics models started to be used in various Natural Language Processing tasks. It was used, among other tasks, for native language identification [15], for learning word classes [16], and for opinion analysis [17]. Supervised versions were developed, named labelled LDA, and applied, for example, for authorship attribution [18]. Experiments that used LDA topic models for a task of cross-language categorization of Wikipedia pages were presented in [19]. In this paper, we focus on the task of automatic text classification into a set of generic topics/subjects using multiple LDA models in the same time, in order to achieve different levels of generalization. Smet at al. [19] also used multiple LDA models (with 10 to 200 topics, with an increment of 20). We developend our method without being aware of their work, initally. The task and datasets that we used are different.

3 DATASETS

In order to properly evaluate our new multi-level LDA text representation, we conducted experiment with two datasets of different genres (social media text and newspaper text).

The first dataset that we prepared for our experiments consists of threads from the FriendFeed social network. We collected main postings (12,450,658) and their corresponding comments (3,749,890) in order to obtain all the discussion threads (a thread consists of a message and its follow up comments). We filtered out the threads with less than three comments. We were left with about 24,000 threads. From these, we used 4,000 randomly-selected threads as background source of data, in order to build the LDA model. We randomly selected 500 threads out of the 4000 and manually annotated them with 10 general classes[2], to use as training and test data for the classification task. The 10 classes are: *consumers, education, entertainment, lifestyle, politics, relationships, religion, science, social_life* and *technology*. We will make the dataset available (the whole corpus that we collected and the manually-annotated part).

We observed that the 10 class labels (general topics) are distributed unevenly over the dataset of 500 threads, in which we had 21 threads for the class *consumers*, 10 threads for *education*, 92 threads for *entertainment*, 28 threads for *incidents*, 90 threads for *lifestyle*, 27 threads for *politics*, 58 threads for *relationships*, 31 threads for *science*, 49 threads for *social_activities*, and 94 threads for *technology*.

The second dataset that we chose for our experiments is the well-known R8 subset of the Reuters-21578 collection (excerpted from the UCI machine learning repository), a typical text classification benchmark. The data includes the 8 most frequent classes of Reuteres-21578;

[2] We used only one annotator, but we had a second annotator check a small subset, in order to validate the quality of annotation. In future work, we plan to have a second annotator label all the 500 threads.

hence the topics that will be considered as class labels in our experiments are *acq, crude, earn, grain, interest, money, ship* and *trade*.

In order to follow Sebastiani's convention [3], we also call the dataset R8. Note that there is also a R10 dataset, and the only difference between R10 and R8 is that the classes *corn* and *wheat*, which are closely related to the class *grain*, were removed. Table 1 shows the distribution of documents per class and the split into training and test data for the R8 subset.

Table 1. Class distribution of training and testing data for R8.

Class	No. of Training Docs	No. of Test Docs	Total
Acq	1596	696	2292
Earn	2840	1083	3923
Grain	41	10	51
Interest	190	81	271
Money-fx	206	87	293
Ship	108	36	144
Trade	251	75	326
Crude	253	121	374
Total	5485	2189	7674

4 METHOD

We trained LDA models for each of the two datasets: the 4000 threads from FriendFeed and the R8 text data. LDA models have two parameters whose values need to be chosen experimentally: the number of topic clusters and the number of words in each cluster. We experimented with various parameter values of the LDA models. The number of cluster is particularly difficult to choose, since it reflects the level of generality of the extracted topics / concepts.

For the first dataset, the number of words in each cluster was set to maximum 15 (because for higher values, the weights of the words in the clusters became very small). For the number of topics, we chose several values: 10, 20, 40, 80, 160, and 320. Therefore we build 6 LDA models. We started with 10 topics because we have 10 classes, then we doubled the number of LDA topics at every model. Instead of choosing one of the models, we used all of them in order to represent each text at multiple levels of generalization at the same time. In this way, we let the classifiers choose the best features for the task.

In LDA models, polysemous words can be member of more than one topical cluster, while synonymous words are normally gathered in the same topics. An example of LDA topic cluster for the first model is: "Google", "email", "search", "work", "site", "services", "image", "click", "page", "create", "contact", "connect", "buzz", "Gmail", "mail". This could be labeled as *Internet*.

As mentioned, our 500 threads were manually annotated with the 10 generic classes. These classes, enumerated in section 3, are a manually generalized version of the top 50 LDA clusters into the 10 generic categories that proved to be sufficient during the manual annotation of the data. For the above example, the annotator placed it under the *technology* and *social_life* categories. The classification task is therefore multi-class, since a thread can be in more than one class. We trained binary classifiers for each class, and averaged the results over all classes.

For the second dataset, R8, we experimented with several parameter values for the number of clusters in the LDA models: 8, 16, 32, 64, 128, and 256 (thus we built 6 models). We chose 20 words in each cluster (because for higher values the weights were becoming too small). The reason we started with 8 clusters is that there are 8 classes in the annotated data. Then we doubled the number of topics several times. Similarly to the representation used for the first dataset, we combined all the models in the feature representation (the multi-level LDA-based representation), leaving up to the classifier to choose an appropriate level of generalization.

For the classification task on both datasets, we chose several classifiers from Weka [20], including Naive Bayes (NB) because it is fast and works well with text, SVM since it is known to obtain high performance on many tasks, and decision trees because we can manually inspect the learned tree.

We applied these classifiers on simple bag-of-words (BOW) representation, on LDA-based representations of different granularities, and on an integrated representation concatenating the BOW features and the LDA features. The values of the LDA-based features for each document are the weights of the clusters associated to the document by the LDA model (probability distributions).

5 Experiments and Results

The results on the first dataset are presented in Table 2. After stopword removal and stemming, the bag-of-words (BOW) representation contained

6573 words as features (TF-IDF values). The lower-dimensional representation based on LDA contained 630 features (10 + 20 + 40 + 80 + 160 + 320), whose values are the weights corresponding to the topic clusters. For the combined representation (BOW integrated with the LDA topics) the number of features was 7203 (6573+630).

The baseline of any classification experiment over this dataset may be considered as 18.8%, for a trivial classifier that puts everything in the most frequent class, *technology*.

On this dataset, due to its relatively small size, we conducted the classification evaluations using stratified 10-fold cross-validations (this means that the classifier is trained on nine parts of the data and tested on the remaining part, then this is repeated 10 times for different splits, and the results are averaged over the 10 folds). We performed several experiments on a range of classifiers and parameter values for each representation, to check the stability of a classifier's performance. We changed the *seed*, a randomization parameter of the 10-fold cross-validation, in order to avoid the accidental over-fitting. The values reported in Table 2 are the accuracies of the classification over all classes.

Table 2. Results on the FriendFeed dataset for different classifiers and representations, by cross-validation.

Representation / Classifier	Accuracy
Baseline	18.8%
BOW / SVM	72.22%
LDA Topics / SVM	75.13%
LDA+BOW / SVM	**80.40%**
BOW / NB	75.93%
LDA Topics / NB	74.63%
LDA+BOW / NB	77.39%
BOW / DT	69.33%
LDA Topics / DT	73.11%
LDA + BOW / DT	75.69%

The SVM classifier was the best for the task. The multi-level LDA-based representation achieved an accuracy of 75.13% compared to the BOW representation at 72.20%. Note that for the BOW representation, the best classifier was Naive Bayes, with an accuracy of 75.93%, but this is due to the use of a variant called complement Naive Bayes that compensates for data imbalance. For the combine LDA and BOW representation, SVM achieved the best accuracy of 80.40%. When using the low-dimensional LDA representation only, the accuracy goes down a bit, but still at the same level as BOW and it has the advantage that the classifiers are faster and other classifiers could be used (that do not usually run on high-dimensional data).

Table 3 presents detailed results for each class, for the best run (SVM classifier with LDA + BOW representation). We present the rate of true positives, the rate of false positives, the precision, recall and F-measure for each class. Since the accuracy results over all the classes is good, we wanted to see if the results are good for all classes, or if they vary by class. We can see that there are a few classes that seem to be more challenging for the classifier: *entertainment* and *lifestyle*. This could be due to these two classes being a bit ambiguous, with overlapping vocabulary and topics among the two of them or with the other classes. Perhaps *lifestyle* might be considered a bit too vague as a class label. The *technology* class is also on the low side; perhaps the vocabulary of this class overlaps with other classes too, since we are using a lot of technology for entertainment and other purposes.

Table 3. Results on the FriendFeed dataset for each class, for the SVM classifier (LDA+BOW representation), by cross-validation.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1.000	0.001	0.989	1.000	0.994	consumers
1.000	0.000	1.000	1.000	1.000	education
0.459	0.040	0.558	0.484	0.504	entertainment
0.993	0.027	0.797	0.989	0.891	incidents
0.419	0.045	0.479	0.469	0.439	lifestyle
1.000	0.001	0.989	1.000	0.994	politics
0.787	0.050	0.629	0.797	0.710	relationships
1.000	0.001	0.989	1.000	0.984	science
0.921	0.019	0.829	0.941	0.878	social_activities
0.553	0.026	0.678	0.593	0.606	technology

The results on the second dataset, R8, are shown in Table 4, for classifiers trained on the training parts of the data and tested on the test part. After stopword removal and stemming, the BOW representation (TF-IDF values) contained 17387 words as the feature space. We experimented with each LDA representation separately, without good results; therefore we chose the combined 6-level representation wit 504 features (8 + 16 +

32 + 64 + 128 + 256), corresponding to the LDA models with 8, 16, 32, 64, 128, and 256 clusters. For the integrated representation BOW with LDA topics we had 17891 features (the 504 LDA topics plus the 17387 word features).

The average classification accuracy is very high, compared to a baseline of 49.47% (of a simplistic 8-way classifier that always chooses the most frequent class, *earn* in this dataset). The SVM classifier achieved the best results. These values are in line with state-of-the art results reports in the literature. We can compare our results with other reported classification results of the same dataset. According to the best of our knowledge, the accuracy of our integrated representation method on the Reuters R8 dataset, 97%, is higher than any simple and combinatory representation method from related work, which reports accuracies of 88%–95% [21–23], while 96% was reached with SVM on a complex representation method based on kernel functions and Latent Semantic Indexing [24].

For our SVM classifier, the LDA-based representation achieved better accuracy (95.89%) than the BOW representation (93.33%). This is due to the multi-level representation. When we experimented with each level separately, the accuracies dropped considerably. The best results over all the experiments were for SVM with the combined BOW and LDA-based representation (97.03%), though the representation based only on LDA is not far behind and it has the advantage of lower dimensionality.

Table 4. Results on the R8 dataset, on the test data (2189 documents).

Representation / Classifier	Accuracy
Baseline	49.47%
BOW / SVM	93.33%
LDA Topics / SVM	95.89%
LDA+BOW / SVM	**97.03%**
BOW / NB	95.20%
LDA Topics / NB	94.61%
LDA+BOW / NB	95.52%
BOW / DT	91.54%
LDA Topics / DT	91.78%
LDA + BOW / DT	92.10%

For more complete experiments, as a second scenario on the R8 dataset, we also trained and tested the same set of classifiers using 10-fold cross-validation on the whole dataset, to check the stability of the results when training and testing sets are rotationally changed by stratified 10-fold cross-validation. The results are presented in Table 5 and they show a similar trend as the results from Table 4. The SVM classifier with LDA + BOW representation achieved the best accuracy, while the representation based only on the multi-level LDA is not far behind and it is better than the BOW representation.

In Table 6 we show detailed results for each class, for the best classifier, SVM with the combined feature representation, for the cross-validation setting. We can see that performance is very good for all classes, with one exception for the class *grain*. This is probably due to the lower number of instances of this class in the training and test data compared to the other classes.

Table 5. Results on the R8 dataset, by cross-validation on the whole data.

Representation / Classifier	Accuracy
Baseline	51.00%
BOW / SVM	94.67%
LDA Topics / SVM	95.89%
LDA+BOW / SVM	**97.29%**
BOW / NB	94.91%
LDA Topics / NB	92.57%
LDA+BOW / NB	94.59%
BOW / DT	90.40%
LDA Topics / DT	91.73%
LDA + BOW / DT	91.88%

6 Conclusions and Future Work

As our experimental results show, we can achieve good classification results by using a low-dimensional representation based on the multi-level LDA. This representation has the advantage that allows the use of classifiers or clustering algorithms that cannot run on high-dimensional feature spaces. By using a multi-level representation (different generalization levels) we achieved better results than the BOW representation on both datasets, with the SVM classifier.

The combined BOW and LDA features representation achieved the best classification performance, and it can be used when there memory

Table 6. Results on the R8 dataset for each class, for the SVM classifier (LDA+BOW representation), by cross-validation.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.982	0.020	0.954	0.982	0.968	acq
0.988	0.011	0.990	0.988	0.989	earn
0.794	0.000	0.986	0.794	0.880	grain
0.914	0.002	0.921	0.914	0.917	interest
0.884	0.002	0.931	0.884	0.907	money-fx
0.878	0.001	0.950	0.878	0.913	ship
0.945	0.003	0.931	0.945	0.938	trade
0.912	0.001	0.977	0.912	0.943	crude

is not a concern, for classifiers that are able to cope with the large vector spaces. Even in this case, the training and test times can be reduced by using only the LDA based representation.

Our results show that the first dataset is more difficult to classify than the second dataset. The reason is that it consists of social media texts, which are very noisy. In future work, we plan to experiment with more training data for the FriendFeed dataset (automatically annotated via the mapping of LDA clusters into the 10 classes), and to design new representation and classification methods that are more appropriate for this kind of data.

ACKNOWLEDGMENTS The authors wish to thank Ontario Centres of Excellence and to The Natural Sciences and Engineering Research Council of Canada for the financial support. We thank Lana Bogouslavski for annotating the FriendFeed data and Dmitry Brusilovsky for his insights in the experiments related to the classification of social media texts.

REFERENCES

1. Razavi, A.H., Inkpen, D., Brusilovsky, D., Bogouslavski, L.: General topic annotation in social networks: A Latent Dirichlet Allocation approach. In: Canadian Conference on Artificial Intelligence. (2013) 293–300
2. Pan, X., Assal, H.: Providing context for free text interpretation. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering. (2003) 704–709
3. Sebastiani, F.: Classification of text, automatic. In Brown, K., ed.: The Encyclopedia of Language and Linguistics, Volume 14, 2nd Edition. Elsevier Science Publishers, Amsterdam, NL (2006) 457–462

4. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research **37**(1) (2010) 141–188
5. Spark Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **1** (1972) 11–21
6. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems, MIT Press (2004) 2003
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Volume 101, Suppl. 1 of Proceedings of the National Academy of Sciences of the United States of America. National Academy of Sciences (2004)
8. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. UAI'02, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2002) 352–359
9. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06, New York, NY, USA, ACM (2006) 424–433
10. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. IJCAI'05, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2005) 786–791
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
12. Blei, D.M., Lafferty, J.D.: Topic models. Text mining: Classification, clustering, and applications **10** (2009) 71
13. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning. ICML'06, New York, NY, USA, ACM (2006) 577–584
14. Mimno, D.M., Hoffman, M.D., Blei, D.M.: Sparse stochastic inference for Latent Dirichlet Allocation. In: Proceedings of the 29th International Conference on Machine Learning (ICML 2012). (2012)
15. Jojo, S.m., Wong, M., Dras, M.J.: Topic modeling for native language identification. In: Proceedings of the Australasian Language Technology Association Workshop, Canberra, Australia (2011) 115–124
16. Chrupala, G.: Efficient induction of probabilistic word classes with LDA. In: Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011). (2011) 363–372
17. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. EMNLP'10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 56–65
18. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language

Processing: Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 248–256
19. Smet, W.D., Tang, J., Moens, M.F.: Knowledge transfer across multilingual corpora via latent topics. In: Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I. PAKDD'11, Berlin, Heidelberg, Springer-Verlag (2011) 549–560
20. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
21. Chen, Y.L., Yu, T.L.: News classification based on experts' work knowledge. In: Proceedings of the 2nd International Conference on Networking and Information Technology (IPCSIT 2011), vol.17, Singapore, IACSIT Press (2011)
22. Aggarwal, C.C., Zhao, P.: Towards graphical models for text processing. Knowledge Information Systems **36**(1) (2013) 1–21
23. Cardoso-Cachopo, A., Oliveira, A.L.: Combining LSI with other classifiers to improve accuracy of single-label text categorization. In: Proceedings of the First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning-EWLSATEL. (2007)
24. Yuan, M., Ouyang, Y., Xiong, Z.: A text categorization method using extended vector space model by frequent term sets. Journal of Information Science and Engineering **29**(1) (2013) 99–114

DIANA INKPEN
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
UNIVERSITY OF OTTAWA,
OTTAWA, ON, K1N 6N5, CANADA
E-MAIL: <DIANA.INKPEN@UOTTAWA.CA>

AMIR H. RAZAVI
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
UNIVERSITY OF OTTAWA,
OTTAWA, ON, K1N 6N5, CANADA
E-MAIL: <ARAZA082@UOTTAWA.CA>