From Text to Insight - A Novel Approach to Measuring Business Model Innovation

Max Gabler Wanshu Jiang Christoph Kiesl Leonard Pöhls
Alexander Rieber Ansgar Scherp

I am the abstract

Table of contents

Introduction

Business model innovation (BMI) is a key activity to maintain competitiveness and even gain a competitive advantage (Pucihar et al. 2019; Teece 2018). It is therefore no surprise that the interest in BMI and methods of measuring it has grown rapidly over the last twenty years. Researchers have recently called for a BMI measurement instrument that is more comprehensive and advanced than already existing ones (Huang and Ichikohji 2023). The scale developed by Spieth & Schneider (2016) provides managers and practitioners with a measurement index for business model innovativeness. This measurement model only validates applicability of BMI theory (Huang and Ichikohji 2023) and is insufficient for longitudinal studies (Clauss 2017). Hence, this measure is not adequate for a time series analysis of BMI. Furthermore, it refers only to BMI as new-to-the-firm and is not able to grasp BMI in the sense of new-to-the-industry and new-to-market. Clauss (2017) developed a similar measure with comparable limitations.

This gap is addressed by proposing a novel approach to measuring BMI. US-based companies are obliged by the United States Security and Exchange Commission (SEC) to submit annual 10-K filings, wherein a detailed description of the company's business operations is required. Hoberg & Phillips (2016) use this filings to create word vectors about the companies products in order to cluster them into industries and thereby proposing a new industry classification. This study builds on their work and methods. We summarize these descriptions with Gemini, utilize the BERTScore as a similarity measure to calculate the similarities between companies and cluster them into industries. We compare our BERTScore clustering with the text-based network industry classification (TNIC) by Hoberg & Phillips (2016). Furthermore, we calculate the BERTScore between the summaries of different years for a single company. This approach enables the measurement of changes in the business model (BM) over time as the distance between the BM summary of one year to another. There is evidence that an increase in BMI is associated with improved firm performance (Cucculelli and Bettinelli 2015; Latifi, Nikou, and Bouwman 2021; White et al. 2022). In order to test the validity of our measure, we regress financial numbers on our measure, hoping to find a positive relationship.

- Key findings (and Contribution, already down below)

Our contribution is made in two ways. Firstly, we build on the concept of alternative industry classification put forth by Hoberg & Phillips (2016) and propose an industry classification system based on a firm's BM. Thirdly, we propose a novel measure for BMI that is sufficient for longitudinal studies.

The SEC mandates that the majority of public companies based in the United States submit specific documents in certain intervals. One such document is the annual 10-K filing. These filings follow a set order of topics and contain a range of information, including details about managerial discussions, risk factors for the company, legal proceedings and financial data. In the first section under the subtitle "Business," a company presents its general business, encompassing information about its products and services. In some instances additional topics may be addressed, such as labor issues or competition (SEC 2024). In conclusion, this section contains the most useful information for describing a company's BM (Lee and Hong 2014). Furthermore, 10-K filings are a reliable source of information, given that US law prohibits false or misleading statements in the filings. The SEC monitors the compliance of the companies with the requirements and comments where disclosure appears to be inconsistent (SEC 2024).

- paragraph 5 (robustness checks)

In spite of the growing interest in BMI and the increasing number of theoretical and empirical studies in this field, the research of BMI is still in a preliminary state (Huang and Ichikohji 2023). Consequently, there is considerable variation in the definitions of BMI, with some definitions being more similar to one another than others (Foss and Saebi 2017). Spieth & Schneider (2016) identify three core dimensions a company's BM is comprised of: its value proposition, its value creation architecture and its revenue model logic. Based on this, BMI can be conceptualized as a change that is new-to-the-firm in at least one of these dimensions. Furthermore, Spieth and Schneider (2016) introduce a measurement model to evaluate these three dimensions of BMI. They develop an index by first specifying the contents, followed by a specification of the indicators and assessing their content validity, assessing the indicators collinearity and finally assessing the external validity. A total of twelve indicators for measuring the innovativeness of the BM were identified through a comprehensive literature review and through engagement with industry practitioners. The external validity of the formative indicators was successfully validated through a survey of 200 experts in strategy and innovation management (Spieth and Schneider 2016). Clauss (2017) employs a very similar approach. After specifying the domain and dimensionality of BMI trough literature research, the author divides his scale into three hierarchical levels consisting of 41 reflective items, 10 subconstructs and three main dimensions, which are similar to the ones mentioned earlier. The scale was validated through two samples from the manufacturing industry and further demonstrated nomological validity (Clauss 2017). However, both measures are subject to three significant limitations. Firstly, both measures lack a temporal component. Consequently, they are inadequate for use in longitudinal studies or ex-post evaluations of BMI. Secondly, BMI is only measured at the new-to-the-firm level rather than at the new-to-the-industry or new-to-the-market level. Thirdly, both measures rely on interviews and questionnaires, which makes conducting large-scale studies time-consuming and reliant on the willingness of the companies to cooperate (Clauss 2017; Spieth and Schneider 2016).

The process of text mining 10-K filings is not a novel concept. Hoberg & Phillips (2016) present a novel approach to defining industry boundaries. This is achieved through the parsing of the product descriptions provided by firm 10-K filings and creating word vectors. Specifically, the authors identify and exclude proper nouns, which include common words and geographic locations.They then create word vectors for each firm and year, which enables the measurement of product similarity over time. In this way the authors demonstrate shortcomings in the traditional industry classification systems such as the Standard Industry Classification (SIC) and the North American Industry Classification System (NAICS), which are not able to account for temporal changes. The new method is capable of capturing changes in industry boundaries and competitor sets over time, thereby providing a dynamic industry classification system. In their study, Lee & Hong (2014) examine the evolution of a firm's BM over time. The authors represent each document as a vector of keywords, which is similar to the approach utilized by Hoberg & Phillips (2016). After identifying the Item 1 part of the 10-K filings as the most crucial part for describing a firm's BM, Lee & Hong (2014) filter these for relevant sentences. Subsequently, the authors construct keyword vectors, which represent the concept of the BM. Therefore, the evolution of the BM is depicted as the change in the distribution of keywords over time. Nevertheless, this approach is not without shortcomings. The authors advocate for a more robust methodology, such as incorporating multi-word phrases in the keyword vectors, to enhance the reliability of the approach (Lee and Hong 2014).

The rest of the paper proceeds as follows. Section 2 describes the preprocessing with Gemini, our data and our methodology. Section 3 compares the industry classification based on product descriptions by Hoberg & Phillips (2016) with our classification based on the BM. Section 4 lays out the BERT-model and our estimations. Section 5 discusses our results, and Section 6 concludes our study.

Data and Methodology

Preprocessing with Gemini

10-K filings are typically very large text documents, and Item 1 of these filings is no exception. Table 1 shows the deskriptive measures of the length of the original Item 1 section in our final sample. The length of a document was measured by the word count without punctuation. The document length ranges from a couple hundred words to tens of thousands. In order to utilise the entirety of the information regarding the BM in the Item 1 section and pass the text to our BERT model, we decided to let Google's GenAI chatbot Gemini summarize them to a maximum length of 512 tokens. The summaries were created between 26 June 2024 and 6 August 2024. The model employed was Gemini Flash 1.5. The prompt was inserted at the beginning of each text file and it was passed via an API to Gemini [1]. We used following prompt: "Summarize the business model from the following text. Answer with a continuous text and with five hundred twelve tokens at max. Set your focus on sources of revenue, the intended customer base, products, distribution channels and details of financing. Use only information from the following the text".[2] "intended customer base" and "product" refer to the value offering, "distribution channels" refers to the value architecture, and "sources of revenue" and "details of financing" refer to the revenue model. Consequently, this prompt covers all aspects of the definition of BMI proposed by Spieth & Schneider (2016). The term 'tokens' was used deliberately in preference to 'words', given that the number of tokens and

[1] We forked and used following Github Repository: https://github.com/skranz/gemini_ex.

[2] The spelling error in the last sentence of the prompt was found after processing Item 1. After evaluating the Summaries, this error did not cause any issues.

the number of words in a text may vary depending on the tokeniser. This way, we wanted to ensure that the whole summary is used by the BERT model. To assess the quality and accuracy of the summaries produced by Gemini, a random sample of 100 filings was selected for comparison with the original text. More precise,the original file was initially read with a focus on the points mentioned in the prompt. Subsequently, the summary was evaluated to ascertain whether it contained these same points. A list of the sample with the summaries is provided in the Appendix.

- result of this check

Table 1: Descriptive Statistics Original Filings

Year	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
2016	7842	6104	155	3705	6026	10271	51227
2017	7542	6320	155	3522	5767	9700	70611
2018	7604	6272	180	3528	5771	9669	71700
2019	8009	6631	189	3669	5971	10410	78270
2020	8660	7195	171	3943	6449	10971	57980
2021	10324	8406	235	4670	7568	13563	78799
2022	9471	7997	171	4309	7042	11897	73937
2023	6646	4771	190	3660	5814	8401	43523

The Dataset

We collect 10-Ks filings from the digital SEC Database, using the category “10-K” as extraction condition. Since the focus of our study lies on company’s BM, we merely use the Item 1 part, since this is the most crucial part of the 10-K filings for describing the companies BM (Lee and Hong 2014). Our observations are limited to an intersection of such companies, which on the one hand has been made available to the SEC since 2001 in a publicly accessible list of 10,284 companies (Appendix), of which 7,590 are currently listed on NASDAQ, NYSE or over-the-counter. We extracted 10-K filings that were submitted between 2017 and 2023 based on underlying Central Index Keys (CIK). According to [...] financial firms are not further considered. Corresponding to Table 2, multiple steps of pre-processing were required to obtain a final amount 21,683 observations. Financial key figures, including net income, total assets and others were originally extracted from the SEC, but also challenged with DataStream.

We exclude companies from the financial sector, namely companies with a SIC Code starting with six.

TODO

- Descriptive Table3 for length of summary (in words and tokens (use tokenizer our Model uses))

- Description of Table3 and the final Dataset

Methodology

Regarding our Replication H1: We expect a similar distribution of firms. H2: We expect a high overlap with Hobergs classification.

Regarding Contribution H2: Our measure for BMI shows a positive relationship with firm performance

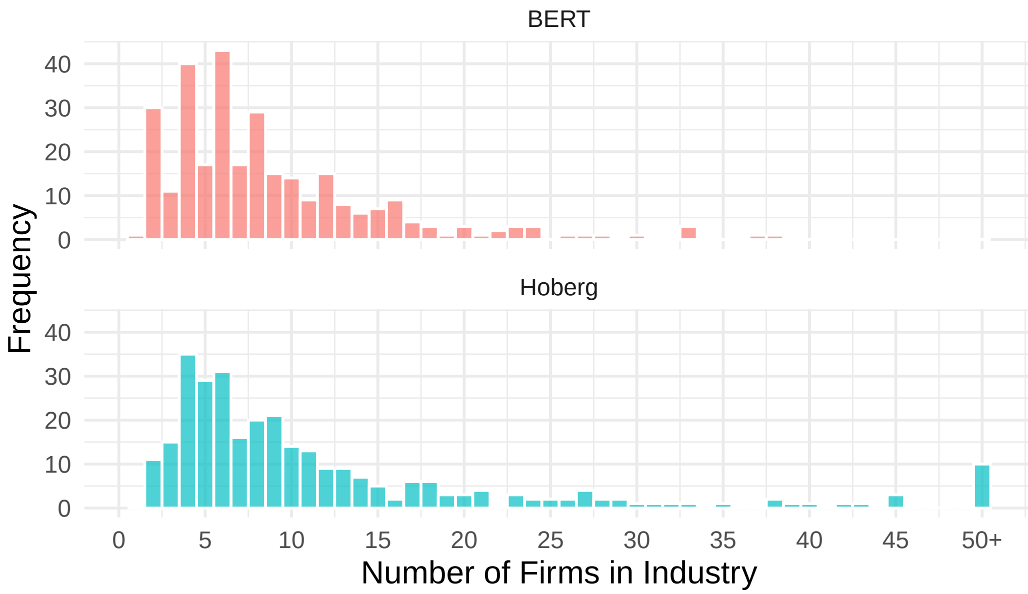
Replication

Our study builds on the idea of Hoberg & Phillips (2016) to utilize text data from 10-K filings to classify companies based on their similarity to each other into dynamic industries. Our approach differs in two ways: Firstly, instead of word-to-vec like the TNIC uses we employ BERT to represent text. Consequently, we use the BERTScore instead of the cosine similarity as our similarity measure. Secondly, we focus on the description of the BM rather than the product descriptions. In order to set a benchmark for our industry classification, we cluster companies into industries with the approach by Hoberg & Phillips (2016). Following this, we compare our industry classification to this benchmark.

- The data for the industry classification with the BERTScore is the same as described in Section 2. For the benchmark data, we use the similarity scores provided by Hoberg-Phillips Data Library.[3] The clustering described in the original study by Hoberg & Phillips (2016) is a hierarchical agglomerative clustering algorithm. -> explain in further detail as well as our replication/comparison; we use 300 industries; cal
- Results: similar picture as Hoberg & Phillips,overall left skewed distribution, most industries with less than 10 firms, but BERT classification with no industry over 36/37 firms. Our classification has more smaller industries

[3]For the database see: https://hobergphillips.tuck.dartmouth.edu.

Figure 1: Comparison between TNIC and BERT Classification



- Transition

Empirical Framework

BERT and BERTScore

Estimation Strategy

Results and Discussion

Conclusion

Acknowledgement

- Jonathan for IT Support
- Prof. Kranz for Repo

Appendix

Clauss, Thomas. 2017. "Measuring Business Model Innovation: Conceptualization, Scale Development, and Proof of Performance." *R&D Management* 47 (3): 385–403. https://doi.org/10.1111/radm.12186.

Cucculelli, Marco, and Cristina Bettinelli. 2015. "Business Models, Intangibles and Firm Performance: Evidence on Corporate Entrepreneurship from Italian Manufacturing SMEs." *Small Business Economics* 45 (2): 329–50. https://doi.org/10.1007/s11187-015-9631-7.

Foss, Nicolai J., and Tina Saebi. 2017. "Fifteen Years of Research on Business Model Innovation: How Far Have We Come, and Where Should We Go?" *Journal of Management* 43 (1): 200–227. https://doi.org/10.1177/0149206316675927.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy* 124 (5): 1423–65. https://doi.org/10.1086/688176.

Huang, WenJun, and Takeyasu Ichikohji. 2023. "A Review and Analysis of the Business Model Innovation Literature." *Heliyon* 9 (7): e17895. https://doi.org/10.1016/j.heliyon.2023.e17895.

Latifi, Mohammad-Ali, Shahrokh Nikou, and Harry Bouwman. 2021. "Business Model Innovation and Firm Performance: Exploring Causal Mechanisms in SMEs." *Technovation* 107 (September): 102274. https://doi.org/10.1016/j.technovation.2021.102274.

Lee, Jihwan, and Yoo S. Hong. 2014. "Business Model Mining: Analyzing a Firm's Business Model with Text Mining of Annual Report." *Industrial Engineering and Management Systems* 13 (4): 432–41. https://doi.org/10.7232/iems.2014.13.4.432.

Pucihar, Andreja, Gregor Lenart, Mirjana Kljajić Borštnar, Doroteja Vidmar, and Marjeta Marolt. 2019. "Drivers and Outcomes of Business Model Innovation—Micro, Small and Medium-Sized Enterprises Perspective." *Sustainability* 11 (2): 344. https://doi.org/10.3390/su11020344.

SEC. 2024. "Investor Bulletin: How to Read a 10-K." https://www.sec.gov/files/reada10k.pdf.

Spieth, Patrick, and Sabrina Schneider. 2016. "Business Model Innovativeness: Designing a Formative Measure for Business Model Innovation." *Journal of Business Economics* 86 (6): 671–96. https://doi.org/10.1007/s11573-015-0794-0.

Teece, David J. 2018. "Business Models and Dynamic Capabilities." *Long Range Planning* 51 (1): 40–49. https://doi.org/10.1016/j.lrp.2017.06.007.

White, Joshua V., Erik Markin, David Marshall, and Vishal K. Gupta. 2022. "Exploring the Boundaries of Business Model Innovation and Firm Performance: A Meta-Analysis." *Long Range Planning* 55 (5): 102242. https://doi.org/10.1016/j.lrp.2022.102242.