Abstract

Purpose: This is an attempt to classify the level of noise in twitter texts which is part of social media data analytics problem. Estimations in recent machine learning & data feeding algorithms researches’ assumptions consider high data quality in social media texts, while they actually lack data accuracy, completeness, and overall quality which leads to the principle of “Garbage In Garbage Out” resulting in bizarre statistical findings. The aim of this project is to predict and classify Twitter data noise levels using a labelled dataset.

Methodology: After data cleaning, a clustering technique was used to find the major dimensions in the data imported, and a dimension reduction algorithm was ran using PCA Weighting and the Wight Guided Feature Selection algorithms. They resulted into 6 most significant features which were used in the implementation. An artificial neural network model was trained to predict the Tweets’ quality classes using R and RStudio. The ANN used is Neural Network (NN) and Naïve Bayes (NB) for the purpose of predicting the Twitter text quality. There will be a comparison between the 2 ANN used in terms of accuracy and precision.
Findings: Three different aspects of text mining were discovered in twitter data. (1) Neural network gives surprisingly good result as compared to Naive Bayes algorithm, (2) With only 3 hidden layers, a network was created which can predict good or bad class, (3) Preprocessing of the data and implementing predictive algorithms take huge data and very high computational complexity and time. Research results show that Neural Network performs well even without Dropout layer and convolutional layers. The accuracy of the Neural Network is 99%.

References

1. Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in adaptive data analysis, 1(01), 1-41.
2. Aggarwal, C. C., & Wang, H. (2011). Text mining in social networks. In Social network data analytics (pp. 353-378). Springer, Boston, MA.
3. McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf
4. Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In ICML (Vol. 99, pp. 200-209). http://www1.cs.columbia.edu/~dplewis/candidacy/joachims99transductive.pdf
5. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2-3), 103-134. https://link.springer.com/article/10.1023/A:1007692713085
6. Baker, L. D., & McCallum, A. K. (1998, August). Distributional clustering of words for text classification. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 96-103). ACM. https://dl.acm.org/citation.cfm?id=290970
7. Göpferich, S. (1995). A pragmatic classification of LSP texts in science and technology. Target. International Journal of Translation Studies, 7(2), 305-326.
8. Mosquera, A., & Moreda, P. (2012, May). Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts. In Proceedings of the ICWSM workshop: Real-Time Analysis and Mining of Social Streams (RAMSS).
9. Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the 2008 international conference on web search and data mining (pp. 183-194). ACM.
10. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.
11. Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in adaptive data analysis, 1(01), 1-41.
12. Aggarwal, C. C., & Wang, H. (2011). Text mining in social networks. In Social network data analytics (pp. 353-378). Springer, Boston, MA.
13. Göpferich, S. (1995). A pragmatic classification of LSP texts in science and technology. Target. International Journal of Translation Studies, 7(2), 305-326.
14. Mosquera, A., & Moreda, P. (2012, May). Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts. In Proceedings of the ICWSM workshop: Real-Time Analysis and Mining of Social Streams (RAMSS). classification tasks. Information Processing & Management, 45(4), 427-437.
15. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), 1289-1305. http://www.jmlr.org/papers/v3/forman03a.html

16. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

17. Clark, E., & Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences, 27, 2-11.

18. Hu, X., & Liu, H. (2012). Text analytics in social media. In Mining text data (pp. 385-414). Springer US.

19. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how different social media sources? In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 356-364)

20. He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.

21. Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the 2008 international conference on web search and data mining (pp. 183-194). ACM.

22. Tan, A.-H. (1997). Cascade ARTMAP: Integrating neural computation and symbolic knowledge processing. IEEE Transactions on Neural Networks, 8(2), 237-250.

23. Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), 45-66. http://www.jmlr.org/papers/v2/tong01a.html

24. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841-842). ACM. https://dl.acm.org/citation.cfm?id=1835643

25. Clark, E., & Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences, 27, 2-11.

26. Hu, X., & Liu, H. (2012). Text analytics in social media. In Mining text data (pp. 385-414). Springer US.

27. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how different social media sources? In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 356-364).

28. He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.

Index Terms

Computer Science  Artificial Intelligence
Keywords

Data Mining, Twitter Text Quality, Twitter Data Classification, Classification Algorithms, Neural Network Algorithm, Text Analysis.