Text Analytics Final Project Part 2: Literature Review
Rachel Rosenberg
November 2019

My project will be multi-label classification of both age group and astrological sign on a dataset of 600,000 blog posts made to blogger.com in 2004 (Schler, Koppel, Argamon, and Pennebaker, 2006). This dataset contains labels for the gender, age, category, and astrological (zodiac) sign of the author. It comes with one file (labelled file name) per author, with all of the author's blog posts included in one text file.

The simplest version of text classification is sentiment analysis, which aims to classify the opinions of writers of sections of text as either "positive" (1) or "negative" (0) (Liu, 2010). As Liu's 2010 introduction to sentiment analysis argues, this can be too general of a method; peoples' opinions are expressed in complex ways, often leading to sections of opinionated text that are misinterpreted or that, even if represented correctly, are irrelevant to the analyst's business need.

Adding a third, neutral, class to sentiment analysis is one good way to improve the accuracy and usability of models (Ding, Liu, and Yu, 2008). The neutral class can either be removed or separately predicted (methodology will depend on future use case); however, either way, adding a neutral case strengthens predictions made toward the other two classes (Ding, Liu, and Yu, 2008; Koppel and Schler, 2006). This brings us to multi-label text classification, which requires little more work and extends the usability of the text classification model (Koppel and Schler, 2006).

There are many types of algorithms that can be used in text classification, as in other applications of classification problems. These can include standard methods like Decision Trees and Logistic Regression, ensemble methods like Random Forests and Boosted Trees, and deep learning methods like Convolutional Neural Networks and LSTMs (Khan, Baharudin, Lee, and Khan, 2010). Different algorithms are good for different use cases (Pawar and Gawande, 2012); for example, kNN clustering for grouping of news articles (Tam, Santoso, and Setiono, 2002), sentence importance labelling and Naïve Bayes clustering for a similar set of news articles (Ko, Park, and Seo, 2004), and dictionary-based categorization of chemical web pages (Liang *et. al*, 2006).

The variety of approaches taken in the past speaks to the long history and extreme flexibility of text classification. It is possible to combine, munge, and parse data in a huge number of ways; the way that works the best for a given application will depend heavily on the data's structure and future use cases of the model. I will likely attempt simple TF-IDF feature creation, as well as sentence importance labelling and a dictionary-based approach, since these seem promising for parsing out only the most important features in my very large dataset.

Related Literature

X. Ding, B. Liu, and P. S. Yu. (2008). A Holistic Lexicon-Based Approach to Opinion Mining in *Proceedings of the 2008 International Conference on Web Search and Data Mining.* (pdf)

A. Khan, B. Baharudin, L. H. Lee, and K. Khan. (2010). A Review of Machine Learning Algorithms For Text-Documents Classification in *Journal of Advances in Information Technology*, 1(1), 4-20. (pdf)

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. Information processing & management, 40(1), 65-79. (link)

M. Koppel and J. Schler. (2006). The Importance of Neutral Examples for Learning Sentiment in *Computational Intelligence*. (pdf)

Liang, C. Y., Guo, L., Xia, Z. J., Nie, F. G., Li, X. X., Su, L., & Yang, Z. Y. (2006). Dictionary-based text categorization of chemical web pages in *Information processing & management*, 42(4), 1017-1029. (link)

B. Liu. (2010). Sentiment Analysis and Subjectivity in *Handbook of Natural Language Processing* (Second ed.). (pdf)

Pawar, P. Y., & Gawande, S. H. (2012). A comparative study on different types of approaches to text categorization in *International Journal of Machine Learning and Computing*, 2(4), 423. (pdf)

J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. (pdf)

Tam, V., Santoso, A., & Setiono, R. (2002). A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization in *Object recognition supported by user interaction for service robots* (Vol. 4, pp. 235-238). IEEE. (pdf)