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
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

<table>
<thead>
<tr>
<th>Direct Information</th>
<th>Keywords</th>
<th>Innovation-related and/or product-related terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Terms related to emerging technologies (from Wikipedia’s list of emerging technologies)</td>
<td></td>
</tr>
<tr>
<td>Latent patterns</td>
<td>Topics generated by the latent Dirichlet allocation (LDA)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indirect Information</th>
<th>Economic sector</th>
<th>Clusters based on website similarities (k-means algorithm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm size</td>
<td>Number of subpages; total amount of characters</td>
<td></td>
</tr>
<tr>
<td>International orientation</td>
<td>Percentage of subpages in English language; number of occurrences of the word ‘German’</td>
<td></td>
</tr>
<tr>
<td>Relationships to other firms</td>
<td>The sum of incoming and outgoing hyperlinks</td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td>Hyperlinks to Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr and Vimeo</td>
<td></td>
</tr>
</tbody>
</table>
**SUMMARY – RESEARCH APPROACH**

### MANNHEIM INNOVATION PANEL 2018 (MIP)

- **Answer to the question:** „*Does your firm plan to introduce a product or process innovation in 2018?*“
- **Focus on product innovation**

### MANNHEIM ENTERPRISE PANEL 2018 (MUP)

- **Website addresses**

### Automated Robot for Generic Universal Scraping (ARGUS) (Web Scraper)

- **September 2018:** Texts with preference for .de
- **January 2019:** Texts and hyperlinks with no preference

### Website characteristics

- **generate**

### 2334 German firms

- **comparison with**
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

Results for website characteristics capturing *indirect information* are comparable
PREDICTING THE PRODUCT INNOVATION STATUS USING ALL IDENTIFIED WEBSITE CHARACTERISTICS COMBINED

<table>
<thead>
<tr>
<th>Model</th>
<th>Innovative</th>
<th>Non-Innovative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>random weighted guess</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>All website characteristics</td>
<td>0.62</td>
<td>0.32</td>
</tr>
</tbody>
</table>

- 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


REFERENCES II


REFERENCES III

METRICS: ACCURACY, PRECISION AND RECALL

predicted

MIP 2018

= innovativ
= non-innovative
= MIP 2018
= predicted

accuracy

precision

recall
LDA – LATENT DIRICHLET ALLOCATION

LDA

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

- \(w\): word
- \(D\): corpus
- \(M\): collection of documents
- \(N\): sum of words
- \(n\): index word
- \(d\): index document
- \(\theta\): distribution of topics in document
- \(z\): topic
- \(\alpha, \beta\): 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

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 10 (4.7% of tokens)

1. teilen
2. teilver
3. gmbh
4. mal
5. teilen teilver
6. autoway
7. bau
8. unternehmer
9. steck
10. sba
11. gerne
12. teilen mal
13. situell
14. kosten
15. mehr
16. teilen
17. konnen
18. mehr
19. konnen
20. partner
21. planung
22. mehr
23. beratung
24. stehen
25. beraten
26. kompetent
27. fragen
28. obr
29. mehr
30. service

Marginal topic distribution

Overall term frequency
Estimated term frequency within the selected topic

1. relevance_term w = frequency(w) * ln(freq(1/p) / ln(freq(1/p)) for topics 1, see Chuang et al. (2013)
2. relevance_term w = topic_d * w*(1 - w) + w*w*(1 - w) for topics 1, see Stewart & Shirley (2014)
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

Random weighted guess
PREDICTING THE INNOVATION STATUS
WEBSITE CHARACTERISTICS - DIRECT INFORMATION
Innovative firms

Baseline model 1

Nova
Techno
Manufactur
Solution
Innovat
Software
System
Inno combined
Inno count
Emerg. techno
LDA

Accuracy

Precision

Recall

Random weighted guess

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
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
PREDICTING THE INNOVATION STATUS
POTENTIAL PROCESS INNOVATION IN 2018

<table>
<thead>
<tr>
<th>Modell</th>
<th>Innovativ</th>
<th></th>
<th>Non-Innovativ</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>All website characteristics</td>
<td>0.75</td>
<td>0.06</td>
<td>0.80</td>
<td>0.99</td>
</tr>
</tbody>
</table>

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