Classifying Hotel Topics and Opinions from Online Guest Review Comments

By:

N. Nobel

Master Thesis

VU University Amsterdam

March, 2013

Master Thesis Business Analytics

Classifying Hotel Topics and Opinions from Online Guest Review Comments

Author:

Nicolaas Nobel

Supervisors:

Mark Hoogendoorn

Piek Vossen

Sandjai Bhulai

VU University Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

Booking.com B.V.
Herengracht 597
1017 CE Amsterdam

March, 2013

PREFACE

The final part of the Master Business Analytics consists of a six months internship at an organisation by choice. During this internship students are required to use their theoretical knowledge of business mathematics and informatics to support the organization with data-driven decision making, optimisation of business processes and data analysis. The deliverables generally consists of a final thesis along with a data analysis, program or simulation. The research for and creation of this thesis took place at the content department of Booking.com from September 2012 until March 2013. During this period a text mining tool is developed and a business focused data analysis is performed.

In the first and foremost place I would like to thank Vladimir Sterngold for his on-going guidance and support, giving me insight into the hotel business and providing me detailed information about the Booking.com website. In the second place I would like to acknowledge my special gratitude to my supervisors from the VU University, Mark Hoogendoorn and Piek Vossen, giving me direction and constructive comments on my thesis. In the third place I would like to thank Sandjai Bhulai, who accepted to take on the task of second reader. In the fourth place, I would like to use the opportunity to thank all the guest review associates that helped me develop a high quality labelled dataset for this project. Finally, I would like to express my grateful appreciations to my colleagues from the business analytics team, content and IT department for their collegiality and discussions that helped me write this thesis.

Nicolaas Nobel

Amsterdam, March 21, 2013

ABSTRACT

[Classified]

Contents

Bibliography

1. *http://www.booking.com.* [Online] Booking.com B.V. [Cited: 08 02 2013.] http://www.booking.com/general.en-gb.html?aid=303948&dcid=1&label=bookings-naam-3jHr6EsDyn1Uq9ttdGmKKwS17149403061%3Apl%3Ata%3Ap1%3Ap2260%2C000%3Aac%3Aap1t1%3Aneg&lang=en-gb&sid=64fd0b6cb562f018c6cc9132ec9cd39f&tmpl=docs%2Fabout.

2. Wiki Booking.com (Internal Document). *www.wiki.booking.com.* [Online] Booking.com. [Cited: 08 02 2013.] https://wiki.booking.com/display/HD01/Organizing+New+Hires+training+in+your+office.

3. *http://www.booking.com.* [Online] Booking.com B.V. [Cited: 08 02 2013.] http://www.booking.com/general.nl.html?tmpl=docs/career_opportunities#cntnt_dtr .

4. *Spam Filtering for Optimization in Internet Promotions using Bayesian Analysis.* **Smeureanu, I. and Zurini, M.** s.l. : Journal of Applied Quantitative, 2010, pp. 198-211.

5. *Applying Supervised Opinion Mining Techniques on Online User Reviews.* **Smeureanu, I. and Bucur, C.** 2012, Informatica Economică, pp. 81-91.

6. **Liu, B.** *Web Data Mining - Exploring Hyperlinks, Contents and Usage.* Chicago : Springer Heidelberg Dordrecht London New York, 2007.

7. *Automatic Text Classification: A Technical Review.* **Dalal, M. K. and Zaveri, M. A.** 2011, International Journal of Computer Applications, pp. 37-40.

8. *Opinion Mining Classification Using Key Word Summarization Based on Singular Value Decomposition.* **Valarmathi, B. and Palanisamy, V.** 2011, International Journal on Computer Science and Engineering, pp. 212-215.

9. *Opinion Mining and Sentiment Analysis.* **Pang, B. and Lee, L.** 2008, Foundations and Trends in Information Retrieval, pp. 1-135.

10. *Mining Opinion Features in Customer Reviews.* **Hu, M. and Liu, B.** 2004, American Association for Artificial Intelligence.

11. *A Survey on Sentiment Analysis of (Product) Reviews.* **Jebaseeli, A. N. and Kirubakaran, E.** 2012, International Journal of Computer Applications, pp. 36-39.

12. *Finding Opinion Strength Using Fuzzy Logic on Web Reviews.* **Kar, A. and Mandal, D.P.** 2011, International Journal of Engineering and Industries, pp. 37-43.

13. *Rule Based System for Enhancing Recall for Feature Mining from Short Sentences in Customer Review Documents.* **Ahmad, T. and Doja, M. N.** 2012, International Journal on Computer Science and Engineering, pp. 1211-1219.

14. *A Comparison of Feature and Semantic-Based Summarization Algorithms for Turkish.* **Güran, A., Bekar, E. and Akyokuş, S.** 2010, International Symposium on Innovations in Intelligent Systems and Applicaitons, pp. 371-375.

15. *Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis.* **Hamouda, A., Marei, M. and Rohaim, M.** 2011, JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, pp. 199-203.

16. *Classification of Opinion Mining Techniques.* **Mishra, N. and Jha, C.K.** 2012, International Journal of Computer Applications.

17. *Opinion Mining: Issues and Challenges (A survey).* **Seerat, B. and Azam, F.** 2012, International Journal of Computer Applications, pp. 42-51.

18. *Fuzzy Logic Based Method for Improving Text Summarization.* **Suanmali, L., Salim, N. and Binwahlan, M. S.** 2009, International Journal of Computer Science and Information Security.

19. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.* **Wilson, T., Wiebe, J. and Hoffmann, P.** 2005, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, pp. 347–354.

20. *Identifying Consumers' Arguments in Text.* **Schneider, J. and Wyner, A.** 2012.

21. *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.* **Pang, B. and Lee, L.** 2004, ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

22. *An Analysis of One-Star Online Reviews and Responses in the Washington, D.C., Lodging Market.* **Levy, S. E., Duan, W. and Boo, S.** 2012, Cornell Hospitality Quarterly 54(1).

23. *Text Classification Using Machine Learning Techniques.* **Ikonomakis, M., Kotsiantis, S. and Tampakas, V.** 2005, Wseas Transactions On Computers, pp. 966-974.

24. *Opinion Mining of M Learning Reviews using Soft Computing Techniques.* **Jebaseeli, A. N. and Kirubakaran, E.** 2012, International Journal of Computer Applications, pp. 44-48.

25. *An Efficient Text Classification Using KNN and Naive Bayesian.* **Sreemathy, J. et al.** 2012, International Journal on Computer Science and Engineering (IJCSE), pp. 392-396.

26. *Automatic Text Categorization by Unsupervised Learning.* **Ko, Y. and Seo, J.** 2000, COLING '00 Proceedings of the 18th conference on Computational linguistics.

27. *Advanced Natural Language Processing - Basic Text Process.* **X., Zhu.** 2010, pp. 1-3.

28. *Ranking System for Opinion Mining of Features.* **Ahmad, T. and Doja, M.N.** 2012, International Journal of Computer Science Issues, pp. 440-447.

29. **Hu, N., Zhang, J. and Pavlou, P. A.** Overcoming the J-shaped distribution of product reviews. 2009, Vol. 52, 10.

30. *communication, Why do online product reviews have a j-shaped distribution? overcoming biases in online word-of-mouth.* **Hu, N., Pavlou, P. A. and Zhang, J.** s.l. : Marketing Science, 2007.

31. **Moson, Marc.** *Wiki Booking.com.* [Online] [Cited: 08 02 2013.] https://wiki.booking.com/display/INT/Content+Request+4+-+Customer+Experience+Tab+in+Extranet+%28gettext%29.

32. *Worditout.* [Online] [Cited: 08 02 2013.] http://www.worditout.com/.

33. *IBM.com.* [Online] [Cited: 15 02 2013.] http://pic.dhe.ibm.com/infocenter/analytic/v2r1m0/index.jsp?topic=%2Fcom.ibm.discovery.es.ta.doc%2Fiiysalgstopwd.htm.

34. *Online reputation management is hot — but is it ethical?* **Hoffman, T.** s.l. : Computerworld, 2008.

35. *The Effect of Word of Mouth on Sales: Online Book Reviews.* **Chevalier, J. and Mayzlin, D.** 2006.

36. **Martin, Hugo.** *Los Angeles Times.* [Online] Los Angeles Times. http://articles.latimes.com/2012/dec/09/business/la-fi-mo-financial-impact-of-positive-online-reviews-20121207.

37. **Mayock, P.** *Measuring the impact of online reviews on rate.* [Online] http://www.hotelnewsnow.com/Articles.aspx/9269/Measuring-the-impact-of-online-reviews-on-rate.

38. *Deriving the Pricing Power of Product Features by Mining Consumer Reviews.* **Archak, N., Ghose, A. and G., Ipeirotis P.** 2011.

39. *Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics.* **Ghose, A. and Ipeirotis, P. G.** 2011.

40. **Freed, J. Q.** How to respond to hotel reviews. *hotelnewsnow.com.* [Online] hotelnewsnow. [Cited: 20 02 2013.] www.hotelnewsnow.com/Articles.aspx/5276/How-to-respond-to-hotel-reviews.

41. **Cooper, Caroline.** Embracing guest feedback leads to success. *hotelnewsnow.com.* [Online] hotelnewsnow, 16 08 2012. [Cited: 20 02 2013.] http://www.hotelnewsnow.com/Articles.aspx/8780/Embracing-guest-feedback-leads-to-success.

42. **Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome.** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2009.

43. *From Data Mining to Knowledge Discovery in Databases.* **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.** s.l. : American Association for Artificial Intelligence, 1996.

44. **Schapire, R.** Machine Learning Algorithms for Classification. [Online] 2006. [Cited: 20 02 2013.] http://dimacs.rutgers.edu/Workshops/DataMiningTutorial/slides/schapire.pdf.

45. **Sohn, J. P.** Introduction to Computational Linguistics Tokenization and Sentence Boundary Detection. *soehn.net.* [Online] 2007. http://www.soehn.net/work/icl/tokenize.pdf.

46. *Binary codes capable of correcting deletions, insertions, and reversals.* **Levenshtein.** s.l. : Soviet Physics Doklady, 1966.

47. *Damerau Levenshtein Distance.* [Online] [Cited: 12 03 2013.] http://software-and-algorithms.blogspot.co.uk/2012/09/damerau-levenshtein-edit-distance.html.

48. **Norouzi, M, Fleet, D. J. and Salakhutdinov, R.** *Hamming Distance Metric Learning.* s.l. : Departments of Computer Sciencey and Statistics University of Toronto.

49. **Salton, G. and McGill, M. J.** *Amazon.co.uk.* s.l. : McGraw-Hill Companies, 1983.

50. *An Introduction to Variable and Feature Selection. .* **Clopinet, I. G. and Elisseeff, A.** s.l. : The Journal of Machine Learning Research, 2003.

51. *A Comparative Study on Feature Selection in Text Categorization.* **Yang, Y. and Pedersen, J.O.** s.l. : Proceedings of the 14th ICML97 1997.

52. *An Experimental Study of Feature Selection Metrics for Text Categorization.* **Forman, G.** s.l. : Journal of Machine Learning Research, 2003.

53. *Understanding Inverse Document Frequency On theoretical arguments for IDF.* **Robertson, S.** s.l. : Journal of Documentation.

54. *A statistical interpretation of term specificity and its application in retrieval.* **Jones, K. S.** s.l. : Journal of Documentation, 1972.

55. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* **Joachims, T.**

56. *Document Classification with Support Vector Machines.* **Mertsalov, K. and McCreary, M.** 2009.

57. *Feature hashing for large scale multitask learning.* **Weinberger, K. Q., et al., et al.** s.l. : Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1113–1120.

58. *An Improved KNN Text Classification Algorithm Based on Clustering.* **Yong, Z., Youwen, L. and Shixiong, X.** s.l. : JOURNAL OF COMPUTERS, 2009.

59. *Techniques for Improving the Performance of Naive Bayes for Text Classification.* **Schneider, K. M.**

60. *Distribution of content words and phrases in text and language modelling.* **Katz, S.M.** s.l. : Natural Language Engineering, Vol. 2.

61. *On the optimality of the simple Bayesian classifier under zeroone loss.* **Domingos, P. and Pazzani, M.** s.l. : Machine Learning, 1997, pp. 103–130.

62. *On bias, variance, 0/1-loss, and the curse-of-dimensionality.* **Friedman, J.H.** s.l. : Data Mining and Knowledge Discovery, 1997, pp. 55–77.

63. *Opinion Miner: A Novel Machine Learning System for Web Opinion Mining and Extraction.* **Jin, W., Hay Ho, H. and Srihari, R.** Paris, France. : Proceeding of International conference on Knowledge Discovery and Data Mining, 2009.

64. *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?* **Manning, C. D.** s.l. : Departments of Linguistics and Computer Science.

65. SentiWordNet. *SentiWordNet.* [Online] [Cited: 20 02 2013.] http://sentiwordnet.isti.cnr.it/.

66. *Sentiment Analysis and Opinion Mining: A Survey.* **Vinodhini, G. and Chandrasekaran, R. M.** s.l. : International Journal of Advanced Research in Computer Science and Software Engineering, 2012.

67. *C4.5: Programs for Machine Learning.* **Quinlan, J. R.** s.l. : Morgan Kaufmann Publishers, 1993.

68. *An Introduction to Information Retrieval, page 182.* **Manning, C. D., Raghavan, P. and Schütze, H.:.** s.l. : Cambridge University Press, 2009.

69. *Reuters Test set.* [Online] http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection.

70. *An Empirical Comparison of Text Categorization Methods.* **Cardoso-Cachopo, A. and Oliveira, A. L.** s.l. : Instituto Superior Tecnico Departamento de Engenharia Informatica.

71. *An evaluation of statistical approaches to text categorization. .* **Yang, Y.** s.l. : Technical Report CMU-CS-97-127, Carnegie Mellon University, 1997.

72. **Gwet, K. L.** *Handbook Of Inter-Rater Reliability.* s.l. : Advanced Analytics, LLC, 2010.

73. Types of Reliability. *socialresearchmethods.* [Online] [Cited: 26 02 2013.] http://www.socialresearchmethods.net/kb/reltypes.php.

74. *CPAN.* [Online] [Cited: 02 02 2013.] http://www.cpan.org/.

75. **Wall, L., Christiansen, T. and Orwant, J.** *Programming Perl.* s.l. : O'Reilly Media, 2000.

76. CPAN. *AI Categorizer.* [Online] http://search.cpan.org/~kwilliams/AI-Categorizer-0.09/lib/AI/Categorizer.pm.

77. PorterStemmer. *The Porter Stemming Algorithm.* [Online] [Cited: 11 03 2013.] http://tartarus.org/martin/PorterStemmer/.

78. *A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews.* **Ghorashi, S. H., et al., et al.** 2012, IJCSI International Journal of Computer Science Issues, pp. 1694-0814.

79. *The Stanford Natural Language Processing Group.* [Online] [Cited: 11 03 2013.] http://nlp.stanford.edu/software/lex-parser.shtml.