Nonparametric Bayesian Topic Modelling with Auxiliary Data

Kar Wai Lim

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University

August 2016

Abstract

The intent of this dissertation in computer science is to study topic models for text analytics. The first objective of this dissertation is to incorporate auxiliary information present in text corpora to improve topic modelling for natural language processing (NLP) applications. The second objective of this dissertation is to extend existing topic models to employ state-of-the-art nonparametric Bayesian techniques for better modelling of text data. In particular, this dissertation focusses on:

- incorporating hashtags, mentions, emoticons, and target-opinion dependency present in tweets, together with an external sentiment lexicon, to perform opinion mining or sentiment analysis on products and services;
- leveraging abstracts, titles, authors, keywords, categorical labels, and the citation network to perform bibliographic analysis on research publications, using a supervised or semi-supervised topic model; and
- employing the hierarchical Pitman-Yor process (HPYP) and the Gaussian process (GP) to jointly model text, hashtags, authors, and the follower network in tweets for corpora exploration and summarisation.

In addition, we provide a framework for implementing arbitrary HPYP topic models to ease the development of our proposed topic models, made possible by modularising the Pitman-Yor processes. Through extensive experiments and qualitative assessment, we find that topic models fit better to the data as we utilise more auxiliary information and by employing the Bayesian nonparametric method.

Bibliography

Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In Grünwald, P. and Spirtes, P., editors, *Proceedings of the Twenty-sixth Conference on Uncertainty in Artificial Intelligence*, UAI 2010, pages 20–29. Corvallis, Oregon, USA. Association for Uncertainty in Artificial Intelligence Press.

Aletras, N. and Stevenson, M. (2014). Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2014, pages 631–636. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

AlSumait, L., Barbará, D., and Domeniconi, C. (2008). On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In Kellenberger, P., editor, *Proceedings of the Eighth IEEE International Conference on Data Mining*, ICDM 2008, pages 3–12. Piscataway, New Jersey, USA. Institute of Electrical and Electronics Engineers.

Aula, P. (2010). Social media, reputation risk and ambient publicity management. *Strategy & Leadership*, 38(6):43–49.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC 2010, pages 2200–2204. Paris, France. European Language Resources Association.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt *[sic]* social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, IJCNLP 2013, pages 356–364. Nagoya, Japan. Asian Federation of Natural Language Processing.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Chichester, England.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, volume 1. Springer-Verlag, Secaucus, New Jersey, USA.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese Restaurant Process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7:1–7:30.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR 2003, pages 127–134. New York City, New York, USA. Association for Computing Machinery.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In Cohen, W. W. and Moore, A., editors, *Proceedings of the 23rd International Conference on Machine Learning*, ICML 2006, pages 113–120. New York City, New York, USA. Association for Computing Machinery.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Broersma, M. and Graham, T. (2012). Social media as beat. *Journalism Practice*, 6(3):403–419.

Bundschus, M., Yu, S., Tresp, V., Rettinger, A., Dejori, M., and Kriegel, H.-P. (2009). Hierarchical Bayesian models for collaborative tagging systems. In Wang, W., Kargupta, H., Ranka, S., Yu, P. S., and Wu, X., editors, *Proceedings of the Ninth IEEE International Conference on Data Mining*, ICDM 2009, pages 728–733. Piscataway, New Jersey, USA. Institute of Electrical and Electronics Engineers.

Buntine, W. L. (2002). Variational extensions to EM and multinomial PCA. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Proceedings of the 13th European Conference on Machine Learning*, ECML 2002, pages 23–34. Berlin, Heidelberg. Springer.

Buntine, W. L. and Hutter, M. (2012). A Bayesian view of the Poisson-Dirichlet process. *ArXiv e-prints arXiv:1007.0296v2*.

Buntine, W. L. and Mishra, S. (2014). Experiments with non-parametric topic models. In Macskassy, S. A., Perlich, C., Leskovec, J., Wang, W., and Ghani, R., editors, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2014, pages 881–890. New York City, New York, USA. Association for Computing Machinery.

Cano Basave, A. E., He, Y., and Xu, R. (2014). Automatic labelling of topic models learned from Twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2014, pages 618–624. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Carpenter, B. (2004). Phrasal queries with LingPipe and Lucene: Ad hoc genomics text retrieval. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of the Thirteenth Text Retrieval Conference*, TREC 2004. Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

Chen, C., Du, L., and Buntine, W. L. (2011). Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECML 2011, pages 296–311. Berlin, Heidelberg. Springer-Verlag.

Çinlar, E. (2011). *Probability and Stochastics*, volume 261. Springer Science & Business Media, New York City, New York, USA.

Correa, T., Hinsley, A. W., and de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2):247–253.

Daumé III, H. (2007). HBC: Hierarchical Bayes Compiler. University of Maryland, USA.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using Twitter hashtags and smileys. In Huang, C. and Jurafsky, D., editors, *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING 2010, pages 241–249. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

De Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC 2006, pages 449–454. Paris, France. European Language Resources Association.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In Najork, M., Broder, A. Z., and Chakrabarti, S., editors, *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM 2008, pages 231–240. New York City, New York, USA. Association for Computing Machinery.

Du, L. (2012). *Non-parametric Bayesian methods for structured topic models*. PhD thesis, The Australian National University, Canberra, Australia.

Du, L., Buntine, W. L., and Jin, H. (2010). A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning*, 81(1):5–19.

Du, L., Buntine, W. L., and Jin, H. (2012a). Modelling sequential text with an adaptive topic model. In Tsujii, J., Henderson, J., and Pasca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2012, pages 535–545. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Du, L., Buntine, W. L., Jin, H., and Chen, C. (2012b). Sequential latent Dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503.

Du, L., Buntine, W. L., and Johnson, M. (2013). Topic segmentation with a structured topic model. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2013, pages 190–200. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Eisenstein, J. (2013). What to do about bad language on the internet. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2013, pages 359–369. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Erosheva, E. A. and Fienberg, S. E. (2005). *Bayesian Mixed Membership Models for Soft Clustering and Classification*, pages 11–26. Springer, Berlin, Heidelberg.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR 2005, pages 524–531. Piscataway, New Jersey, USA. Institute of Electrical and Electronics Engineers.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from Google's image search. In Sebe, N., Lew, M. S., and Huang, T. S., editors, *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 2 of *ICCV 2005*, pages 1816–1823. Piscataway, New Jersey, USA. Institute of Electrical and Electronics Engineers.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton, Florida, USA.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University, California, USA.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2005). Interpolating between types and tokens by estimating power-law generators. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, NIPS 2005, pages 459–466. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2335–2382.

Green, P. J. and Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3):1391–1403.

Gruber, A., Weiss, Y., and Rosen-Zvi, M. (2007). Hidden topic Markov models. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, AISTATS 2007, pages 163–170. Brookline, Massachusetts, USA. Microtome Publishing.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28:55–61.

Han, B., Cook, P., and Baldwin, T. (2012). Automatically constructing a normalisation dictionary for microblogs. In Tsujii, J., Henderson, J., and Pasca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2012, pages 421–432. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Han, B., Cook, P., and Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):5:1–5:27.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Chen, H., Wactlar, H. D., Chen, C., Lim, E., and Christel, M. G., editors, *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL 2004, pages 296–305. New York City, New York, USA. Association for Computing Machinery.

Han, H., Zha, H., and Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In Marlino, M., Sumner, T., and III, F. M. S., editors, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL 2005, pages 334–343. New York City, New York, USA. Association for Computing Machinery.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., and Giles, C. L. (2009). Detecting topic evolution in scientific literature: How can citations help? In Cheung, D. W., Song, I., Chu, W. W., Hu, X., and Lin, J. J., editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM 2009, pages 957–966. New York City, New York, USA. Association for Computing Machinery.

He, Y. (2012). Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2):4:1–4:19.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *ArXiv e-prints 1309.6835v1*.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*, volume 28. Cambridge University Press, Cambridge, England.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1999, pages 50–57. New York City, New York, USA. Association for Computing Machinery.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA 2010, pages 80–88. New York City, New York, USA. Association for Computing Machinery.

Hospedales, T., Gong, S., and Xiang, T. (2012). Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision*, 98(3):303–323.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In McGuinness, D. L. and Ferguson, G., editors, *Proceedings of the Nineteenth National Conference on Artifial Intelligence*, AAAI 2004, pages 755–760. Palo Alto, California, USA. Association for the Advancement of Artificial Intelligence Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jagarlamudi, J., Daumé III, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. In Bouma, G. and Parmentier, Y., editors, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, pages 204–213. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

Jaynes, E. T. and Kempthorne, O. (1976). Confidence intervals vs Bayesian intervals. In Harper, W. L. and Hooker, C. A., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, pages 175–257. Springer, Dordrecht, Netherlands.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent Twitter sentiment classification. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ACL-HLT 2011, pages 151–160. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., and Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In Macdonald, C., Ounis, I., and Ruthven, I., editors, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM 2011, pages 775–784. New York City, New York, USA. Association for Computing Machinery.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In King, I., Nejdl, W., and Li, H., editors, *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM 2011, pages 815–824. New York City, New York, USA. Association for Computing Machinery.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, NIPS 2007, pages 641–648. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

Jurafsky, D. and Martin, J. H. (2000). *Speech & Language Processing*. Prentice-Hall, Upper Saddle River, New Jersey, USA.

Karimi, S., Yin, J., and Paris, C. (2013). Classifying microblogs for disasters. In Culpepper, S., Zuccon, G., and Sitbon, L., editors, *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS 2013, pages 26–33. New York City, New York, USA. Association for Computing Machinery.

Kataria, S., Mitra, P., Caragea, C., and Giles, C. L. (2011). Context sensitive topic models for author influence in document networks. In Walsh, T., editor, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three*, IJCAI 2011, pages 2274–2280. Palo Alto, California, USA. Association for the Advancement of Artificial Intelligence Press.

Kim, D., Kim, S., and Oh, A. (2012). Dirichlet process with mixed random measures: A nonparametric topic model for labeled data. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning*, ICML 2012, pages 727–734. New York City, New York, USA. Omnipress.

Kinsella, S., Murdock, V., and O'Hare, N. (2011). "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In Cantador, I., Carrero, F. M., Cortizo, J. C., Rosso, P., Schedl, M., and Troyano, J. A., editors, *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC 2011, pages 61–68. New York City, New York, USA. Association for Computing Machinery.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide*

Web, WWW 2010, pages 591–600. New York City, New York, USA. Association for Computing Machinery.

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, Mahwah, New Jersey, USA.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ACL-HLT 2011, pages 1536–1545. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Li, A. Q., Ahmed, A., Ravi, S., and Smola, A. J. (2014). Reducing the sampling complexity of topic models. In Macskassy, S. A., Perlich, C., Leskovec, J., Wang, W., and Ghani, R., editors, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2014, pages 891–900. New York City, New York, USA. Association for Computing Machinery.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., and Yu, H. (2010). Structure-aware review mining and summarization. In Huang, C. and Jurafsky, D., editors, *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING 2010. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Li, T., Zhang, Y., and Sindhwani, V. (2009). A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In Su, K., Su, J., and Wiebe, J., editors, *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*, ACL 2009, pages 244–252. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Lim, K. W. and Buntine, W. L. (2014a). Bibliographic analysis with the Citation Network Topic Model. In Phung, D. and Li, H., editors, *Proceedings of the Sixth Asian Conference on Machine Learning*, ACML 2014, pages 142–158. Brookline, Massachusetts, USA. Microtome Publishing.

Lim, K. W. and Buntine, W. L. (2014b). Twitter Opinion Topic Model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In Li, J., Wang, X. S., Garofalakis, M. N., Soboroff, I., Suel, T., and Wang, M., editors, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM 2014, pages 1319–1328. New York City, New York, USA. Association for Computing Machinery.

Lim, K. W. and Buntine, W. L. (2016). Bibliographic analysis on research publications using authors, categorical labels and the citation network. *Machine Learning*, 103(2):185–213.

Lim, K. W., Buntine, W. L., Chen, C., and Du, L. (2016). Nonparametric Bayesian topic modelling with the hierarchical Pitman-Yor processes. *International Journal of Approximate Reasoning*, 1(1):1–40.

Lim, K. W., Chen, C., and Buntine, W. L. (2013). Twitter-Network Topic Model: A full Bayesian treatment for social network and text modeling. In *Advances in Neural Information Processing Systems: Topic Models Workshop*, NIPS Workshop 2013, pages 1–5. Lake Tahoe, Nevada, USA.

Lim, K. W., Sanner, S., and Guo, S. (2012). On the mathematical relationship between expected n-call@k and the relevance vs. diversity trade-off. In Hersh, W. R., Callan, J., Maarek, Y., and Sanderson, M., editors, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2012, pages 1117–1118. New York City, New York, USA. Association for Computing Machinery.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In Cheung, D. W., Song, I., Chu, W. W., Hu, X., and Lin, J. J., editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM 2009, pages 375–384. New York City, New York, USA. Association for Computing Machinery.

Lindsey, R. V., Headden III, W. P., and Stipicevic, M. J. (2012). A phrase-discovering topic model using hierarchical Pitman-Yor processes. In Tsujii, J., Henderson, J., and Pasca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2012, pages 214–222. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

Liu, L., Tang, J., Han, J., Jiang, M., and Yang, S. (2010). Mining topic-level influence in heterogeneous networks. In Huang, J., Koudas, N., Jones, G. J. F., Wu, X., Collins-Thompson, K., and An, A., editors, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM 2010, pages 199–208. New York City, New York, USA. Association for Computing Machinery.

Liu, S., Li, F., Li, F., Cheng, X., and Shen, H. (2013). Adaptive co-training SVM for sentiment classification on tweets. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R., editors, *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM 2013, pages 2079–2088. New York City, New York, USA. Association for Computing Machinery.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community. In Danyluk, A. P., Bottou, L., and Littman, M. L., editors, *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 665–672. New York City, New York, USA. Association for Computing Machinery.

Lloret, E. and Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review*, 37(1):1–41.

Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, NIPS 2012, pages 998–1006. Curran Associates, Rostrevor, Northern Ireland.

Low, A. A. (1991). *Introductory Computer Vision and Image Processing*. McGraw-Hill, New York City, New York, USA.

Lu, J., Plataniotis, K. N., and Venetsanopoulos, A. N. (2003). Face recognition using LDA-based algorithms. *IEEE Transactions on Neural Networks*, 14(1):195–200.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL 2012, pages 25–30. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Mai, L. C. (2010). Introduction to image processing and computer vision. Technical report, Institute of Information Technology, Hanoi, Vietnam.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York City, New York, USA.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

Mao, X.-L., Ming, Z.-Y., Zha, Z.-J., Chua, T.-S., Yan, H., and Li, X. (2012). Automatic labeling hierarchical topics. In Chen, X., Lebanon, G., Wang, H., and Zaki, M. J., editors, *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM 2012, pages 2383–2386. New York City, New York, USA. Association for Computing Machinery.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. In Melero, M., editor, *Proceedings of @NLP can u tag #user_generated_content*, LREC Workshop 2012, pages 15–22. Istanbul, Turkey.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Rostrevor, Northern Ireland.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Cornell University, New York, USA.

McCord, M. and Chuah, M. (2011). Spam detection on Twitter using traditional classifiers. In *Proceedings of the 8th International Conference on Autonomic and Trusted Computing*, ATC 2011, pages 175–186. Berlin, Heidelberg. Springer-Verlag.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

Mehdad, Y., Carenini, G., Ng, R. T., and Joty, S. R. (2013). Towards topic labeling with phrase entailment and aggregation. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2013, pages 179–189. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Mehrotra, R., Sanner, S., Buntine, W. L., and Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In Jones, G. J. F., Sheridan, P., Kelly, D., de Rijke, M., and Sakai, T., editors, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2013, pages 889–892. New York City, New York, USA. Association for Computing Machinery.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW 2007, pages 171–180. New York City, New York, USA. Association for Computing Machinery.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mimno, D. and McCallum, A. (2007). Mining a digital library for influential authors. In Rasmussen, E. M., Larson, R. R., Toms, E. G., and Sugimoto, S., editors, *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL 2007, pages 105–106. New York City, New York, USA. Association for Computing Machinery.

Minka, T. P. (2001). Expectation Propagation for approximate Bayesian inference. In Breese, J. and Koller, D., editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI 2001, pages 362–369. San Francisco, California, USA. Morgan Kaufmann.

Minka, T. P., Winn, J. M., Guiver, J. P., Webster, S., Zaykov, Y., Yangel, B., Spengler, A., and Bronskill, J. (2014). Infer.NET 2.6. Microsoft Research, Cambridge, UK.

Moghaddam, S. and Ester, M. (2010). Opinion Digger: An unsupervised opinion miner from unstructured product reviews. In Huang, J., Koudas, N., Jones, G. J. F., Wu, X., Collins-Thompson, K., and An, A., editors, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM 2010, pages 1825–1828. New York City, New York, USA. Association for Computing Machinery.

Moghaddam, S. and Ester, M. (2011). ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In Ma, W., Nie, J., Baeza-Yates, R. A., Chua, T., and Croft, W. B., editors, *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2011, pages 665–674. New York City, New York, USA. Association for Computing Machinery.

Moghaddam, S. and Ester, M. (2012). On the design of LDA models for aspect-based opinion mining. In Chen, X., Lebanon, G., Wang, H., and Zaki, M. J., editors, *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM 2012, pages 803–812. New York City, New York, USA. Association for Computing Machinery.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. In Teh, Y. W. and Titterington, D. M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, AISTATS 2010, pages 541–548. Brookline, Massachusetts, USA. Microtome Publishing.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the Second*

Joint Conference on Lexical and Computational Semantics, Volume 2: Seventh International Workshop on Semantic Evaluation, SemEval 2013, pages 312–320. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In Li, Y., Liu, B., and Sarawagi, S., editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2008, pages 542–550. New York City, New York, USA. Association for Computing Machinery.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–741.

Newman, D., Karimi, S., and Cavedon, L. (2009). External evaluation of topic models. In Kay, J., Thomas, P., and Trotman, A., editors, *Proceedings of the 14th Australasian Document Computing Symposium*, ADCS 2009, pages 11–18. NSW, Australia. University of Sydney.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

Oldham, K. B., Myland, J., and Spanier, J. (2009). *An Atlas of Functions: With Equator, the Atlas Function Calculator*. Springer Science and Business Media, New York City, New York, USA.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2013, pages 380–390. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC 2010, pages 1320–1326. Paris, France. European Language Resources Association.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Springer-Verlag, Berlin Heidelberg.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, DSC 2003. Vienna, Austria.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP 2005, pages 339–346. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Qiu, M., Zhu, F., and Jiang, J. (2013). It is not just what we say, but how we say them: LDA-based behavior-topic model. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, SDM 2013, pages 794–802. Philadelphia, Pennsylvania, USA. Society for Industrial and Applied Mathematics.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Upper Saddle River, New Jersey, USA.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). 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 2009, pages 248–256. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1524–1534. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In Halpern, J. and Meek, C., editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI 2004, pages 487–494. Arlington, Virginia, USA. Association for Uncertainty in Artificial Intelligence Press.

Sato, I. and Nakagawa, H. (2010). Topic models with power-law using Pitman-Yor process. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2010, pages 673–682. New York City, New York, USA. Association for Computing Machinery.

Schnober, C. and Gurevych, I. (2015). Combining topic models for corpus exploration: Applying LDA for complex corpus research tasks in a digital humanities project. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, TM 2015, pages 11–20. New York City, New York, USA. Association for Computing Machinery.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, 29(3):93–106.

Sethuraman, J. (1991). A constructive definition of Dirichlet priors. Technical report, Florida State University, USA.

Si, X. and Sun, M. (2009). Tag-LDA for scalable real-time tag recommendation. *Journal of Information and Computational Science*, 6(2):1009–1016.

Suominen, H., Hanlen, L., and Paris, C. (2014). Twitter for health – building a social media search engine to better understand and curate laypersons' personal experiences. In Neustein, A., editor, *Text Mining of Web-based Medical Content*, chapter 6, pages 133–174. De Gruyter, Berlin, Germany.

Suominen, H., Pyysalo, S., Hiissa, M., Ginter, F., Liu, S., Marghescu, D., Pahikkala, T., Back, B., Karsten, H., and Salakoski, T. (2008). Performance evaluation measures for text mining. *Handbook of Research on TextWeb Mining Technologies*, pages 724–747.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In IV, J. F. E., Fogelman-Soulié, F., Flach, P. A., and Zaki, M. J., editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2009, pages 807–816. New York City, New York, USA. Association for Computing Machinery.

Teh, Y. W. (2006a). A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, National University of Singapore.

Teh, Y. W. (2006b). A hierarchical Bayesian language model based on Pitman-Yor processes. In Calzolari, N., Cardie, C., and Isabelle, P., editors, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, COLING-ACL 2006, pages 985–992. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics: Principles and Practice*, chapter 5. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Titov, I. and McDonald, R. T. (2008a). A joint model of text and aspect ratings for sentiment summarization. In McKeown, K., Moore, J. D., Teufel, S., Allan, J., and Furui, S., editors, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL 2008, pages 308–316. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Titov, I. and McDonald, R. T. (2008b). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW 2008, pages 111–120. New York City, New York, USA. Association for Computing Machinery.

Tsai, F. S. (2011). A tag-topic model for blog mining. *Expert Systems with Applications*, 38(5):5330–5335.

Tsur, O., Davidov, D., and Rappoport, A. (2010). ICWSM-A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In Cohen, W. W. and Gosling, S., editors, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM 2010, pages 162–169. Palo Alto, California, USA. Association for the Advancement of Artificial Intelligence Press.

Tu, Y., Johri, N., Roth, D., and Hockenmaier, J. (2010). Citation author topic model in expert search. In Huang, C. and Jurafsky, D., editors, *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING 2010, pages 1265–1273. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Walck, C. (2007). Handbook on statistical distributions for experimentalists. Technical Report SUF-PFY/96-01, University of Stockholm, Sweden.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In Cohen, W. W. and Moore, A., editors, *Proceedings of the 23rd International Conference on Machine Learning*, ICML 2006, pages 977–984. New York City, New York, USA. Association for Computing Machinery.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009a). Rethinking LDA: Why priors matter. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, NIPS 2009, pages 1973–1981. Curran Associates, Rostrevor, Northern Ireland.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). Evaluation methods for topic models. In Danyluk, A. P., Bottou, L., and Littman, M. L., editors, *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 1105–1112. New York City, New York, USA. Association for Computing Machinery.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous time dynamic topic models. In McAllester, D. and Myllymaki, P., editors, *Proceedings of the Twenty-Fourth Conference Conference on Uncertainty in Artificial Intelligence*, UAI 2008, pages 579–586. Corvallis, Oregon, USA. Association for Uncertainty in Artificial Intelligence Press.

Wang, C. and Blei, D. M. (2012). A split-merge MCMC algorithm for the hierarchical Dirichlet process. *ArXiv e-prints 1201.1657v1*.

Wang, H., Zhang, D., and Zhai, C. (2011a). Structural topic model for latent topical structure analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ACL-HLT 2011, pages 1526–1535. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Wang, X. and McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In Grossman, R., Bayardo, R. J., and Bennett, K. P., editors, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2006, pages 424–433. New York City, New York, USA. Association for Computing Machinery.

Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Kawada, S., editor, *Proceedings of the Seventh IEEE International Conference on Data Mining*, ICDM 2007, pages 697–702. Piscataway, New Jersey, USA. Institute of Electrical and Electronics Engineers.

Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. (2011b). Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In Macdonald, C., Ounis, I., and Ruthven, I., editors, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM 2011, pages 1031–1040. New York City, New York, USA. Association for Computing Machinery.

Wei, X. and Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In Efthimiadis, E. N., Dumais, S. T., Hawking, D., and Järvelin, K., editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2006, pages 178–185. New York City, New York, USA. Association for Computing Machinery.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). TwitterRank: Finding topic-sensitive influential Twitterers. In Davison, B. D., Suel, T., Craswell, N., and Liu, B., editors, *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM 2010, pages 261–270. New York City, New York, USA. Association for Computing Machinery.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP 2005, pages 347–354. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Wood, F. and Teh, Y. W. (2009). A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In van Dyk, D. A. and Welling, M., editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, AISTATS 2009, pages 607–614. Brookline, Massachusetts, USA. Microtome Publishing.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In Schwabe, D., Almeida, V. A. F., Glaser, H., Baeza-Yates, R. A., and Moon, S. B., editors, *Proceedings of the 22nd International Conference on World Wide Web*, WWW 2013, pages 1445–1456. Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In King, I., Nejdl, W., and Li, H., editors, *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM 2011, pages 177–186. New York City, New York, USA. Association for Computing Machinery.

Zhao, W. X. and Jiang, J. (2011). An empirical comparison of topics in Twitter and traditional media. Technical report, Singapore Management University, Singapore.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing Twitter and traditional media using topic models. In Clough, P. D., Foley, C., Gurrin, C., Jones, G. J. F., Kraaij, W., Lee, H., and Murdock, V., editors, *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR 2011, pages 338–349. Berlin, Heidelberg. Springer-Verlag.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2010, pages 56–65. Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

Zheng, B., McLean, D. C., and Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics*, 7(1):1–10.

Zhu, X., Blei, D., and Lafferty, J. (2006). TagLDA: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, USA.

Zhu, Y., Yan, X., Getoor, L., and Moore, C. (2013). Scalable text and link analysis with mixed-topic link models. In Dhillon, I. S., Koren, Y., Ghani, R., Senator, T. E., Bradley, P., Parekh, R., He, J., Grossman, R. L., and Uthurusamy, R., editors, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2013, pages 473–481. New York City, New York, USA. Association for Computing Machinery.