How to Predict Bestsellers
and
What This Tells Us About Literature

Joris van Zundert[†1]
joris.van.zundert@huygens.knaw.nl

Marijn Koolen[†]
marijn.koolen@huygens.knaw.nl

Karina van Dalen-Oskam[†⋆]
karina.van.dalen@huygens.knaw.nl

[†]Huygens Institute for the History of the Netherlands
Royal Netherlands Academy of Arts and Sciences
[⋆]University of Amsterdam

Abstract

Publishing companies are experiencing difficult times. Although more books are published than ever before (Segura 2017), especially publishers of literary prestigious fiction are struggling to make a profit in the book business. An analysis of the Dutch market showed that often revenue from just a few bestsellers has to cover losses incurred by a large number of titles that do not sell well (Buss 2015). The same study suggested, inter alia, to look into computational methods to remedy the situation (p.84). Acknowledging the publishing industry as an application ground

1. We would like to thank the people at WPG Publishers (http://www.wpg.nl) who kindly provided us with the sales data that made this research possible. We also would like to extend our gratitude to our colleagues at the National Library of the Netherlands who provided the digital research corpus and secured environment that was used to execute this research. We thank also Hermann Buss and Michel Blaauw of Driven By Data (http://driven-by-data.com) and Emile den Tex, who all contributed significantly to our research.

for digital humanities methods we hypothesized that data on published titles combined with machine learning techniques may give us clues as to the features of bestselling and non selling fiction, while the same research process might simultaneously come to the aid of publishers of literary prestigious materials. Obviously if we would be able to infer the features of what makes a bestseller we might be able to predict the selling capability of a title and thereby improve the profitability of the publishing process. Apart from Jodi Archer's and Matthew Jocker's widely acknowledged *The Bestseller Code* (2016) there seems to have been almost no work done in this area.

From a computer science perspective the problem of predicting bestsellers can be cast as a binary classification problem: can an algorithm given a large enough training and test set distinguish between known bestsellers and non selling works of fiction? To establish a baseline we have created a training set of 400 novels (Dutch and translated works of fiction) for which sales numbers for the years 2010–2016 are known; 200 of these are bestsellers (more than 12,000 copies sold), 200 have sold few copies (between 0 and a 100 copies).[2] Apart from cutting front matter (title page, imprint, copyright, contents and so forth) no preprocessing was applied. Also all other metadata such as author name, genre, date of first publishing etc. have been left out of the analysis. This 'naive raw data' approach is on purpose as we are looking for a method that can be easily deployed and thus applied with as little preprocessing labor as possible. In the same way a test set of 50 bestsellers and 50 non sellers was created.

The text of each novel was used to construct a term frequency–inverse document frequency (tf–idf) matrix that for each document represents the relative importance of a word's occurrence within a text. This tf–idf matrix, plus for each novel the knowledge if it was a bestseller or non seller (expressed respectively as 1 or 0), then served as the input for a feed forward multi-layer perceptron (MLP, see Beam 2017) with just the minimum of three layers (input, hidden, and output). With this most naive approach possible we found a success rate of ~80% accuracy. That is: after training the neural network model was able in 80% of cases to predict correctly if a novel from the test set, which it had not previously seen, was a bestseller or a non seller.

This initial success warrants further cross validation as well as performance comparison with other possible algorithms. For this we want to

2. Note that these numbers reflect the Dutch publishing market, which is in comparison relatively small to, for instance, the US market. Sales numbers of bestsellers are therefore in absolute sense smaller than what one would expect for e.g. the US market.

know, for instance, how many examples are needed to train a good predictor? What is impact of data set size? Does stability increase linearly with set size? Lastly and most saliently perhaps we would like to know what features of a novel are effective as input for training a predictor, as presumably these features reveal what properties of texts makes them bestseller material.

A particular interesting model to compare the MLP model with would be the Ružička or MinMax metric (Schubert and Telcs 2014), which has been recently applied in stylometry exercises with impressively accurate results (Kestemont et al. 2016). Because MinMax is a similarity metric we can slightly recast our research question from a binary classification problem to a question of which books are similar to current bestsellers. Concretely, the classifier determines whether a novel to be classified is more similar to a training set of N bestsellers or a set of N non sellers.

To study the impact of training set size and to cross validate, both models are trained on sets of $N_{train} = 40$ (20 top, 20 bottom), $N_{train} = 100$ (50 top, 50 bottom) and $N_{train} = 200$ (100 top, 100 bottom) novels. The validation or test set in each setting is 40 novels (20 bestsellers, 20 non sellers). The top 120 novels and bottom 120 novels in terms of sales figures are randomly sampled and split across 20/20 novels for validation and $N_{top} = 20, 50, 100, N_{bottom} = 20, 50, 100$ for training. This setup is chosen so that the different training set sizes sample from the same base set. We use repeated random sub-sampling of the training and validation sets, with 10 iterations for each of the experimental settings. This setup yields the results as shown in figure 1.

Our results show that both algorithms perform equally well on the task of classifying bestsellers and non sellers. Furthermore performance of both models is stable and improves only slightly with training set size. This means that indeed we would be able to support publishers in judging the selling capabilities of new materials proposed by authors—not so much maybe to benefit bestsellers but primarily to once more examine carefully manuscripts for which the algorithm predicts low selling numbers.

Being able to predict to a certain extend the selling capabilities of a novel may be beneficial to publishers, but what may this process gain us as to knowledge about aspects of literary prestigious fiction? An interesting difference between the reinforcement learning of the MLP and the MinMax plain document similarity measure in this respect is that the latter performed markedly less impressive at identifying non bestsellers. This indicates that there is indeed a certain stylistic 'norm' for bestselling material that both models are able to learn, which is corroborated by the fact that MinMax is apparently unable to infer any common stylistic features for

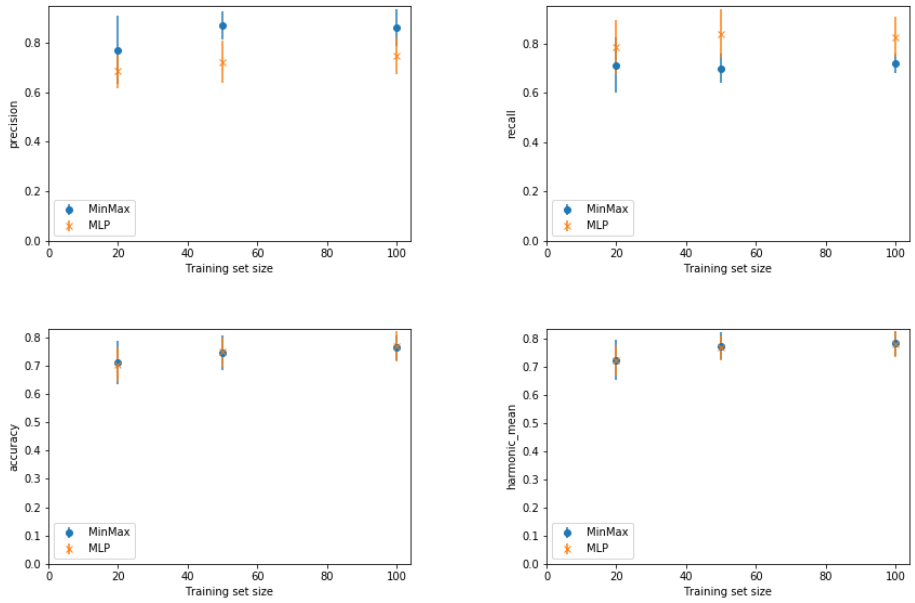

Figure 1: Evaluation results for different training set sizes using the MLP and MinMax algorithms. Reported measures are Precision, Recall, Accuracy and F1 score. The error bars indicate mean and standard deviation based on 10 iterations with each training set size.

non sellers. We note that this is different from saying that literary prestigious fiction shares the same stylistic 'norm'. Both algorithms converge on the same set of predicted bestsellers. However, that set does not comprise solely highly prestigious literary novels. Rather to the contrary such works are spread among bestsellers and non sellers alike. This may be taken to corroborate the often made contention that literary prestige is a purely social construct (for a comprehensive overview see e.g. Verboord 2003). Obviously it also indicates that literary prestige alone does not determine bestselling capacity. To sell well a novel also has to conform to certain stylistic features.

The extent to which such stylistic features are text immanent or to what extent they are also mostly determined by social signals we may start to gauge at a more concrete level of analysis where we examine particular features of high in-demand literary prestigious fiction by delving into and comparing the vocabulary of bestselling and non selling titles. To this end we explored this vocabulary by determining which words are used relatively more often by top selling titles—by taking the top 1,000 terms used in titles with a more than 80% probability of being a well selling title and comparing their relative frequencies with the relative frequencies of these terms in non selling titles. When we discard character names and function words our results reveal—as we will demonstrate in our paper—that words appearing relatively more in bestselling general fiction and prestigious literature tend to be noticeable 'masculine' in nature. This suggests, as is corroborated by research of a number of colleagues in the field (e.g. Koolen 2018; Smeets and Sanders 2018), that there is a still strong cultural tendency to prefer masculine themes and motives in both general and literary prestigious fiction.

In closing our paper we would like to reflect on the theme of the conference. It will be obvious that for both ethical and commercial reasons we cannot fully disclose specific data about authors, individual titles, and sales numbers used in our research. Our research data and results are therefore less open than we would like. However, engaging across institutional borders with a commercial partner that was prepared to share their proprietary data allowed us to significantly progress our methods and our understanding of key features of literary fiction. Having to work with such proprietary data also was a key factor for establishing a data secure experimental research environment in the Dutch National Library to enable research towards closed text corpora. Even if we prefer open data on principle grounds we have to acknowledge that some data cannot be given in open access. Nevertheless we were able to progress the abilities for researchers to work with such data and derive meaningful results from it.

References

Archer, Jodie, and Matthew L. Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel.* New York: St. Martin's Press, September. ISBN: 978-1-250-08827-7.

Beam, Andrew L. 2017. *Deep Learning 101 - Part 2: Multilayer Perceptrons.* Online book, February. Accessed November 20, 2017. http://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part2.html.

Buss, Hermann. 2015. *Bestsellers en Badsellers: Naar andere strategiën voor het uitgeven van boeken* [in Dutch]. Desset Publishers. ISBN: 987-90-823869-0-5.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. "Authorship Verification with the Ruzicka Metric." In *Digital Humanities 2016: Conference Abstracts*, 246–249. Kraków: DH Benelux, July. Accessed November 16, 2017. http://hdl.handle.net/20.500.11755/5f7d08c9-f8fe-44b0-be8a-49364f390d7b.

Koolen, C.W. 2018. "Reading Beyond the Female: the relationship between perception of author gender and literary quality." Phd, University of Amsterdam. Accessed April 27, 2018.

Schubert, András, and András Telcs. 2014. "A note on the Jaccardized Czekanowski similarity index." *Scientometrics* 98, no. 2 (February): 1397–1399. ISSN: 1588-2861, accessed November 20, 2017. doi:10.1007/s11192-013-1044-2.

Segura, Jonathan. 2017. "Print Book Sales Rose Again in 2016." *Publishers Weekly* (January). Accessed November 20, 2017. https://www.publishersweekly.com/pw/by-topic/industry-news/bookselling/article/72450-print-book-sales-rose-again-in-2016.html.

Smeets, Roel, and Eric Sanders. 2018. "Character Centrality in Present-Day Dutch Literary Fiction." In *DHBenelux 2018.* Amsterdam: Royal Netherlands Academy of Arts and Sciences Humanities Cluster, June. Accessed June 18, 2018. http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/RoelSmeets_EricSanders_CharacterCentrality_DHBenelux2018.pdf.

Verboord, Marc. 2003. "Classification of authors by literary prestige." *Poetics* 31, no. 3 (August): 259–281. Accessed November 20, 2017. doi:10.1016/S0304-422X(03)00037-8.