Industrial Implementations of Natural Language Processing

Abstract

Natural language processing (NLP) is becoming increasingly important with the penetration of social media platforms. There're many industrial implementations of NLP, including direct ones and indirect ones that helps solve down-stream problems.

Natural language processing (NLP) is a field targeting at finding implicit relationships in natural language data. Though in the big data era, it could still be a problem to find useful numerical data, as some measurements are hard to be quantified and expensive to get. However, there are a lot of natural language data available. For example, Twitter with billions of users, serves as a good data pool for NLP. Without much numerical data at hand, NLP, is another approach to make use of resourceful natural language data to discover certain patterns behind data.

The development of NLP, along with penetration of social media platforms makes it possible to use natural language data to find more implicit correlation besides explicit ones in numerical data. NLP guarantees the possibility to discover patterns behind natural language data that is complementary to numerical data and social media platforms provides rich natural language data to gather. Therefore, there are many industrial implementations of natural language processing.

For example, natural language processing can be used to model user personality. There's hardly non-textual data to describe personality. However, if we make use of natural language data, we could find many posts about a user and then model his/her personality based on this resourceful data. It is where natural language processing techniques play the role. Oberlander and Nowson [2006], Celli [2012], Maheshwari et al. [2017] all have proposed successful methods. Usually people put forward certain traits used to describe personality and then regard the problem as a classification task, where regular classifiers like support vector machine, logistic regression, and random forest can be applied. Similar implementations occur in health industry to modeling user health profile or gender and ethnicity.

In the above cases, natural language processing techniques are used to directly solve the tasks. On the other hand, it could also be implemented to solve some down-stream tasks. We could extract useful information from natural language data and then use the extracted information to make predictions. In this way, natural language processing serves as an intermediary tool to help improve the down-stream forecasting. This is an indirect way of natural language processing implementation.

Take finance industry as an example, behavioral economics focuses on the relationship between public mood and economic indicators. However, it is hard to quantify public mood.

People therefore apply natural language processing techniques to process text data from online social platform, such as Twitter. For example, Bollen et al. [2010] uses Twitter feeds to represent public mood and then investigates the correlations between collective mood states and Dow Jones Industrial Average. To present public mood in a numerical way, Bollen et al. [2010] uses two mood tracking tools. They first analyze positive and negative mood of text contents and then project text data into six particularly designed mood dimensions. As public mood is an important factor in finance industry, there are many other successful attempts. Porshnev et al. [2013] in another way uses the lexicon-based approach to evaluate presence of eight predefined emotions, which also provides many extra information from natural language data and therefore improves the accuracy of prediction.

While natural language data do provide complementary information that is useful and sometimes critical in certain applications, they are not the panacea. Bollen et al. [2010] finds that information about public mood improves the accuracy of some DJIA predictions yet not all the predictions. There could be several reasons. First, information extracted from natural language data could be misleading. In the case of Bollen's work, natural language data are not directly quantifying public mood. On the other hand, they are extracted and projected by some NLP techniques, which could be misleading. In addition, information from natural language data may be inherently not useful in certain applications. After all, they are just relevant data without strict causal relationships.

Reference

[1] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: Classifying author personality from weblog text. In Proceedings of COLING/ACL 2006 (Posters), pages 627–634. Association for Computational Linguistics, 2006. http://www.aclweb.org/anthology/P06-2081.pdf.DOI: 10.3115/1273073.1273154 110

[2] Fabio Celli. Unsupervised personality recognition for social network sites. In The Sixth International Conference on Digital Society ICDS 2012, January 2012. http://www.worldcat.org/isbn/978-1-61208-176-2. 110

[3] Tushar Maheshwari, Aishwarya N. Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 731–741, Valencia, Spain, April 2017. Association for Computational Linguistics. http://www.aclweb.org/anthology/E17-1069.DOI: 10.18653/v1/e17-1069 110

[4] Jonah Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Computing Research Repository (CoRR), abs/1010.3003, 2010. http://arxiv.org/abs/1010.3003.DOI: 10.1016/j.jocs.2010.12.007 97

[5] Alexabder Porshnev, Ilyia Redkin, and Alexey Shevchenko. Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, pages 440–444, Dec 2013. DOI: 10.1109/ICDMW.2013.111 97