Proceedings of the AINL FRUCT 2016 Conference, 2017

Predicting the age of social network users from user-generated texts with word embeddings

Alekseev A., Nikolenko S.

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

Abstract

© 2016 FRUCT. Many web-based applications such as advertising or recommender systems often critically depend on the demographic information, which may be unavailable for new or anonymous users. We study the problem of predicting demographic information based on user-generated texts on a Russian-language dataset from a large social network. We evaluate the efficiency of age prediction algorithms based on word2vec word embeddings and conduct a comprehensive experimental evaluation, comparing these algorithms with each other and with classical baseline approaches.

References

[1] Modeling Interestingness with Deep Neural Networks. EMNLR 2014.

[2] C. Aggarwal and R Zhao. Graphical models for text: A new paradigm for text representation and processing. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pages 899-900, New York, NY, USA, 2010. ACM.

[3] R. Al-Rfou, B. Perozzi, and S. Skiena. Polyglot: Distributed word representations for multilingual nip. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 183192, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[4] N. Arefyev, A. Panchenko, A. Lukanin, O. Lesota, and P. Romanov. Evaluating three corpus-based semantic similarity systems for Russian. In Proceedings of International Conference on Computational Linguistics Dialogue, 2015.

[5] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.

[6] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3: 1137-1155, 2003.

[7] Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, pages 137-186. Springer, 2006.

[8] M. Berggren, J. Karlgren, R. Ostling, and M. Parkvall. Inferring the location of authors from words in their texts. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), pages 211-218, Vilnius, Lithuania, May 2015. Linkoping University Electronic Press, Sweden.

[9] S. Bergsma and B. Van Durme. Using conceptual class attributes to characterize social media users. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 710-720, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] E. Bloedorn and I. Mani. Using {NLP} for machine learning of user profiles. Intelligent Data Analysis, 2 (l-4): 3-18, 1998.

[11] G. Boleda, S. Pado, and J. Utt. Regular polysemy: A distributional model. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval '12, pages 151-160, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] S. E. M. E. Bouanani and I. Kassou. Article: Authorship analysis studies: A survey. International Journal of Computer Applications, 86 (12): 22-29, January 2014. Full text available.

[13] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. Comput. Linguist., 18 (4): 467-479, 1992.

[14] E. Bruni, N. K. Tran, and M. Baroni. Multimodal distributional semantics. J. Artif. Int. Res., 49 (1): 1-7, 2014.

[15] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL '96, pages 310-318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[16] Z. Chen, W. Lin, Q. Chen, X. Chen, S. Wei, H. Jiang, and X. Zhu. Revisiting word embedding for contrasting meaning. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 106-115, Beijing, China, 2015. Association for Computational Linguistics.

[17] A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni. Linguistic profiling based on general-purpose features and native language identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 207-215, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[18] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In International AAAI Conference on Weblogs and Social Media, 2013.

[19] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu. Gender identification on twitter using the modified balanced winnow. 2012.

[20] N. Djuric, H. Wu, V. Radosavljevic, M. Grbovic, and N. Bhamidi-pati. Hierarchical neural language models for joint representation of streaming documents and their content. In Proceedings of the 24th International Conference on World Wide Web, WWW ' 15, pages 248255, New York, NY, USA, 2015. ACM.

[21] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson. Author profiling for english emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACUNG'07), pages 263-272, 2007.

[22] L. Flekova and I. Gurevych. Personality profiling of fictional characters using sense-level links between lexical resources. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1805-1816, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[23] P. Forner, R. Navigli, and D. Tufis. Clef 2013 evaluation labs and workshop-working notes papers, 23-26 September, Valencia, spain (2013). URL http. /fwww. clef-initiative. eu/publicationAvorking-notes.

[24] Y. Goldberg. A primer on neural network models for natural language processing. CoRR, abs/1510. 00726, 2015.

[25] J. T. Goodman. A bit of progress in language modeling. Comput. Speech Lang., 15 (4): 403-134, 2001.

[26] R. M. Green and J. W. Sheppard. Comparing frequency-and style-based features for twitter author identification. In FLAIRS Conference, 2013.

[27] B. Han, P. Cook, and T. Baldwin. A stacking-based approach totwitter user geolocation prediction. In Proceedings of the 51st Annual Meeting of the Association for Computational linguistics: System Demonstrations, pages 7-12, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[28] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 873-882. Association for Computational Linguistics, 2012.

[29] M. Kageback, O. Mogren, N. Tahmasebi, and D. Dubhashi. Extractive summarization using continuous vector space models. In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL, pages 31-39, 2014.

[30] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, volume 1, pages 181-184 vol. 1, 1995.

[31] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate. 2006.

[32] V. Lampos, D. PreoAiuc-Pietro, and T. Conn. A user-centric model of voting intention from social media. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 993-1003, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[33] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. CoRR, abs/1405. 4053, 2014.

[34] J. Li, A. Ritter, and E. Hovy. Weakly supervised user profile extraction from twitter. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 165174, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[35] R. Ling, T. F. Bertel, and R R. Sunds0y. The socio-demographics of texting: An analysis of traffic data. New Media & Society, 14 (2): 281-298, 2012.

[36] J. Liu and D. Inkpen. Estimating user location in social media with stacked denoising auto-encoders. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 201-210, Denver, Colorado, June 2015. Association for Computational Linguistics.

[37] R Liu, X. Qiu, and X. Huang. Learning context-sensitive word embeddings with neural tensor skip-gram model. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, pages 1284-1290. AAAI Press, 2015.

[38] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, pages 2418-2424. AAAI Press, 2015.

[39] Q. Luo and W. Xu. Learning word vectors efficiently using shared representations and document representations. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, pages 4180-4181. AAAI Press, 2015.

[40] Q. Luo, W. Xu, and J. Guo. A study on the cbow model's overfitting and stability. In Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval 38; Reasoning, Web-KR '14, pages 9-12, New York, NY, USA, 2014. ACM.

[41] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. ACM Transactions on Information Systems (TOIS), 22 (I): 54-88, 2004.

[42] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301. 3781, 2013.

[43] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. INTERSPEECH, 2: 3, 2010.

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

[45] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. CoRR, abs/1310. 4546, 2013.

[46] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In Advances in neural information processing systems, pages 1081-1088, 2009.

[47] A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In C. J. C. Bulges, L. Bottou, M. Welling, Z. Ghabramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 2265-2273. Curran Associates, Inc., 2013.

[48] S. I. Nikolenko and A. Alekseyev. User profiling in text-based recommender systems based on distributed word representations. In Proc. 5tInternational Conference on Analysis of Images, Social Networks, and Texts, 2016.

[49] J. Oberlander and S. Nowson. Whose thumb is it anyway: classifying author personality from weblog text. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 627-634. Association for Computational Linguistics, 2006.

[50] W. Paik, S. Yilmazel, E. Brown, M. Poulin, S. Dubon, and C. Amice. Applying natural language processing (nip) based metadata extraction to automatically acquire user preferences. In Proceedings of the 1st International Conference on Knowledge Capture, K-CAP '01, pages 116-122, New York, NY, USA, 2001. ACM.

[51] A. Panchenko, N. Loukacheviteh, D. Ustalov, D. Paperno, C. M. Meyer, and N. Konstantinova. Russe: The first workshop on Russian semantic similarity. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue), pages 89-105, 2015.

[52] T. Pedersen. Screening twitter users for depression and ptsd with lexical decision lists. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 46-53, Denver, Colorado, June 5 2015. Association for Computational Linguistics.

[53] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[54] D. Preotiuc-Pietro, V. Lampos, and N. Aletras. An analysis of the user occupational class through twitter content. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1754-1764, Beijing, China, July 2015. Association for Computational Linguistics.

[55] M. Qiu, L. Yang, and J. Jiang. Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 401-410, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[56] A. Rahimi, T. Cohn, and T. Baldwin. Twitter user geolocation using a unified text and network prediction model. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 630-636, Beijing, China, July 2015. Association for Computational Linguistics.

[57] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin. Exploiting text and network context for geolocation of social media users. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1362-1367, Denver, Colorado, May-June 2015. Association for Computational Linguistics.

[58] F. Rangel and P. Rosso. On the impact of emotions on author profiling. Information Processing & Management, 52 (1): 73-92, 2016. Emotion and Sentiment in Social and Expressive Media.

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

[60] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. Overview of the 3rd author profiling task at pan 2015. In CLEF, 2015.

[61] F. Rangel, P. Rosso, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daeleman, et al. Overview of the 2nd author profiling task at pan 2014. In CEUR Workshop Proceedings, volume 1180, pages 898-927. CEUR Workshop Proceedings, 2014.

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

[63] J. Reisinger and R. J. Mooney. Multi-prototype vector-space models of word meaning. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT ' 10, pages 109-117, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[64] X. Rong. word2vec parameter learning explained. CoRR, abs/1411. 2738, 2014.

[65] E. Stamatatos. A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60 (3): 538-556, 2009.

[66] E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. Lopez-Lopez, M. Potthast, and B. Stein. Overview of the author identification task at pan 2015.

[67] D. Tang, B. Qin, and T. Liu. Learning semantic representations of users and products for document level sentiment classification. Li Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1014-1023, Beijing, China, July 2015. Association for Computational Linguistics.

[68] W. Tau Yih, X. He, and C. Meek. Semantic parsing for single-relation question answering. In Proceedings of ACL. Association for Computational Linguistics, 2014.

[69] S. Volkova, G. Coppersmith, and B. Van Durme. Inferring user political preferences from streaming communications. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 186-196, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[70] Z. Wang, S. LI, F. Kong, and G. Zhou. Collective personal profile summarization with social networks. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 715-725, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[71] B. Wing and J. Baldridge. Hierarchical discriminative classification for text-based geolocation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 336-348, Doha, Qatar, October 2014. Association for Computational Linguistics.

[72] Z. Wu and C. L. Giles. Sense-aware semantic analysis: A multi-prototype word representation model using wiMpedia. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAT15, pages 2188-2194. AAAI Press, 2015.

[73] W.-t. Yih, G. Zweig, and J. C. Piatt. Polarity inducing latent semantic analysis. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 1212-1222, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[74] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In Intelligence and Security Informatics, pages 59-73. Springer, 2003.