Computacion y Sistemas 2017 vol.21 N2, pages 227-241

Demographic prediction based on user reviews about medications

Tutubalina E., Nikolenko S.

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

Drug reactions can be extracted from user reviews provided on the Web, and processing this information in an automated way represents a novel and exciting approach to personalized medicine and wide-scale drug tests. In medical applications, demographic information regarding the authors of these reviews such as age and gender is of primary importance; however, existing studies usually assume that this information is available or overlook the issue entirely. In this work, we propose and compare several approaches to automated mining of demographic information from user-generated texts. We compare modern natural language processing techniques, including feature rich classifiers, extensions of topic models, and deep neural networks (both convolutional and recurrent architectures) for this problem.

http://dx.doi.org/10.13053/CyS-21-2-2736

Keywords

Demographic prediction, Medications, User reviews

References

[1] Arkhipenko, K., Kozlov, I., Trofimovich, J., Skorniakov, K., Gomzin, A., & Turdakov, D. (2016). Comparison of neural network architectures for sentiment analysis of Russian tweets. Proceedings of International Conference, Computational Linguistics and Intellectual Technologies.

[2] Bayot, R. & Gonçalves, T. (2016). Author profiling using SVMs and word embedding averages-notebook for PAN at CLEF. CLEF Evaluation Labs and Workshop-Working Notes Papers, Évora, Portugal.

[3] Bouanani, S.E.M.E. & Kassou, I. (2014). Authorship analysis studies: A survey. International Journal of Computer Applications, Vol. 86, No. 12, pp. 22-29.

[4] Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., & Lai, J.C. (1992). Class-based n-gram models of natural language. Computational linguistics, Vol. 18, No. 4, pp. 467-479.

[5] Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., & Nissim, M. (2016). Gronup: Groningen user profiling-Notebook for PAN. CLEF 2016, Evaluation Labs and Workshop-Working Notes Papers, pp. 5-8, Évora, Portugal.

[6] Feldman, R., Netzer, O., Peretz, A., & Rosenfeld, B. (2015). Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, ACM, New York, USA, pp. 1779-1788.

[7] Forner, P., Navigli, R., & Tufis, D. (2013). CLEF evaluation labs and workshop-Working notes papers. pp. 23-26, Valencia, Spain.

[8] Goldberg, Y. (2015). A primer on neural network models for natural language processing. CoRR, abs/1510.00726.

[9] Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. ACM Comput. Surv., Vol. 47, No. 4, 56:1-56:39.

[10] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[11] Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 437-442.

[12] Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate.

[13] Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. (2010). Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the Workshop on Biomedical Natural Language Processing, BioNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 117-125.

[14] Marcus, A. D. (2014). Researchers fret as social media lift veil on drug trials. Wall Street Journal.

[15] Martinez, P., Martinez, J. L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., & Revert, R. (2016). Turning user generated health-related content into actionable knowledge through text analytics services. Computers in Industry, Vol. 78, pp. 43-56.

[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.

[17] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model, INTERSPEECH, Vol. 2, No. 3.

[18] Mikolov, T., Kombrink, S., Burget, L., Cernocký, J. H., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 5528-5531.

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributedrepresentations of words and phrases and their compositionality. CoRR, abs/1310.4546.

[20] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, pp. 3111-3119.

[21] Mnih, A. & Hinton, G. E. (2009). A scalable hierarchical distributed language model. Advances in neural information processing systems, pp. 1081-1088.

[22] Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). Author age prediction from text using linear regression. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, pp. 115-123.

[23] Pedersen, T. (2015). Screening twitter users for depression and ptsd with lexical decision lists. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, pp. 46-53.

[24] Plachouras, V., Leidner, J. L., & Garrow, A. G. (2016). Quantifying self-reported adverse drug events on twitter: Signal and topic analysis. Proceedings of the 7th 2016 International Conference on Social Media & Society, SMSociety '16. ACM, New York, NY, USA, pp. 6:1-6:10.

[25] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing. Proceedings of Languages in Biology and Medicine.

[26] Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 457-465.

[27] Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at pan 2013. CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, pp. 352-365.

[28] Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. CLEF.

[29] Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daeleman, W., et al. (2014). Overview of the 2nd author profiling task at PAN 2014. CEUR Workshop Proceedings, Vol. 1180, CEUR Workshop Proceedings, pp. 898-927.

[30] Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF.

[31] Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, pp. 37-44.

[32] Rastegar-Mojarad, M., Liu, H., & Nambisan, P. (2016). Using social media data to identify potential candidates for drug repurposing: A feasibility study. JMIR Res Protoc, Vol. 5, No. doi:10.2196/resprot.5621.

[33] Rong, X. (2014). Word2vec parameter learning explained. CoRR, abs/1411.2738.

[34] Sarker, A. & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, Vol. 53, pp. 196-207.

[35] Sarker, A., Nikfarjam, A., & Gonzalez, G. (2016). Social media mining shared task workshop. Proc. Pacific Symposium on Biocomputing, pp. 581-592.

[36] Segura-Bedmar, I., Martinez, P., Revert, R., & Moreno-Schneider, J. (2015). ExploringSpanish health social media for detecting drug effects. BMC Medical Informatics and Decision Making, Vol. 15, No. 2, pp. 1-9. doi:10.1186/1472-6947-15-S2-S6.

[37] Shaywitz, D. & Mammen, M. (2011). The next killer app. The Boston Globe.

[38] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, Vol. 60, No. 3, pp. 538-556.

[39] Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). Overview of the author identification task at PAN.

[40] Tutubalina, E. & Nikolenko, S. I. (2016). Automated prediction of demographic information from medical user reviews. Proc. 4th International Conference on Mining Intelligence and Knowledge Exploration, Lecture Notes in Artificial Intelligence, Springer.

[41] Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, Vol. 35, No. 3, pp. 399-433.

[42] Yang, C. C., Yang, H., Jiang, L., & Zhang, M. (2012). Social media mining for drug safety signal detection. Proceedings of the International Workshop on Smart Health and Wellbeing, SHB '12, ACM, New York, USA, pp. 33-40. doi:10.1145/2389707.2389714.

[43] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, MIT Press, Cambridge, USA, pp. 649-657.

[44] Zhang, Z., Nie, J.-Y., & Zhang, X. (2016). An ensemble method for binary classification of adverse drug reactions from social media. Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing.

[45] Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime investigation. Intelligence and Security Informatics, Springer, pp. 59-73.