Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 15-2018

David Lenz and Peter Winker

Measuring the Diffusion of Innovations from Technology News using Topic Modeling

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Measuring the Diffusion of Innovations from Technology News using Topic Modeling

David Lenz[a,1] **and Peter Winker**[a,1]

[a]Department of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

This manuscript was compiled on May 2, 2019

Measuring the diffusion of innovations from textual data sources besides patent data has not been studied extensively. However, early and accurate indicators of innovation and the recognition of trends in innovation are mandatory to successfully promote economic growth through technological progress via evidence-based policy making. In this study, we propose Paragraph Vector Topic Model (PVTM) and apply it on technology related news articles to analyze innovation related topics over time and gain insights regarding their diffusion process. PVTM represents documents in a semantic space, which has been shown to capture latent variables of the underlying documents, e.g. the latent topics. Clusters of documents in the semantic space can then be interpreted and transformed into meaningful topics by means of Gaussian mixture modeling. Using PVTM we identify innovation related topics from 170 thousand technology news articles published over a span of 20 years and gather insights about their diffusion state by measuring the topics importance in the corpus over time. Thereby, we find that PVTM diffusion indicators for certain topics are Granger causal to Google Trends indices with matching search terms. Further, our results suggest PVTM is well suited to discover latent topics in (technology related) news articles and that the diffusion of innovations could be assessed using topic importance measures derived from PVTM.

Topic Model | R&D | R&I | STI | Innovation | Indicators | Text Mining | Natural Language Processing | NLP |

JEL Classification: O30, C81, C83

The rapidly growing amount of digital information provides novel data sources for economic analysis, e.g., with regard to identifying and measuring innovation trends. In order to exploit this valuable information, there is a growing need for automated information retrieval from large text corpora (1). Meanwhile, great progress has been made in machine learning and neural network theory that led to the emergence of new methods to extract high quality indicators from text. However, the opportunities created by the ongoing digitization have not been fully acknowledged yet nor have they been extensively studied in the economic literature. Following (2), who suggested that machine learning methods should be more widely known to and used by economists, more research needs to be conducted in the fields of innovation economics that appropriately incorporates these new data sets and methods. For a few early exceptions see the broader economic use cases discussed below.

In natural language processing (NLP), topic modeling describes a set of methods to extract the latent topics from a collection of documents. The results yielded by topic models are typically twofold, 1) a list of topics, where each topic is associated to certain words that are especially relevant in the context of the topic, and 2) a document-topic matrix, where every document in the text corpus is assigned with a probability of belonging to each of the topics determined in 1). Topic models have been applied to extract stock related topics from financial news with an application to predict abnormal returns (3), to categorize 8-k fillings and determine which topics are associated to abnormal returns (4), to measure the novelty of financial news (5) and to determine the key issues faced by firms (6). It has been analysed how monetary policy is affected by increased transparency (7) and how topics found in a Norwegian business newspaper can be used to model the impact of news on the business cycle (8). Scientific discourse modeling is also within the scope of topic models, for example (9) identify topics in the Journal of Economics and Statistics to study whether or not the scientific discussion of topics correlates with the actual development of economic key indicators. (10) examine how central bank communications affect real economic variables using topic modeling.

We propose a topic model architecture based on neural embedding methods, which is able to generate meaningful and coherent topics. In particular, we use Paragraph Vector (also known as Doc2Vec, (11)) to compute vector space representations of text documents and Gaussian mixture models (GaussMM) to cluster the resulting document vectors into meaningful semantic topics. We call this combination of embedding and clustering Paragraph Vector Topic Modelling (PVTM). The applicability of the approach is demonstrated on a corpus of news articles from the German IT-publisher *heise news* from the last 20 years.

Our results suggest that PVTM is well suited for topic modeling this type of text data. It became clear that the topic probabilites generated by PVTM might serve as a proxy to measure the diffusion of certain types of innovations.

The remainder of this article is organized as follows. Section 1 details the methodological background for our analysis. The news ticker dataset and the experimental design are reviewed in Section 2. In Section 3 we discuss our findings. Section 5 summarizes our results and describe avenues for future research.

1. Paragraph Vector Topic Modeling

The following section introduces the PVTM methodology to generate topics and topic membership probabilities. Doc2Vec borrows the main ideas from Word2Vec, therefore it is useful to discuss the Word2Vec mechanics for encoding single words

[1]To whom correspondence should be addressed. E-mail: david.lenz@wi.jlug.de or peter.winker@wirtschaft.uni-giessen.de

before detailing Doc2Vec. In the last part of this section the document clustering algorithm is discussed.

A. Neural Embeddings of Words and Documents. Neural network based embedding methods like Word2Vec (12, 13) play an increasingly vital role for encoding the semantic and syntactic meaning of words based on their context. Word2Vec builds low-dimensional dense vector space representations which encode the syntactic and semantic meaning of a word in a given context. As similar words tend to appear in similar contexts (14), encoding words based on their local context captures interesting properties in the resulting vectors, which have been shown to represent the way in which we use these words. Intuitively, words that share many contexts are more similar than words that share fewer contexts.

Vector calculations on the resulting word vectors yield useful results, for example $v_{Paris} - v_{France} + v_{Italy} = v_{Rome}$ (12), where v_{word} is the vector representation for a given word. These meaningful representations of words can be used as features in a variety of NLP tasks, including topic modeling. Word2Vec comes in two architectural variants: Skip-Gram (SG) and Continuous Bag of Words. We discuss the SG architecture in more detail, as this is relevant for our application of the Doc2Vec model.

Skip-Gram. During model training, the SG architecture iterates over a given text corpus in fixed-sized sliding windows and generates (*context / target*) word pairs, where q target words are considered on both sides of the context word, which is the word in the middle. Assume the following 5-word window ($q = 2$): *Innovation is good for business.* The context word w_c would be *good*, and the target words $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ are *Innovation, is, for*, and *business.* This results in 4 *(context / target)* pairs, *(good / Innovation)*, *(good / is)*, *(good / for)* and *(good / business)*.

Given such pairs of context and target words resulting from the sliding window for a document, the construction of the vector representation is done as follows. First, each word in the vocabulary is represented as a one-hot encoded sparse vector v_{word} of size $V \times 1$, where V is the number of different words in the vocabulary. In a one-hot encoding scheme, a 0 indicates the absence of a word while a 1 indicates the presence of a word. Thus, $\mathbf{v}_{word}$ will be a sparse vector consisting of zeros and a single 1 for the row corresponding to the word.

Second, for each context word $\mathbf{v}_c$ a lower dimensional (denser) vector representation $\mathbf{d}_c$ is obtained using a projection matrix $\mathbf{W_c}$ of dimension $D \times V$, where $D \ll V$:

$$\mathbf{d}_c = \mathbf{W_c}\mathbf{v}_c \,. \qquad [1]$$

Equation (1) extracts the column of $\mathbf{W}_c$ corresponding to the non-zero entry in $\mathbf{v}_c$ to use it as a vector representation $\mathbf{d}_c$ for word v_c in a D-dimensional vector space.

The same procedure is repeated for each target word using a different projection (weight) matrix $\mathbf{W}_t$ of size $D \times V$ and the one-hot representation of the target word $\mathbf{v}_t$ resulting again in a D-dimensional vector $\mathbf{d}_t$, which is the vector representation of the target word:

$$\mathbf{d}_t = \mathbf{W_t}\mathbf{v}_t \,. \qquad [2]$$

The probability of finding a target word t close to the context word c is modelled based on the similarity between the corresponding vectors $\mathbf{d}_t$ and $\mathbf{d}_c$ in the low dimensional vector space. The dot product of both vectors $s = \mathbf{d}_t'\mathbf{d}_c$ is used as a measure of similarity, which is closely related to the cosine similarity, but also takes into account the lengths of both vectors. Then, the probability that a target word t is observed close to context word c is defined making use of the softmax function (15):

$$p(t|c) = \frac{e^s}{\sum e^s}, \qquad [3]$$

where the sum in the denominator is over all V different words in the text, i.e. all potential target words.

The projection matrices $\mathbf{W_c}$ and $\mathbf{W_t}$ are obtained by maximizing the joint (log-)probability of observing the target words for all context words in the corpus:

$$\sum_{j=1}^{R}\sum_{k=1}^{n_j}\sum_{l=\max(k-q,1),l\neq k}^{\min(k+q,d_j)} \log(p(v_l^j|v_k^j)) \qquad [4]$$

$C = \{D_1, \ldots, D_R\}$ is the corpus of R news articles, where each article consists of several words, i.e. $D_j = \{v_1^j, \ldots, v_{n_j}^j\}$. The number of words per article D_j is denoted by n_j, and the size of the corpus is denoted by R. Thus, given an article D_j, v_k^j is the current context word used to predict target word v_l^j, that lies in a given range around the context word. The size of the context window is denoted by q. Gradient descent is used to iteratively update the weights in $\mathbf{W}_c$ until some convergence criteria is met. After training the weights in $\mathbf{W}_c$ act as a lookup table for the vector representations of words.

B. Paragraph Vector (Doc2Vec). Doc2Vec expands the ideas of Word2Vec to longer pieces of texts. Instead of word vectors, document vectors are learned during the training process. The resulting vector representations have been shown to capture latent semantic properties of the text fragments, for instance the underlying semantic topic of a document. Doc2Vec also comes in two variants: Distributed Bag of Words (DBOW) and Distributed Memory (DM). In our experiments we focus on the DBOW methodology, as it has been shown to produce slightly better results compared to DM.*

Doc2Vec-DBOW builts upon the Word2Vec-SG architecture but replaces the center word with a unique document ID. Thus, instead of conditioning a single word on its surrounding words, the whole document is conditioned on the words appearing in this document.

C. Gaussian Mixture Clustering. Topic modeling uses clusters of important words to define topics, where different topics may share some words. In our approach this is done by clustering the vector representations obtained from Doc2Vec and then determining the most relevant words per cluster. A Gaussian mixture model (GaussMM, see for example 17) is a parametric probability density function represented as a weighted sum of Gaussian component densities (18, p. 827). Gaussian mixture models employ the expectation maximization algorithm (19) to fit a mixture of Gaussian models to a given dataset and can be used to represent normally distributed subpopulations within an overall population.

GaussMMs have been used to track multiple objects in video sequences (20), to extract features from speech data

*Though (11) report that the DM architecture seems to perform better, subsequent research came to different conclusions (16).

(21) and for speaker verification (22). Compared to frequently used clustering techniques such as k-means (23) or mean-shift (24), GaussMMs offer the advantage of soft-clustering the data. Soft clustering allows multiple cluster memberships per document, so each document can be represented as a probability distribution over the cluster memberships. The result of the process is a matrix with one row per document and one column per identified cluster, where each entry represents the probability of belonging to a certain cluster. Given that Doc2Vec captures latent topics in the corpus, it is reasonable to suggest that clustering the resulting document vectors can be seen as identifying latent topics.

Particularly, given a D-dimensional vector representation $\mathbf{d}_r$ of a news article D_r, and a pre-set number of Gaussian components M with mixture weights w_i, the Gaussian mixture model is defined as a weighted sum over the M Gaussian components, where the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$:

$$p(\mathbf{d}_r|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{d}_r|\mu_i, \Sigma_i) \quad [5]$$

Thereby, each mixture component $g(\mathbf{d}_r|\mu_i, \Sigma_i)$, $i = 1, \cdots, M$ is defined as a D-variate Gaussian function of the form

$$g(\mathbf{d}_r|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{d}_r - \mu_i)^T \Sigma_1^{-1}(\mathbf{d}_r - \mu_i)\right\} \quad [6]$$

with μ_i and $\mathbf{\Sigma}_i$ representing the mean vector and covariance matrix respectively, and λ collects all parameters of all mixture componentes, i.e.

$$\lambda = \{w_i, \mu_i, \Sigma_i | i = 1, \cdots, M\} \quad [7]$$

The optimal parameter configuration λ is estimated through iteratively updating the model components to best fit the training data using the EM algorithm. $p(\mathbf{W}_c|\lambda) = \prod_{r=1}^{R} p(d_r|\lambda)$ is the GaussMM likelihood given the training data $W_c = (\mathbf{d}_1 \dots \mathbf{d}_R)$. Starting with an initial configuration λ_{old}, a new configuration λ_{new} is computed such that $p(\mathbf{W}_c|\lambda_{new}) \geq p(\mathbf{W}_c|\lambda_{old})$ for R training vectors collected in $\mathbf{W}_c$. The initial configuration is computed using k-means. The mixture components are updated according to

$$w_i^{new} = \frac{1}{R}\sum_{r=1}^{R} p(i|\mathbf{d}_r, \lambda_{old}) \quad [8]$$

$$\mu_i^{new} = \frac{\sum_{r=1}^{R} p(i|\mathbf{d}_r, \lambda_{old})\mathbf{d}_r}{\sum_{r=1}^{R} p(i|\mathbf{d}_r, \lambda_{old})} \quad [9]$$

$$\sigma_i^{2new} = \frac{\sum_{r=1}^{R} p(i|\mathbf{d}_c^r, \lambda_{old})\mathbf{d}_r^2}{\sum_{r=1}^{R} p(i|\mathbf{d}_c^r, \lambda_{old})} - (\mu_i^{new})^2, \quad [10]$$

where Eq. (8) is the update for weight w_i, the means are updated according to Eq. (9) and Eq. (10) details the variance re-estimation, in this case for a diagonal covariance. The a posteriori probability for component i is given by

$$p(i|\mathbf{d}_r, \lambda) = \frac{w_i g(\mathbf{d}_r|\mu_i, \mathbf{\Sigma}_i)}{\sum_{i=1}^{M} w_i g(\mathbf{d}_r|\mu_i, \mathbf{\Sigma}_i)} \quad [11]$$

The downside of GaussMMs is that the number of mixture components M needs to be specified beforehand, and the algorithm is always going to use all M components. This gives rise to the need of external validation methods. One way of evaluating the quality of a given GaussMM clustering is to use theoretical criteria like the Bayesian Information Criterion (BIC, 25), which is the approach we take.

2. Discovering innovation related topics from news articles

We apply PVTM to news articles in an effort to discover innovation related topics and measure their diffusion by means of topic probabilities over time.

A. Technology related news corpus. The data set is formed by news articles published by the German IT-magazine *Heise*[†] in their news-ticker archive from 1997 to 2016. The total number of news articles is 174, 532, resulting in an average of 8727 articles per year. However the number of articles per year before the 2000s was considerably lower compared to subsequent periods. The average news article consists of 278 words. Figure 1 details the number of documents per year and the number of words per document.

Fig. 1. Text Corpus Statistics

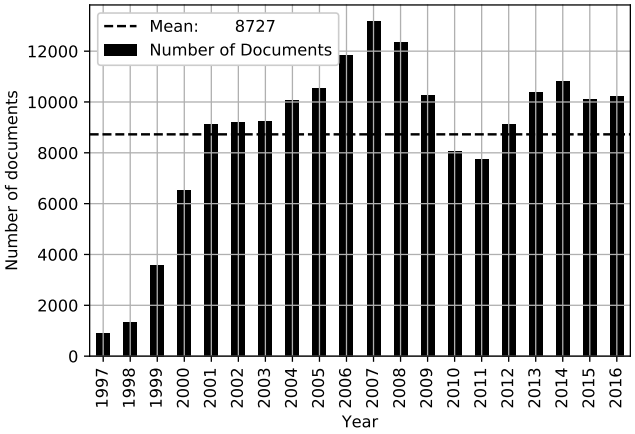
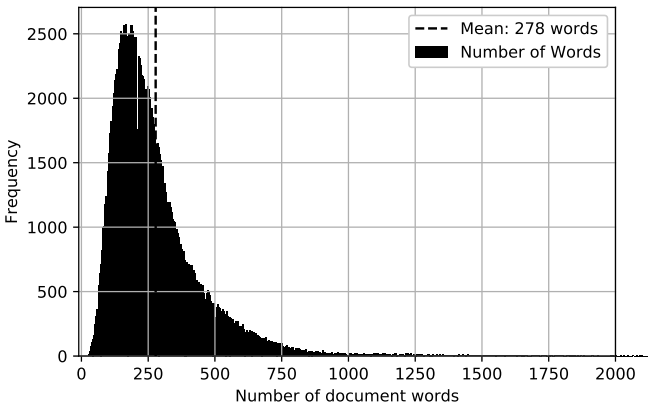


Fig. 2. Text Corpus Statistics



[†]https://www.heise.de/newsticker/

B. Data Preprocessing & Parameter Settings. We removed all non-alphanumeric characters and lowercased the resulting words. Next, we apply popularity based pre-filtering, which is a commonly applied technique in recommendation systems (26) This can be seen as removing corpus-specific stop words from the vocabulary by setting an upper and lower threshold, defining in how many documents a word can occur before it is considered to be too frequent or infrequent. Thus, all words that occur in more than 65% of documents and in less than 0.5% of documents are filtered, as these words appear to be to common/rare to be considered useful for topic modeling. The thresholds have been determined empirically, i.e. they have been adjusted until all corpus-specific stop words were removed.

The algorithm[‡] to compute vector respresentations from news articles is run for 10 epochs, where each epoch consists of going over all articles once. Following the default settings, we choose the dimensionality of the document vectors to be 100. We used the BIC to find the optimal number of Gaussian mixture components and the best approach to construct the covariance matrices Σ_i. Four methods are considered to estimate the covariance matrix, i.e. *Full* uses the full covariance matrix for each individual component, i.e. each cluster can experience any shape, while *Diagonal* only uses the diagonal of the covariance matrix per component, resulting in cluster shapes that are orientated along the coordinate axes. *Tied* uses a genereal covariance matrix for all mixture components, thus all clusters have an identical shape. *Spherical* makes use of a single variance per component instead of a covariance matrix, resulting in spherical cluster forms in higher dimensions

All possible combinations of K and Σ_i for $K \in \{50, 1000\}$ and $\Sigma_i \in \{Diagonal, Full, Spherical, Tied\}$ have been tested.

From there we used a two step procedure to find the optimal model parameters. We first iterated over the parameter space of K in steps of 50 and starting from 50 i.e. $50, 100, \ldots, 1000$. At every step during the parameter optimization procedure, all four options to construct the covariance matrices have been evaluated. The best result K^* was then used to construct a smaller search space $K \in \{K^* \pm 50\}$, which was searched in steps of 5. The optimal number of Gaussian mixture components K (=topics) was found to be 675 after this run which we kept as the final number of clusters. Eventually, the covariance matrices Σ_i of the GaussMM are constrained to be *Diagonal* as this resulted in the lowest BIC scores for the data set at hand.

3. Approximating the adoption of innovations from innovation related topics

The OSLO manual (29) defines innovation as the implementation of a new or significantly improved product or process. Diffusion can be described as the process by which an innovation is communicated through certain channels over time among the members of a social system (30). The diffusion curve is typically drawn as a hamp shaped line as shown in Figure 3.

The diffusion of new ideas consists of four main elements: (1) an innovation (2) that is communicated through certain channels (3) over time (4) among the members of a social system. Thereby, mass media outlets are effective tools to create knowledge about innovations.(30) In our approach, the communication channel is the internet via a media outlet and the social system consists of the users of the media outlet.

Using topic modeling, we attempt to measure the diffusion of topics related to innovations by aggregating topic probabilities in large technology related news corpora over time. In a large news corpus, even though it is technology news, certainly not every topic is related to an innovation. However, if a topic only becomes relevant after a certain point in time it can be interpreted as representing something novel that does not align with previously available categories. Newly emerging topics can thus often be interpreted as being related to new and innovative processes. This way, conclusions about an innovation regarding adoption time and status can be drawn based on the aggregated probabilities for this topic in the corpus. We exemplary present three topics that are related to innovative activities. In particular, we use the evolution of their weights over time as an approximation of the diffusion process of topic relevance.

The topics related to innovative activities are visualized in Figures 4 - 6. Thereby, the TABLET topic in Figure 4 represents a product innovation, the WIKIPEDIA topic in Figure 5 refers to a process innovation and the VIRTUAL REALITY topic in Figure 6 can be described as being both, product and process innovation.

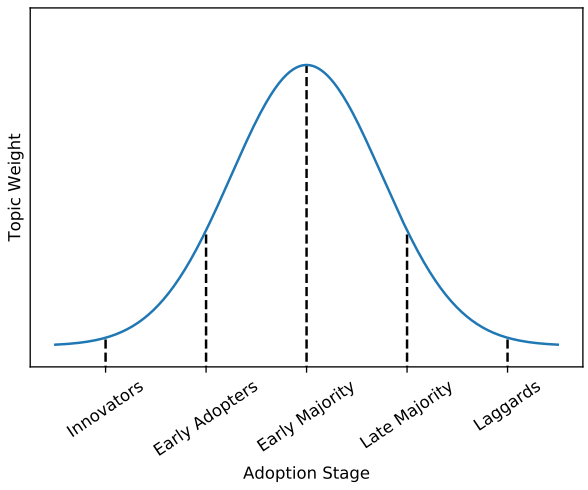
Each figure consists of two panels. The left panel exhibits a wordcloud representing the most relevant words in the topic excluding stopwords[§]. We measure the importance of a word to a topic by means of cosine similarity between the topic vector and the respective word vectors. This importance is reflected by the font size of the words in the wordcloud.

The right panel in each figure shows a line plot of the evolution of topic weights over time. We differentiate two dimensions of importance for a topic. Given a certain time interval, we quantify 1) the probability, that a topic appears in the text corpus and 2) the number of articles that are hard assigned to a topic. The hard assignment is done using

[‡]The gensim package (27) was used to generate document embeddings with Doc2Vec. To cluster the constructed document vectors we used the GMM implementation from Sklearn (28).

[§]The list of stopwords is compiled as follows. We use stopwords from the python package `stop-words` (https://pypi.python.org/pypi/stop-words), add stopwords from the Natural Language Toolkit (NLTK, 31) and also include a list of common stopwords from https://github.com/6/stopwords-json.

Fig. 3. The Diffusion of Innovations



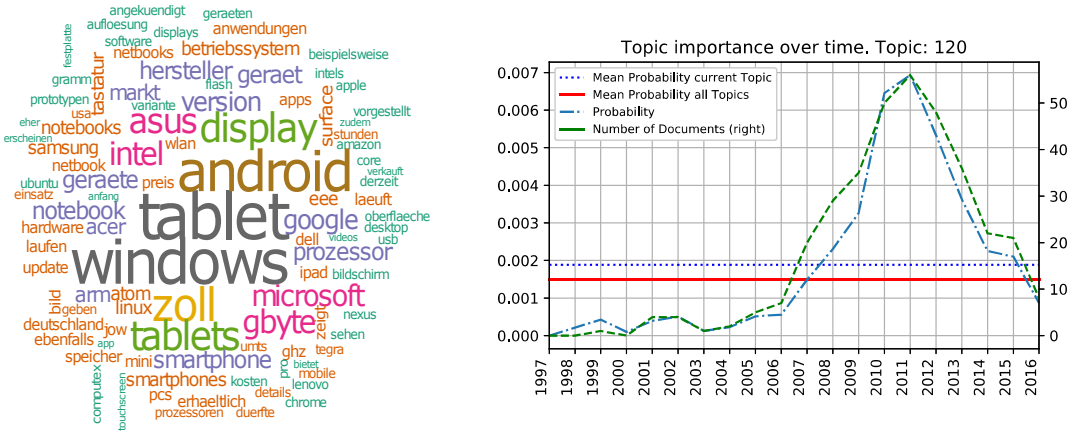


Fig. 4. Topic TABLET *Left*: Most relevant words. *Right*: Importance over time.

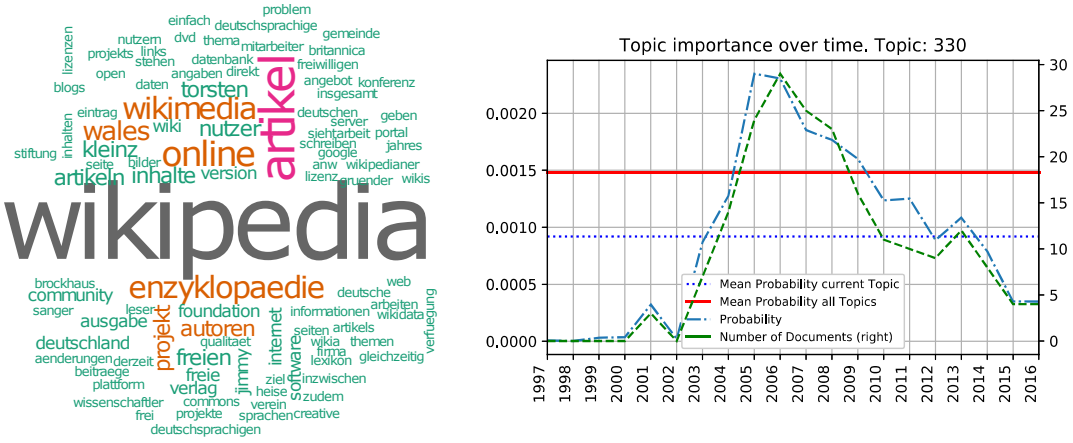


Fig. 5. Topic WIKIPEDIA *Left*: Most relevant words. *Right*: Importance over time.

a mechanism where articles are allocated according to their highest membership probability, therefore in this setting each article can only belong to one topic T_i. Given a text corpus C at a certain time frame t, i.e. C_t, the probability for a single topic $p(T_i|C_t)$ is defined as the sum of the topic probabilities for all articles in the corpus in that time frame, weighted by the overall number of articles in period t, D_t.

$$p(T_i|C_t) = \frac{\sum_{d \in C_t} p(T_i|d)}{D_t} \quad [12]$$

The horizontal lines display the average probability of the topic and the average probability of all topics for comparison.

4. Discussion

The first of the exemplary topics represented by the wordcloud in Figure 4 appears to be about TABLETS, as TABLET, ANDROID and WINDOWS are amongst the top words for this topic. This topic received major attention starting around 2006/2007 before peaking in 2010-2012. Since then, the importance of the topic in the corpus has been decreasing and, given the topic timeline, one could conclude the adoption to the tablet technology is almost completed. The evolution of the topic's importance over time resembles quite closely a prototypical diffusion curve, spanning about 10 years.

In Figure 5 the WIKIPEDIA topic is detailed, which has been labeled according to top words such as WIKIPEDIA, ENZYKLOPAEDIE ("encyclopedia") or WIKIMEDIA. Compared to the TABLET topic, the peak appears earlier, around 2006. Furthermore, the increase in the topic weights from 2002 to 2006 appears steeper than in the TABLET topic, implying a faster rate of adoption. Similarly to the TABLET topic, adoption can be considered as almost complete, which aligns with intuition. The observation of a rather fast adoption rate of the WIKIPEDIA topic represents one of the advantages of the topic model based approach to determine the diffusion state of innovations as it allows for a quite granular and timely approximation of this indicator. For the application, a steeper might correspond to a more disruptive innovation compared to innovations with flatter curves, as adoption is quicker. Starting with this observation, one might build a measure for the disruptiveness of an innovation based on the rate of adoption measured through the importance of a topic over time.

The third topic shown in Figure 6 is labeled the VIRTUAL REALITY topic. The most relevant words in this topic are OCULUS, VIRTUAL and REALITY. Considering the importance of the topic over time, it becomes obvious, that the topic VIRTUAL REALITY has not peeked yet. However, it appears to be close to its peak, considering the overall shape of the aggregated weights over time, which resembles a Gaussian probability curve close to its maximum. Therefore, based

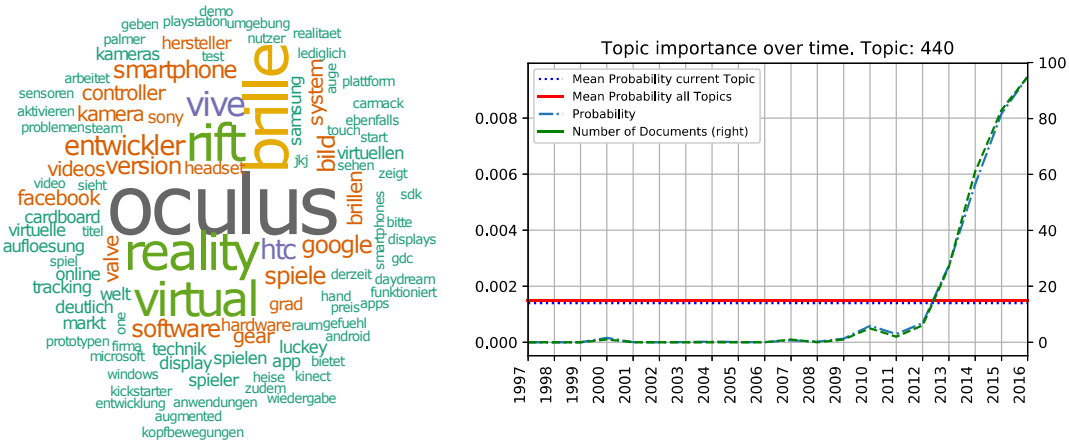


Fig. 6. Topic VIRTUAL REALITY *Left*: Most relevant words. *Right*: Importance over time.

Table 1. Johansen Trace Test for Cointegration

	Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
Wikipedia	None *	0.191	32.953	20.261	0.0005
	At most 1	0.018	2.632	9.164	0.6513
Tablet	None *	0.152	24.028	12.320	0.0004
	At most 1	0.002	0.315	4.129	0.6365
Virtual Reality	None *	0.132	25.425	20.261	0.0089
	At most 1	0.035	5.127	9.164	0.2697

Notes: Trace test indicates 1 cointegrating eqn(s) at the 0.05 level. * denotes rejection of the hypothesis at the 0.05 level. **MacKinnon-Haug-Michelis(1999) p-values(35). Maximum Eigenvalue Tests lead to qualitatively equal results and are not reported here.

on visual inspection, one might expect that after reaching the peek in 2018 or 2019, the relevance of this topic in the news corpus will decrease. In order to go one step beyond the visual inspection and intuitive reasoning regarding the information content of our topic based innovation measures, we compare their evolution over time with Google Trends Indices (GT)[¶], which have been used in the literature as a leading indicator for a variety of economic quantities (32), including the diffusion of product innovations(33, 34). In particular, (33) find that Google Trends data indicates a change in the interest for a product or technology well before it is visible in the purchasing behavior. The rational for using these indicators is that potential users, in a first step, will search for information on new products, possibly making use of a search engine. The Google Trends Index provides a measure of the relative popularity of a specific search term in Google's search engine over time. Unfortunately, it is not known in detail how these measures are constructed. The period of highest interest in the time span under consideration is set to 100, and the remaining observations are adjusted according to the relative frequency compared to the reference period. Thus, the Google Trends Indices provide information about the relative importance of a topic over time, but not compared with other topics.

For our application, we consider monthly observations for the three terms WIKIPEDIA, VIRTUAL REALITY, and TABLET, which have been obtained for the period between 01/2004 and 12/2016 in Germany. Figure 7 provides a graphical comparison of the Google Trends Index with the corresponding monthly relevance measures from our PVTM implementation. Thereby, the PVTM based measures have been smoothed to reduce noise using an exponentially weighted moving average over the last 12 month. In each figure, the left y-axis corresponds to the PVTM importance measure, while the scale of the Google Trends Index is given on the right y-axis.

The visual inspection of the plots in Figure 7 points at some lead of the PVTM indicators as compared to GT, which is most pronounced for the TABLET topic. To substantiate these findings, we apply Granger causality tests (36, pp. 48ff). The application of the Granger causlity test assumes stationary or at least cointegrated time series.

Therefore, we first tested all series for stationarity using the augmented Dickey-Fuller test with automatic lag length selection based on Schwarz' information criterion assuming a maximum lag length of 13. According to these tests, the null hypotheses of non-stationarity cannot be rejected for any of the variables at the 5 percent level as expected. Consequently, Johansen's cointegration tests was applied to identify potential long-run, cointegration relationships between the PVTM and corresponding GT series. Results for Johansens's trace test are presented in Table 1. The results point to the existence of one cointegration relationship for each pair of topic importance measure and Google Trends Index. The maximum eigenvalue tests lead to the same qualitative findings and are not reported here.

According to the test results, the application of the Granger

[¶] https://trends.google.com/trends/

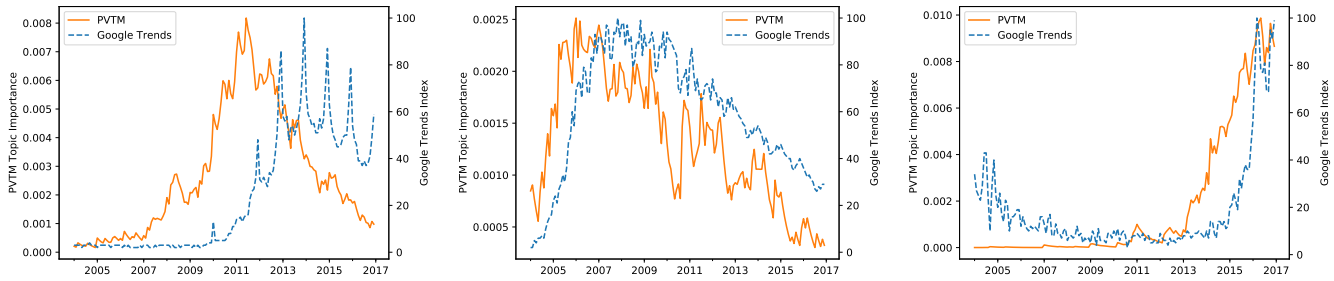


Fig. 7. PVTM topic importance measures compared with Google Trends Index. The upper image details TABLET, the image in the middle shows the WIKIPEDIA topic and the lower image compares VIRTUAL REALITY measures. We found statistical evidence that PVTM topic importance measures Granger cause Google Trends Index for each of the presented topics.

Table 2. VAR Granger Causality/Block Exogeneity Wald Test

Topic	Null Hypothesis	Chi-sq	df	Prob.
Wikipedia	PVTM does not Granger Cause Google Trends	0.645233	3	0.8860
	Google Trends does not Granger Cause PVTM	24.27469	3	0.0000
Tablet	PVTM does not Granger Cause Google Trends	14.86981	14	0.3871
	Google Trends does not Granger Cause PVTM	45.88463	14	0.0000
Virtual Reality	PVTM does not Granger Cause Google Trends	14.79103	7	0.0388
	Google Trends does not Granger Cause PVTM	41.47566	7	0.0000

causality test appears feasible, although the actual distribution of its Wald statistics might be involved. As a an alternative remedy for obtaining standard asymptotic distributions, according to (36, p. 49) one extra lag might be added to the VAR model underlying the Granger causality test. For each pair of variables, we select the lag length of a VAR model in levels based on Schwarz' information criterion. This lag length is increased by one for running the Granger causality tests. The results – including the lag length used – are shown in Table 2.

The null hypothesis that PVTM does not Granger cause the Google Trends Index may be rejected at the 1% significance level, thus suggesting derived PVTM topic indicators Granger cause Google Trends Index for the presented topics. Intuitively, news outlets and investigative journalists are likely to catch novel ideas faster than others. As reporting of topics in news media rises, public interest rises accordingly, however with some lag, during which people inform themselves about these new technologies and products on search engines like Google. Thereby we suggest that topic model importance measures from news corpora are available more early compared to Google Trends and, consequently, are more appropriate as early and leading indicators for the diffusion of innovations.

5. Conclusion and Outlook

There is an increasing interest in topic modeling, driven and ignited by the fast growing amount of textual data sources. In natural language processing, neural embedding methods have been shown to outperform standard methods on many tasks. They are therefore viable candidates for information retrieval from big text corpora, for example for topic modeling.

We proposed Paragraph Vector Topic Modelling, which uses Doc2Vec to construct document vectors and Gaussian mixture clustering to cluster the resulting vectors into meaningful topics. The applicability of our approach has been demonstrated by the emergence of coherent topics from technology related news articles. Thereby, it became apparent that PVTM offers a useful alternative to discover latent topics in text corpora.

First empirical examples derived from the application of PVTM to (technology) news ticker data demonstrate the potential relevance for innovation economics. In particular, it enables the measurement of diffusion of innovations over time. It appeared that the topic model measurements Granger cause Google Trends Indices for the presented topics. It remains a task for further research to derive methods for prediction of diffusion and for the assessment of entities involved in innovative activities with regard to their stage of technology adoption. Taking the ongoing digitization into account, early identification and measurement of innovations will become more important. Therefore, it is planned to analyze to what extent the method is applicable also in a dynamic setting. From an economic point of view it would be interesting to know which entities[‖] are being the main players in a topic. A simple possibility would be counting how often an entity has been mentioned. Going further, one could use entity related sentiment analysis to determine whether entities have a positive or negative impact on the topic. By identifying the first mentions of companies, we could classify them into categories such as innovators, early adopters, early majority, late majority, and laggards with respect to specific topics. Using live news feeds could offer the possibility to capture the diffusion of innovations with very little delay, while including more news sources might inherit the potential to cover a larger share of ongoing innovative activities.

[‖] Named entities can be denoted with a proper name and are real-world objects, such as persons, locations, organizations, products etc.

1. Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, and Robert A. Nisbet. Chapter 4 - applications and use cases for text mining. In Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, and Robert A. Nisbet, editors, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, pages 53 – 72. Academic Press, Boston, 2012. ISBN 978-0-12-386979-1. URL https://www.sciencedirect.com/science/article/pii/B9780123869791000049.
2. Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, May 2014. URL http://www.aeaweb.org/articles?id=10.1257/jep.28.2.3.
3. Ryohei Hisano, Didier Sornette, Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. High quality topic extraction from business news explains abnormal financial market volatility. *PLOS ONE*, 8(6):1–12, 06 2013. URL https://doi.org/10.1371/journal.pone.0064846.
4. Stefan Feuerriegel and Nicolas Prollochs. Investor Reaction to Financial Disclosures Across Topics: An Application of Latent Dirichlet Allocation. Papers, arXiv.org, May 2018. URL https://ideas.repec.org/p/arx/papers/1805.03308.html.
5. Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. Novel and topical business news and their impact on stock market activity. *EPJ Data Science*, 6(1):26, Oct 2017. ISSN 2193-1127. URL https://doi.org/10.1140/epjds/s13688-017-0123-7.
6. Nicolas Pröllochs and Stefan Feuerriegel. Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information Management*, 2018. ISSN 0378-7206. URL http://www.sciencedirect.com/science/article/pii/S0378720617309254.
7. Stephen Hansen, Michael McMahon, and Andrea Prat. Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870, 2018. URL https://ideas.repec.org/a/oup/qjecon/v133y2018i2p801-870..html.
8. Vegard H. Larsen and Leif Anders Thorsrud. The Value of News. Working Papers No 6/2015, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School, June 2015. URL https://ideas.repec.org/p/bny/wpaper/0034.html.
9. Jochen Lüdering and Peter Winker. Forward or backward looking? the economic discourse and the observed reality. *Journal of Economics and Statistics*, 236(4):483–515, 2016.
10. Stephen Hansen and Michael McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(S1):114–133, 2016. URL https://ideas.repec.org/a/eee/inecon/v99y2016is1ps114-s133.html.
11. Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *31st International Conference on Machine Learning, ICML 2014*, 4, 05 2014.
12. Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.
13. Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.
14. Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. URL https://doi.org/10.1080/00437956.1954.11659520.
15. John S. Bridle. *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*, pages 227–236. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990. URL https://doi.org/10.1007/978-3-642-76153-9_28.
16. Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *CoRR*, abs/1607.05368, 2016. URL http://arxiv.org/abs/1607.05368.
17. Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics, Second Edition*, pages 827–832. 2015. . URL https://doi.org/10.1007/978-1-4899-7488-4_196.
18. Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning and data mining*. Springer, 2017.
19. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984875.
20. Harihara Dadi, P Venkatesh, P Poornesh, Narayana Rao L, and N Kumar. Tracking multiple moving objects using gaussian mixture model. *International Journal of Soft Computing and Engineering (IJSCE)*, 3:114–119, 01 2013.
21. Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2014. ISBN 1447157788, 9781447157786.
22. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19 – 41, 2000. URL http://www.sciencedirect.com/science/article/pii/S1051200499903615.
23. Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Information Theory*, 28:129–136, 1982.
24. Keinosuke Fukunaga and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, January 1975. ISSN 0018-9448. URL http://dx.doi.org/10.1109/tit.1975.1055330.
25. Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. URL http://dx.doi.org/10.2307/2958889.
26. Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 241–248, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4035-9. . URL http://doi.acm.org/10.1145/2959100.2959167.
27. Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.
28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
29. OECD and Eurostat. *Oslo Manual 2018*. 2018. URL https://www.oecd-ilibrary.org/content/publication/9789264304604-en.
30. Everett M. Rogers. *Diffusion of innovations*. Free Press, New York, NY [u.a.], 5th edition, 08 2003. ISBN 0-7432-2209-1, 978-0-7432-2209-9.
31. Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL https://doi.org/10.3115/1118108.1118117.
32. Hyunyoung Choin and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x.
33. D. Duwe, F. Herrmann, and D. Spath. Forecasting the diffusion of product and technology innovations: Using google trends as an example. In *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1–7, Aug 2018. URL 10.23919/PICMET.2018.8481971.
34. Won Sang Lee, Hyo Shin Choi, and So Young Sohn. Forecasting new product diffusion using both patent citation and web search traffic. *PLOS ONE*, 13(4):1–12, 04 2018. URL https://doi.org/10.1371/journal.pone.0194723.
35. Joharji Ghazi A. and Martha Starr. Fiscal policy and growth in saudi arabia. *Review of Middle East Economics and Finance*, 6(3):24–45, 2011. URL https://EconPapers.repec.org/RePEc:bpj:rmeecf:v:6:y:2011:i:3:n:2.
36. Lutz Kilian and Helmut Lütkepohl. *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge University Press, 2017.