WHAT DO WEBSITES SAY ABOUT FIRM-LEVEL INNOVATION? - A MACHINE LEARNING APPROACH

Janna Axenbeck
Department of Digital Economy, ZEW Mannheim

Part of the research project TOBI (Text Data Based Output Indicators as Base of a New Innovation Metric), funded by the German Federal Ministry of Education and Research

NAEC conference, 16th April 2019, Paris

Leibniz Gemeinschaft

ZEW

MOTIVATION

Drawbacks of traditional firm-level innovation indicators that are based on large-scale questionnaire-based surveys:

- Lack of geographical coverage
- Costly
- It takes time to process the data
- Firm participation is required

➢ However, innovation indicators with information scraped from firm websites would allow an automatized, timely and comprehensive analysis of firm-level innovation activities.

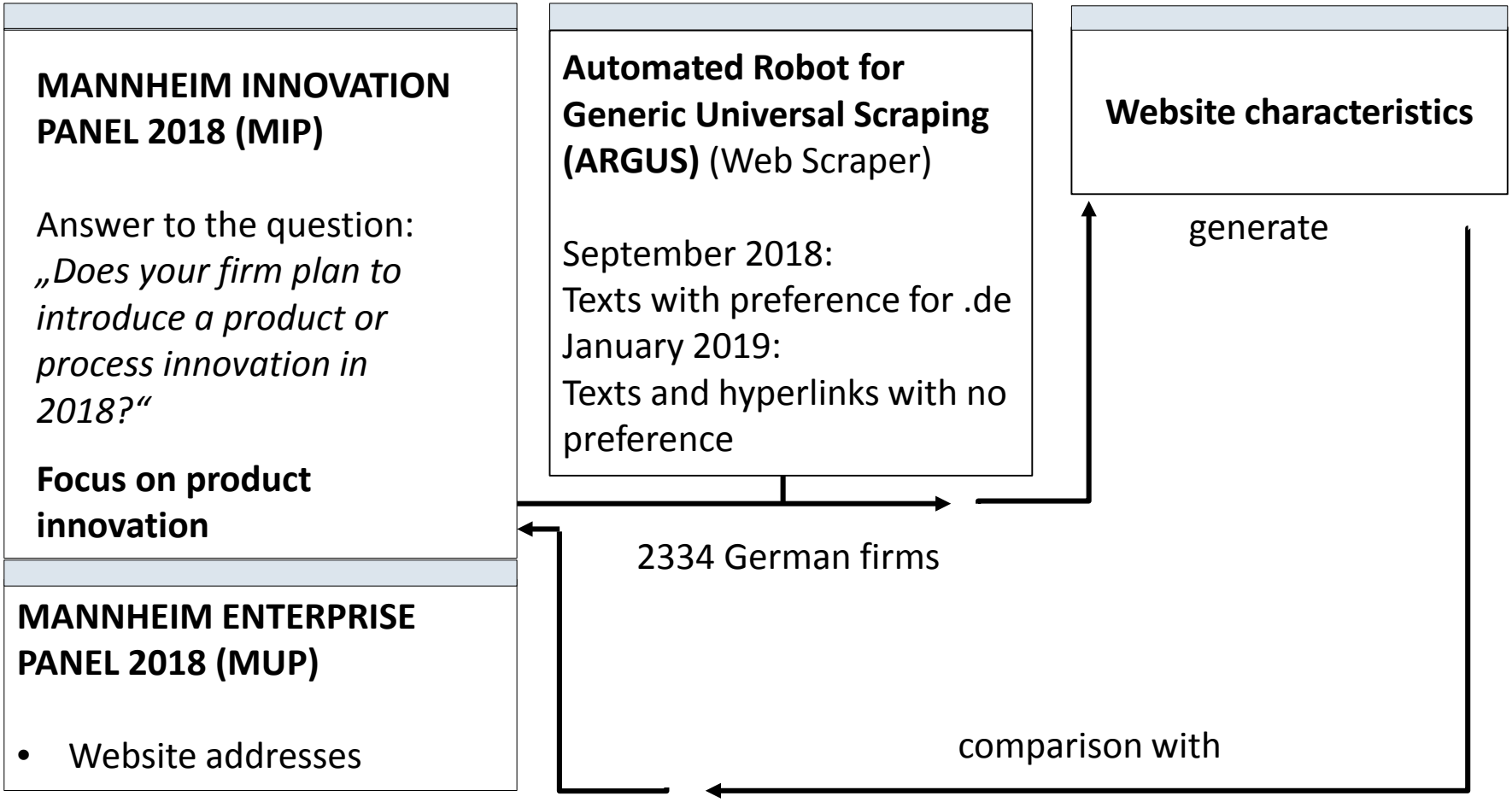
But do websites contain measurable information about firm-level innovation activities and which website characteristics best predict a firm‘s innovation status?

WEBSITE CHARACTERISTICS

Direct Information	Keywords	Innovation-related and/or product-related terms
		Terms related to emerging technologies (*from Wikipedia's list of emerging technologies)*
	Latent patterns	Topics generated by the latent Dirichlet allocation (LDA)

Indirect Information	Economic sector	Clusters based on website similarities (k-means algorithm)
	Firm size	Number of subpages; total amount of characters
	International orientation	Percentage of subpages in English language; number of occurrences of the word '*German*'
	Relationships to other firms	The sum of incoming and outgoing hyperlinks
	Social media	Hyperlinks to Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr and Vimeo

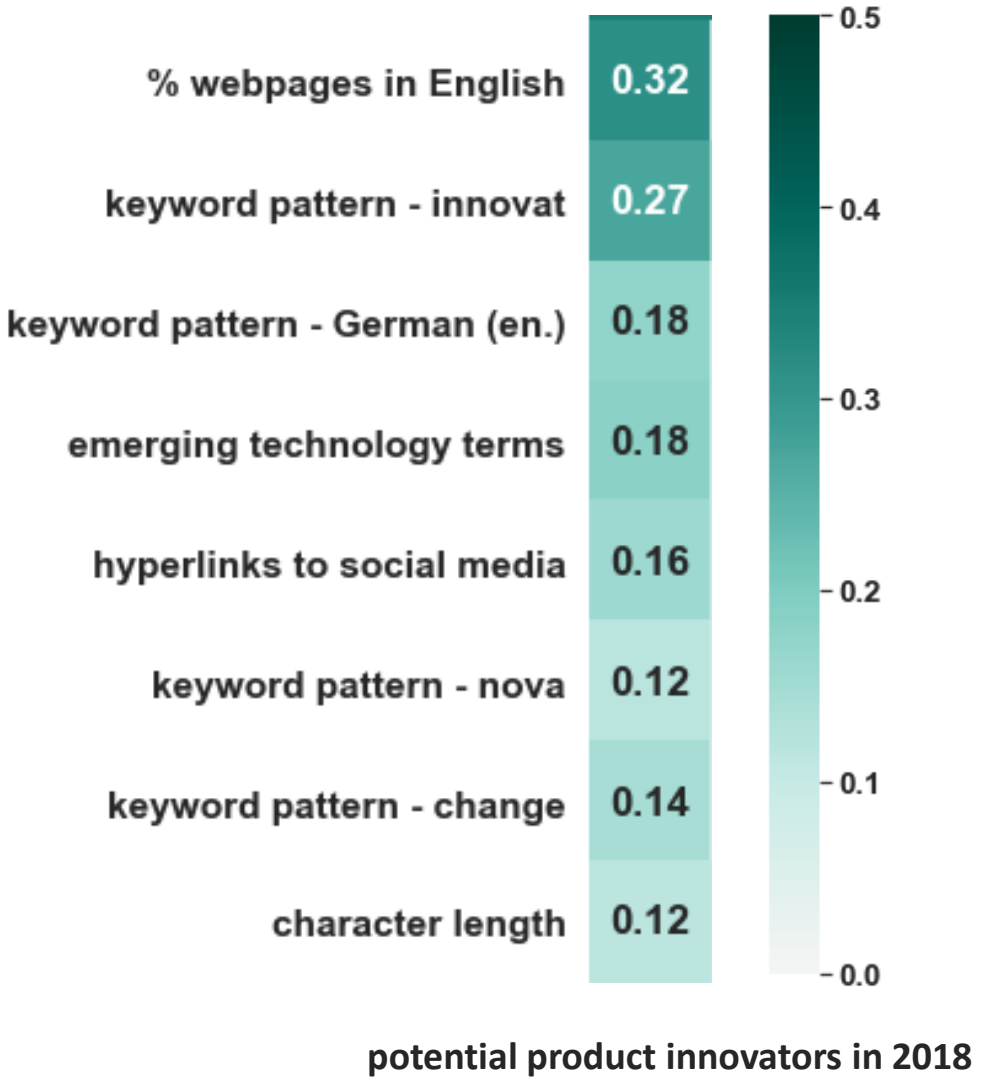
SUMMARY – RESEARCH APPROACH



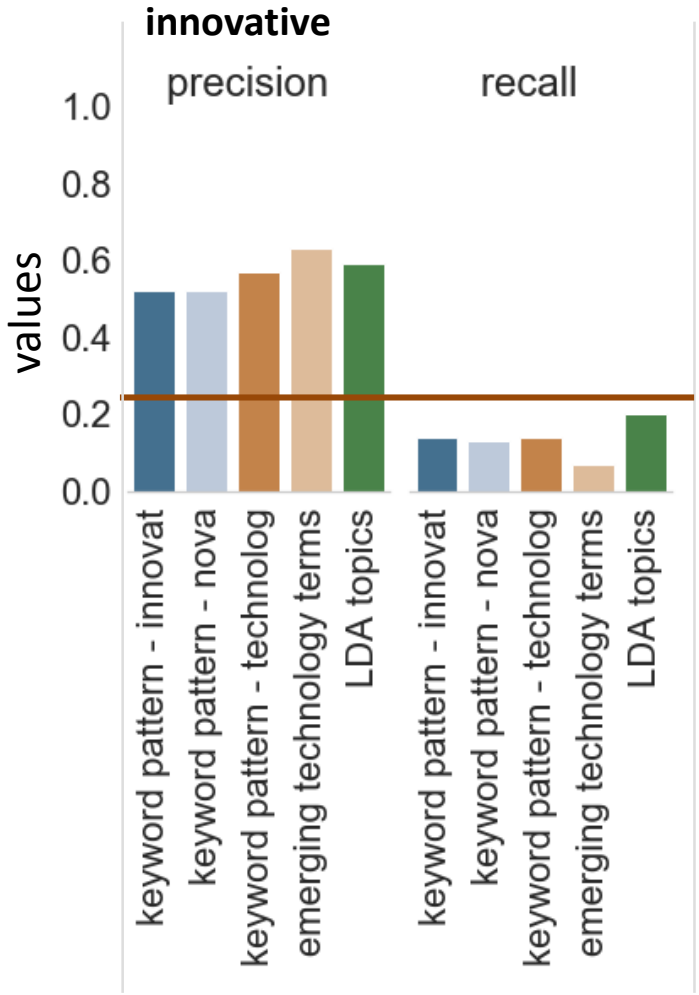
WEBSITE CHARACTERISTICS AND PRODUCT INNOVATION STATUS
Pearson correlation coefficients

The term *innovat* and the percentage of webpages in English language show the strongest correlation with potential product innovators in 2018.

The p-values of all correlation coefficients are <0.01.



PREDICTING THE PRODUCT INNOVATION STATUS WEBSITE CHARACTERISTICS CAPTURING DIRECT INFORMATION



- Logistic regression model with LASSO-regularization
- 10-fold cross-validation
- Cut-off point: 0.5
- Dependent variable: potential product innovators in 2018
- 24% of all firms planned to introduce a product innovation

— Random weighted guess

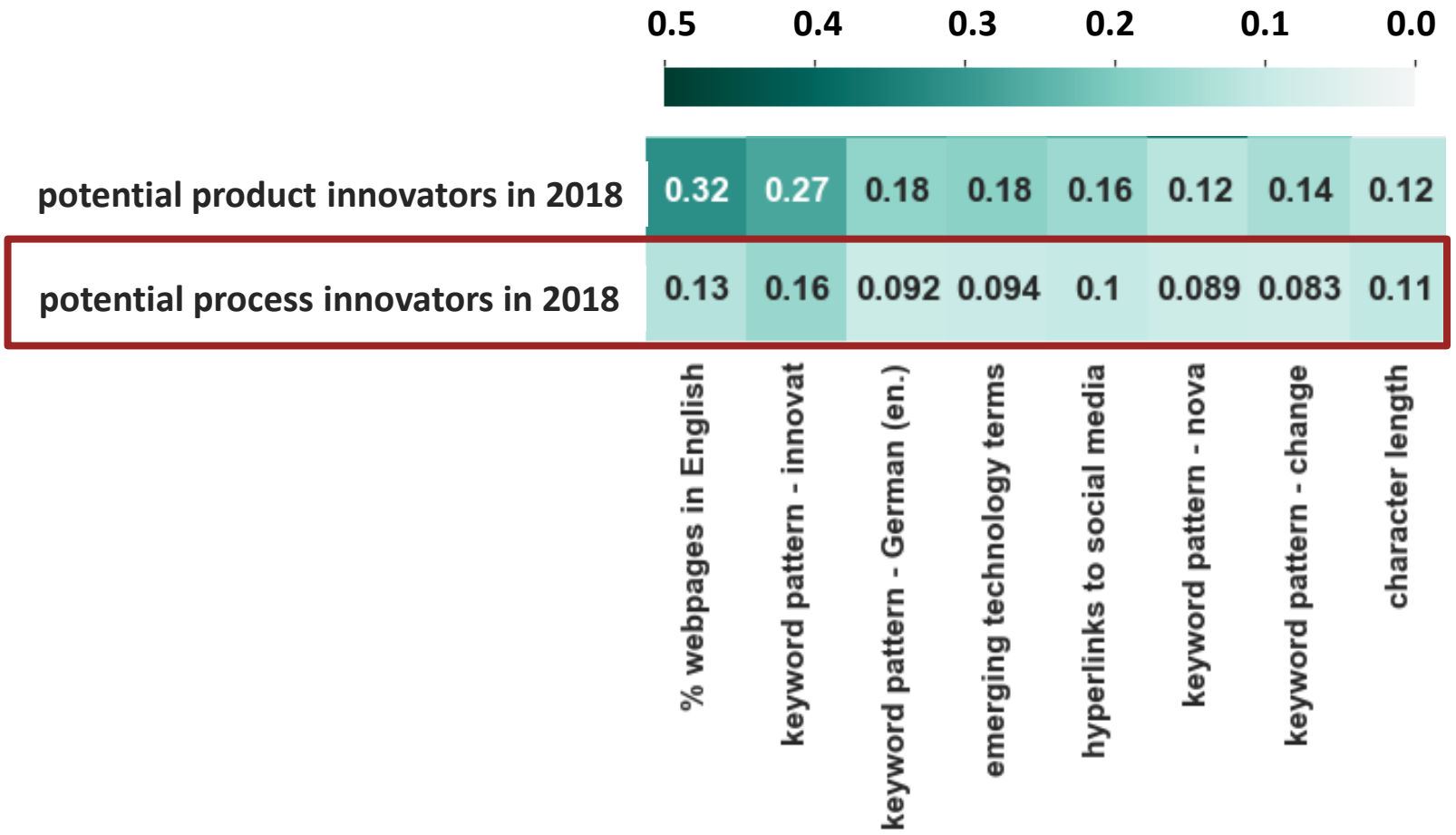
Results for website characteristics capturing *indirect information* are comparable

PREDICTING THE PRODUCT INNOVATION STATUS USING ALL IDENTIFIED WEBSITE CHARACTERISTICS COMBINED

Model	Innovative		Non-Innovative	
	Precision	Recall	Precision	Recall
random weighted guess	0.24	0.24	0.76	0.76
All website characteristics	0.62	**0.32**	0.82	0.94

- Logistic regression model with LASSO-regularization
- 10-fold Cross-validation
- Cut-off point: 0.5
- Dependent variable: potential product innovators in 2018
- Independent variables: number of occurrences of the keyword patterns nova, techno, innovat, German (en.), software, system, solution; emerging technology terms; combinations of the keyword patterns inno combined, inno count; LDA topics; number of subpages; total amount of characters; dummies indicating k-means clusters; % webpages in English language; sum of incoming and outgoing hyperlinks; hyperlinks to social media websites

WEBSITE CHARACTERISTICS AND POTENTIAL PROCESS INNOVATORS 2018
Pearson correlation coefficients



CONCLUSION

- Innovation-related keywords as well as the keyword `German‘, the share of English language, latent patterns and clusters based on website similarities, etc. are all related to firm-level product innovations.
- No single feature is able to detect a sufficient amount of innovative firms, but combinations of different website characteristics improve predictions and perform better than simple solutions.
- The innovation indicator “potential product innovators in 2018” provides downward-biased evaluation metric scores because some firms deviate and do innovate/not innovate.
- Moreover, website characteristics are more strongly related to product innovators than to process innovators.

Thank you for your attention!

REFERENCES I

- Ackland, R., Gibson, R., Lusoli, W. & Ward, S. (2010), Engaging with the public? Assessing the online presence and communication practices of the nanotechnology industry, *Social Science Computer Review* **28**(4), 443–465.
- Archibugi, D. & Planta, M. (1996), Measuring technological change through patents and innovation surveys, *Technovation* 16(9), 451 – 519. URL: *http://www.sciencedirect.com/science/article/pii/0166497296000314*
- Arora, S. K., Youtie, J., Shapira, P., Gao, L. & Ma, T. (2013), 'Entry strategies in an emerging technology: a pilot web-based study of graphene firms', *Scientometrics* 95(3), 1189–1207.
- Beaudry, C., Heroux-Vaillancourt, M. & Rietsch, C. (2016), Validation of a web mining technique to measure innovation in high technology Canadian industries, OECD Blue Sky Forum on Science and Innovation Indicators, Ghent, Belguim.
- Becker, W. & Dietz, J. (2003): R&D Cooperation and innovation activities of firms – evidence for the German manufacturing industry, *Research Policy*, 33, 209-223.
- Bersch, J., Gottschalk, S., Müller, B. & Niefert, M. (2014), 'The mannheim enterprise panel (mup) and firm statistics for germany'. URL: *https://www.zew.de/en/publikationen/zew-discussion-papers/*
- Bertschek, I. & Reinhold K. (2017), Let the user speak: Is feedback on Facebook a source of firms' innovation?, ZEW Discussion Paper No. 17-015, Mannheim.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
- Cassiman, B. & Golovko, E. (2011), 'Innovation and internationalization through exports', *Journal of International Business Studies* 42(1), 56–75.
- Choi, H. & Varian, H. (2012), 'Predicting the present with google trends', *Economic Record* 88, 2–9.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.
- Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

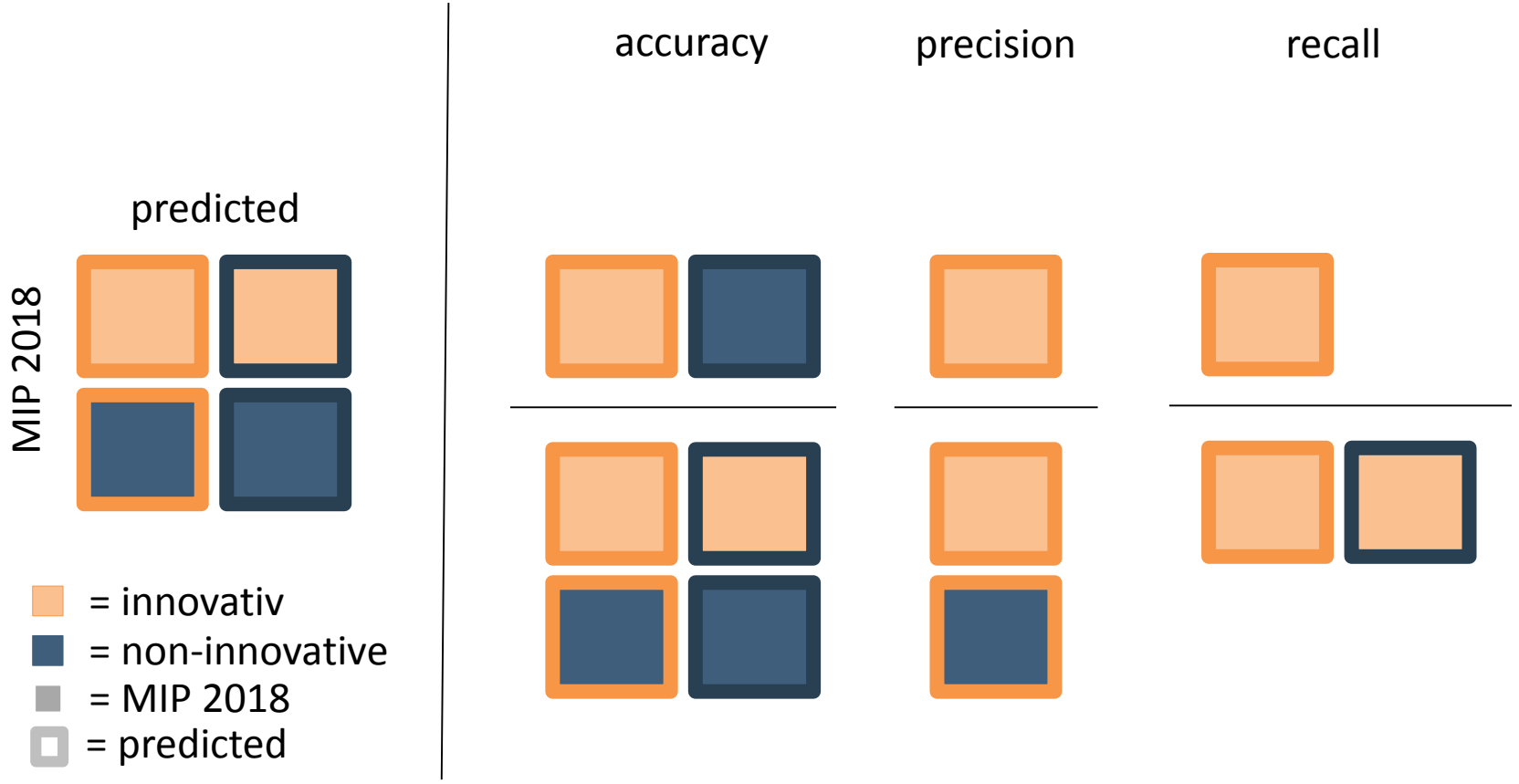
REFERENCES II

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, **1**(10). New York: Springer series in statistics.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463-484.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data', *Nature* 457(7232), 1012.
- Gök, A., Waterworth, A. & Shapira, P. (2015), 'Use of web mining in studying innovation', *Scientometrics* 102(1), 653–671.
- Hoberg, G. & Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy* 124(5), 1423–1465.
- Katz, J. S. & Cothey, V. (2006), 'Web indicators for complex innovation systems', *ReR search Evaluation* 15(2), 85–95.
- Kinne J., Axenbeck J. (2018), Web mining of firm websites: A framework for web scraping and a pilot study for Germany, ZEW Discussion Paper No. 18-033, Mannheim.
- Kinne J., Lenz D. (2019), Predicting innovative firms using web mining and deep learning, ZEW Discussion Paper No. 19-001, Mannheim.
- Kirbach & Schmiedeberg (2006): Innovation and export performance: Adjustment and remaining differences in East and West German manufacturing, *Economics of Innovation and New Technology*, 17, S. 435-457.
- Lenz, D. & Winker, P. (2018), Measuring the diffusion of innovations with paragraph vector topic models, Technical report.
- Martin, J. H., & Jurafsky, D. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson/Prentice Hall.

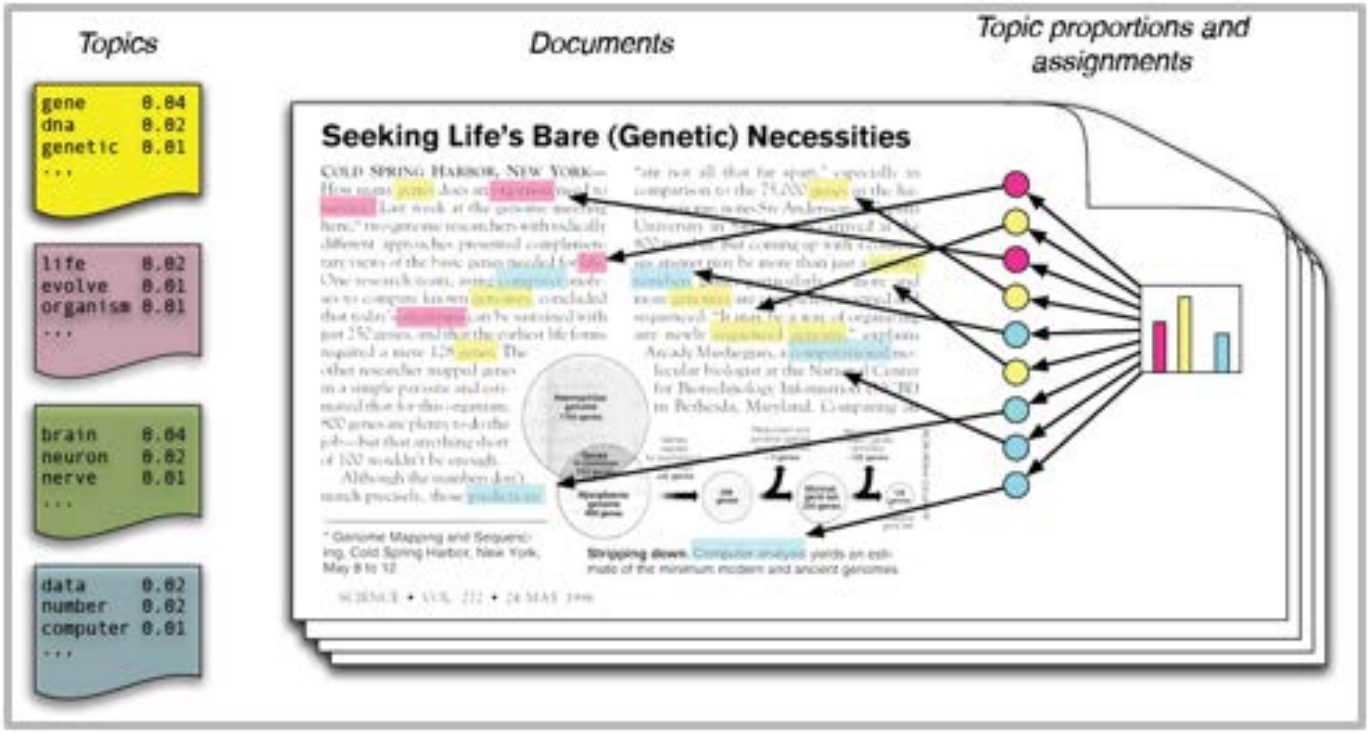
REFERENCES III

- Miner, G., Elder IV, J. & Hill, T. (2012), *Practical text mining and statistical analysis for non-structured text data applications*, Academic Press.
- Nathan, M. & Rosso, A. (2017), Innovative events, Technical report, Centro Studi Luca d'Agliano Development Studies Working Paper No. 429.
- OECD/Eurostat (2018), Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg.
- Peters, B. & Rammer, C. (2013), Innovation panel surveys in Germany, *in* 'Handbook of Innovation Indicators and Measurement', Edward Elgar Publishing, chapter 6, pp. 135– 177. URL: *https://EconPapers.repec.org/RePEc:elg:eechap:14427*6
- Shepherd, W. G. & Shepherd, J. M. (1979), *The economics of industrial organization*, Waveland Press.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation,* 28(1), 11-21.
- Rammer, C., Behrens, V., Doherr, T., Hud, M., Köhler, M., Krieger, B., ... & von der Burg, J. (2019). Innovationen in der deutschen Wirtschaft: Indikatorenbericht zur Innovationserhebung 2018. ZEW Innovationserhebungen-Mannheimer Innovationspanel (MIP), Mannheim.
- Wößmann, L. & Lachenmaier, S. (2006), 'Does innovation cause exports? Evidence from exogenous innovation impulses and obstacles using German micro data', *Oxford Economic Papers* 58(2), 317–350. URL: *https://dx.doi.org/10.1093/oep/gpi043*
- Youtie, J., Hicks, D., Shapira, P. & Horsley, T. (2012), 'Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies', *Technology Analysis & Strategic Management* 24(10), 981–995.

METRICS: ACCURACY, PRECISION AND RECALL



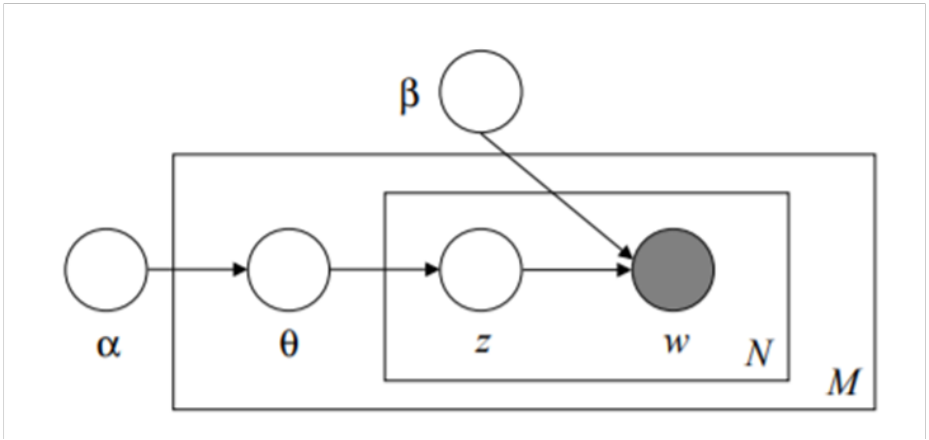
LDA – LATENT DIRICHLET ALLOCATION



Source: https://medium.com/@connectwithghosh/topic-modelling-with-latent-dirichlet-allocation-lda-in-pyspark-2cb3ebd5678e, accessed: 07.04.2019

LDA

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d.$$

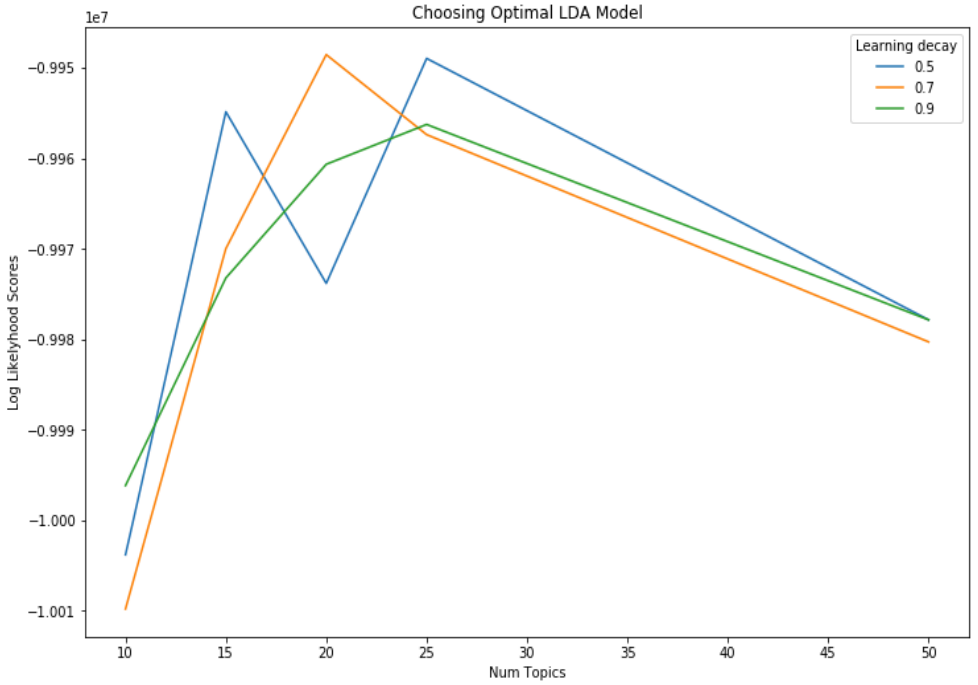


w:	word
D:	corpus
M:	collection of documents
N:	sum of words
n:	index word
d:	index document
θ	distribution of topics in document
z:	topic
α,β:	hyperparameters of Dirichlet allocation

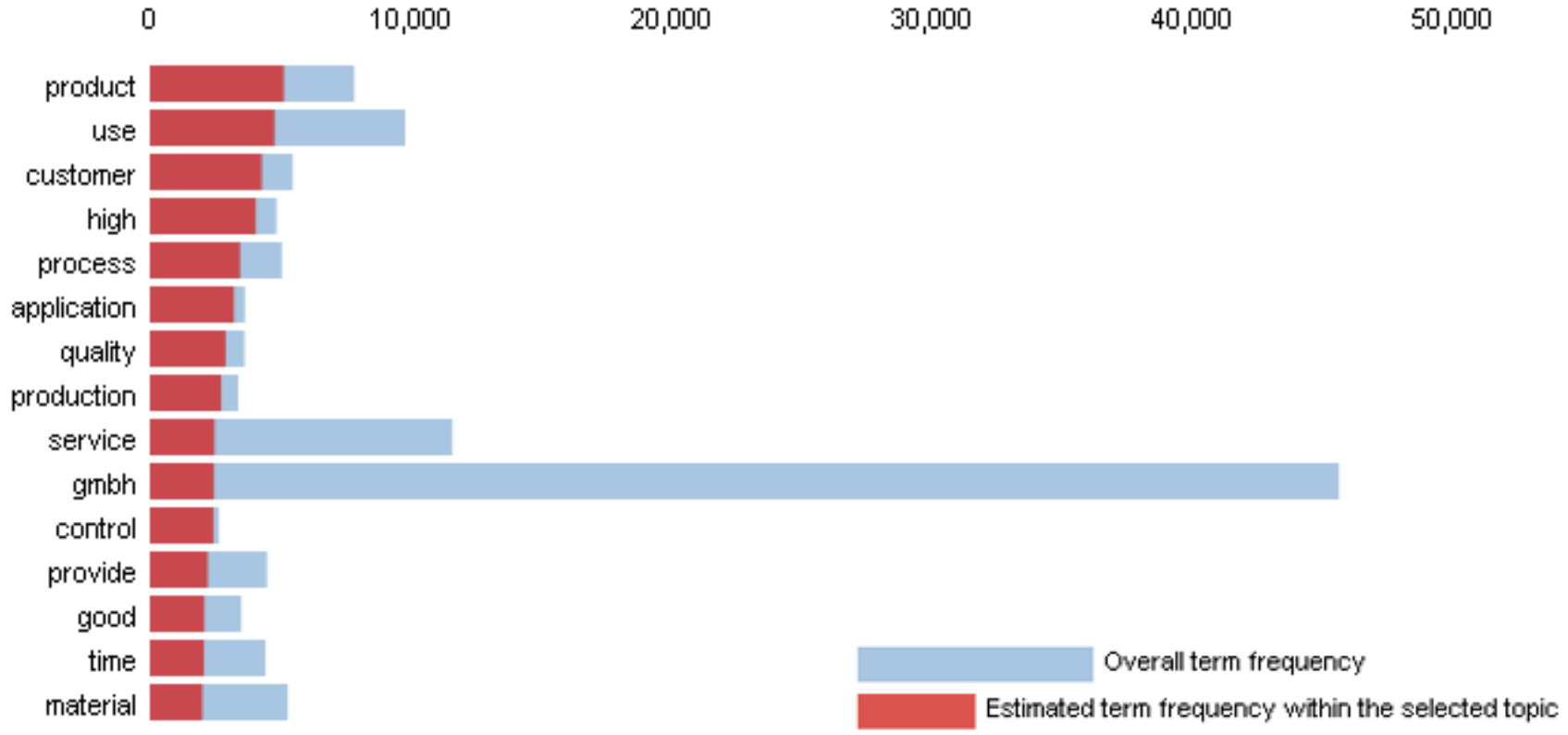
LDA

Preprocessing:

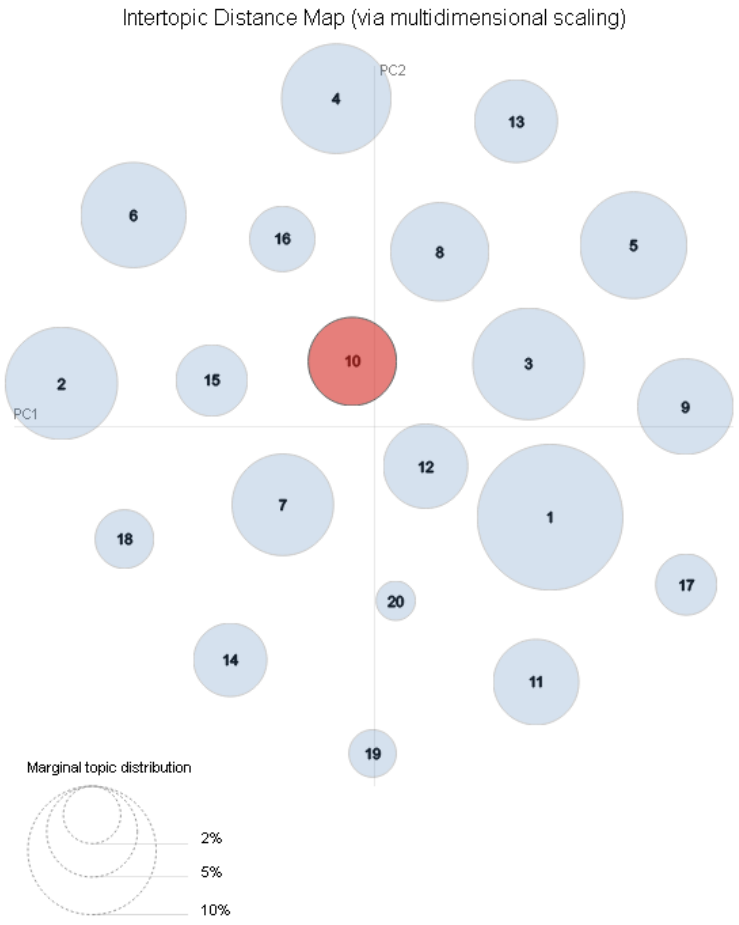
- Exclusion of all subpages mentioning in their URL terms like: *imprint, contact, location* or *general terms and conditions*
- Only words that appear at least 150 times are included
- Texts are lemmatized
- Optimal number of topics is chosen by means of a grid-search

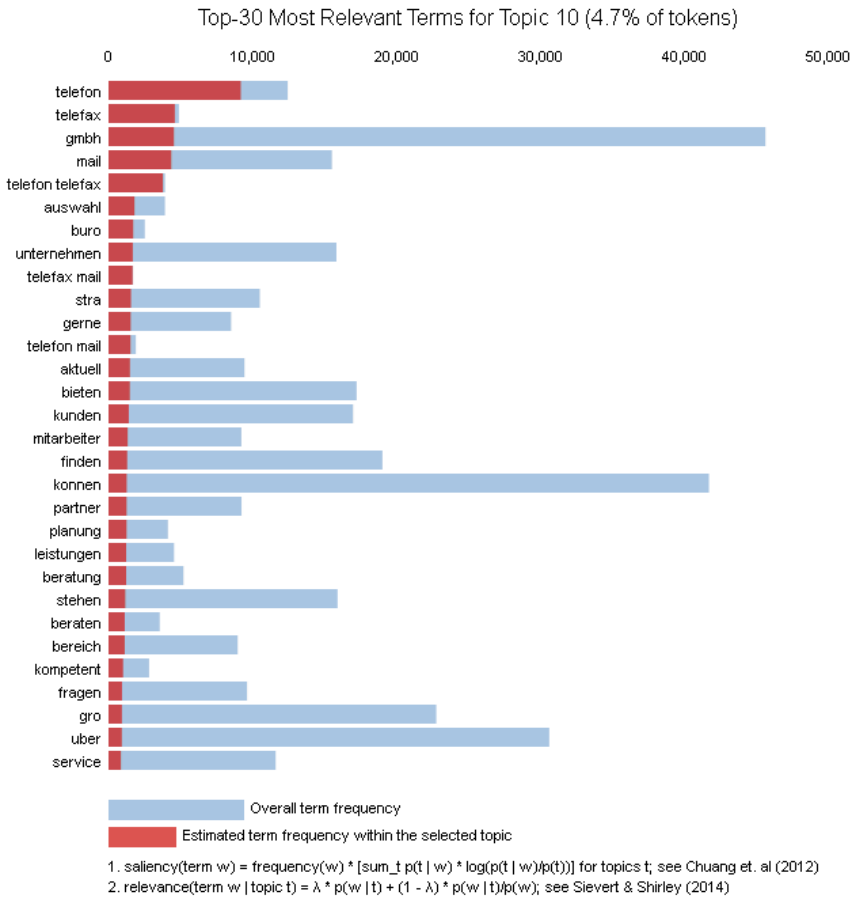


RESULTS: TOP 15 MOST SALIENT WORDS FOR THE TOPIC ‘PRODUCT DESCRIPTIONS IN ENGLISH LANGUAGE’ (EXAMPLE FOR A POSITIVELY-CORRELATED LDA TOPIC WITH POTENTIAL PRODUCT INNOVATORS IN 2018).

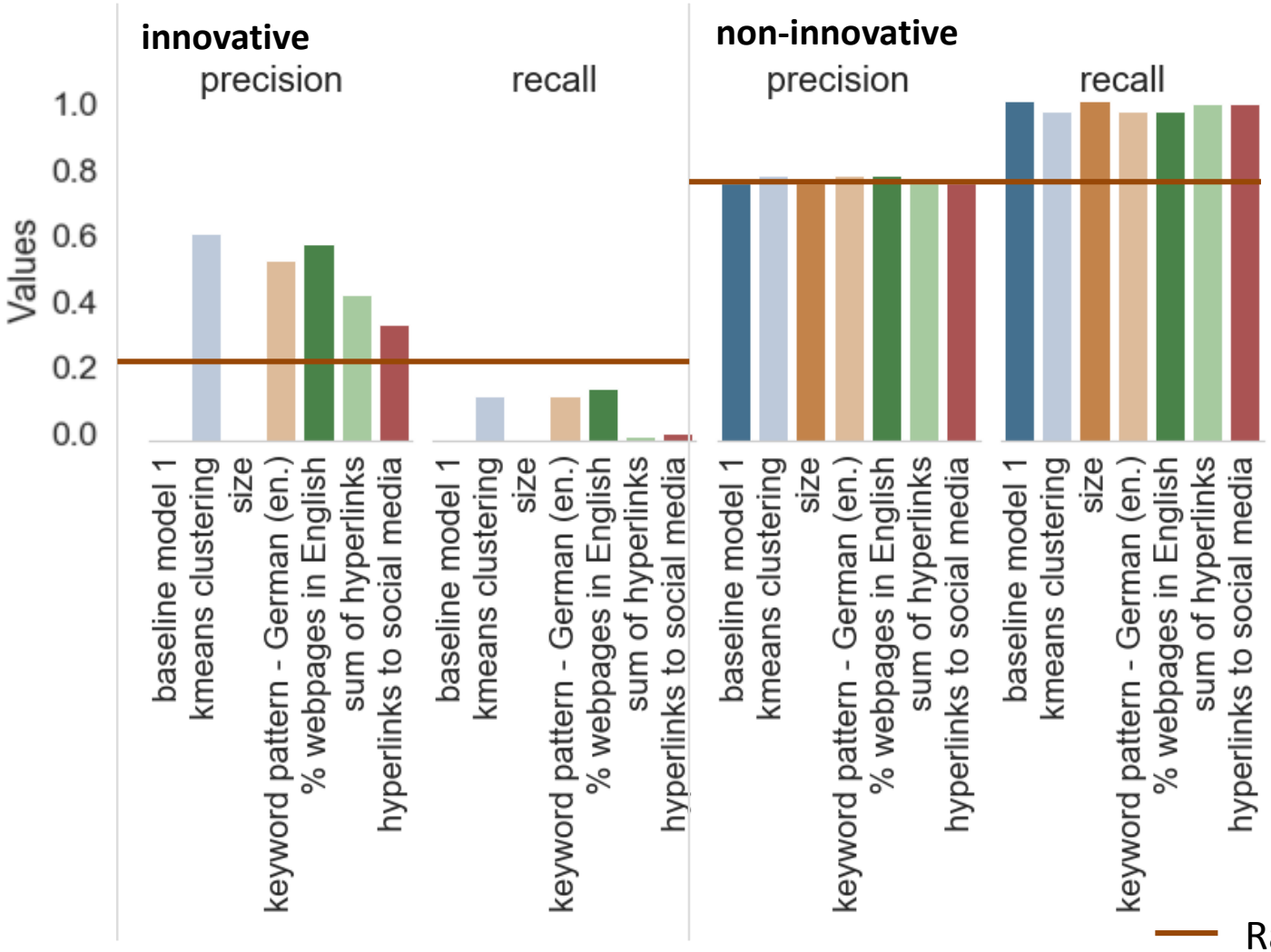


LDA – TOPIC 10 – NEGATIVE CORRELATION



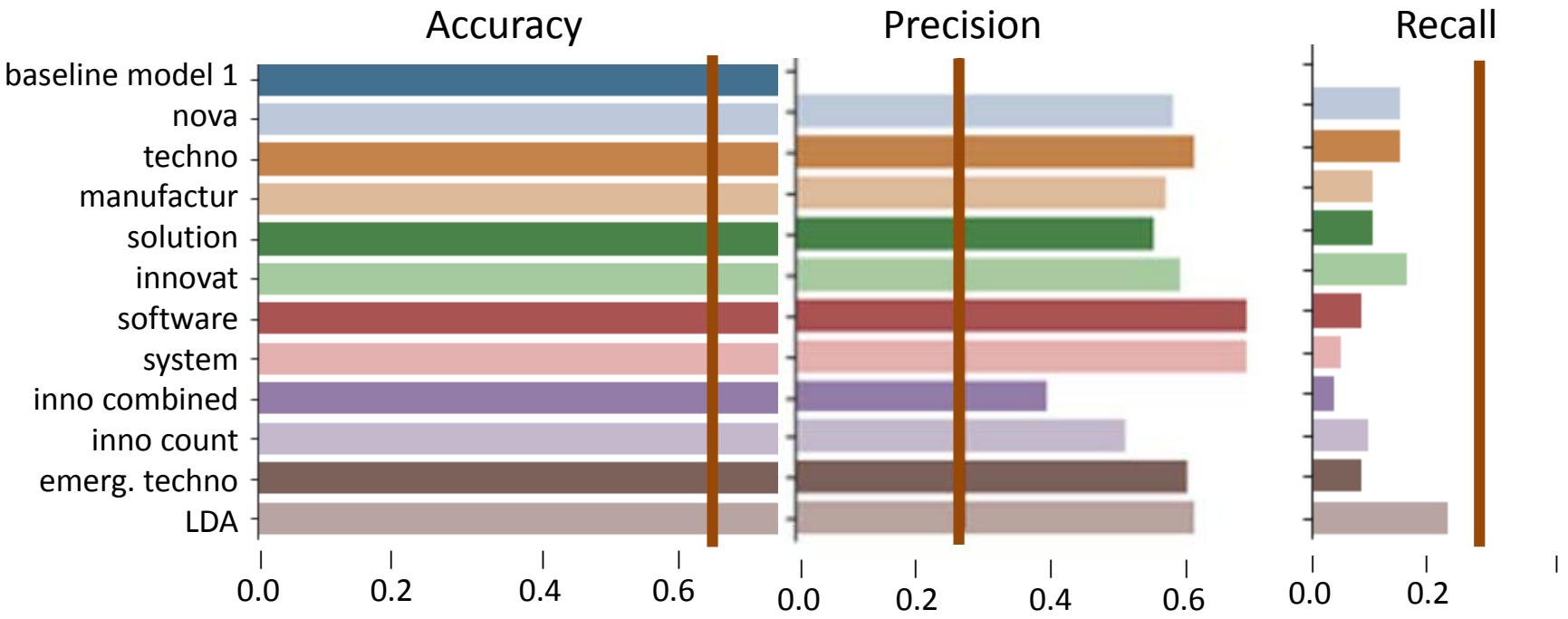


PREDICTING THE INNOVATION STATUS
WEBSITE CHARACTERISTICS - INDIRECT INFORMATION



- Logistic regression model with LASSO-regularization
- 10-fold cross-validation
- Cut-off point: 0.5
- Dependent variable: potential product innovators in 2018

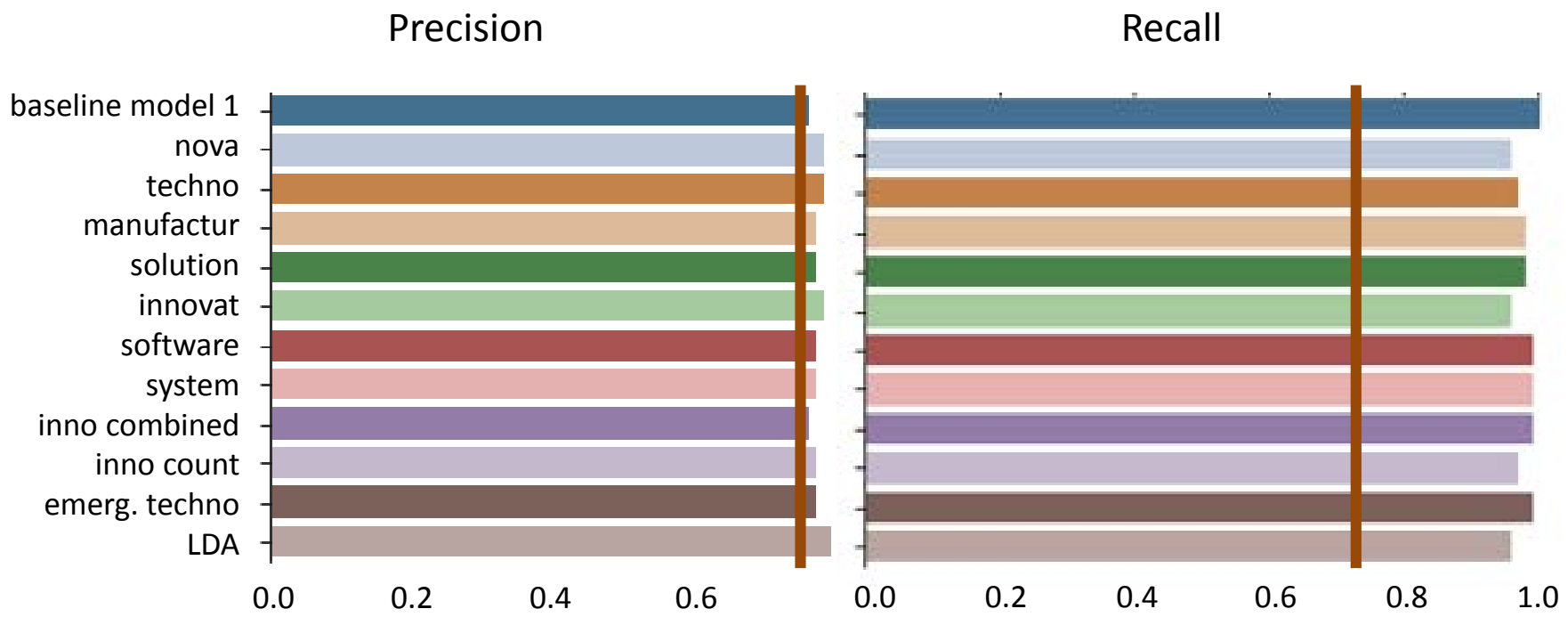
PREDICTING THE INNOVATION STATUS
WEBSITE CHARACTERISTICS - DIRECT INFORMATION
Innovative firms



Logistic regression model with LASSO-regularization
10-fold cross-validation
Cut-off point: 0.5
Dependent variable: potential product innovators in 2018

Random weighted guess

PREDICTING THE INNOVATION STATUS
WEBSITE CHARACTERISTICS - DIRECT INFORMATION
Non-innovative firms



Logistic regression model with LASSO-regularization
10-fold cross-validation
Cut-off point: 0.5
Dependent variable: potential product innovators in 2018

Random weighted guess

PREDICTING THE INNOVATION STATUS
POTENTIAL PROCESS INNOVATION IN 2018

Modell	Innovativ		Non-Innovativ	
	Precision	Recall	Precision	Recall
All website characteristics	0.75	0.06	0.80	0.99

Logistic regression model with LASSO-regularization
Cut-off point: 0.5
Dependent variable: potential process innovators in 2018
N=2334, mean of dependent variable: 0.23