Stockholm University

Master's thesis

Department of Statistics

Masteruppsats, Statistiska institutionen

Nr 2021:01

Scale development using LDA text analysis

Philip Dahlqvist-Sjöberg

Master's Thesis 30 ECTS, VT 2021

Supervisor: Edgar Mauricio Bueno Castellanos

Abstract

In market research, the topic and formulation of questions, is an important aspect to consider when developing a survey. For many cases, a survey's intention is to measure and answer a specific question, but some aspects are too broad and undefined to measure using single questions. Svenskt Kvalitetsindex is a market research company, operating on the Swedish market, and uses a structural equation model to measure perception of aspects, e.g., sustainability. A combination of multiple questions are used to measure perceived sustainability, and these existing questions have been constructed using a standard approach, Scale Development process. This process uses field experts to determine what questions best represents the latent aspect. There has been concerns about the questions performance, evidenced by high proportions of don't know answers. We have the hypothesis that this may be due to the respondents not "feeling identified" with the questions constructed by the field experts.

This thesis investigates an alternative approach of finding and formulating items (i.e. questions), where experts are substituted by LDA text analysis in an attempt of capturing the respondents' perception of the aspect sustainability. These newly formulated items are validated in two stages. Firstly from the validation process used in the Scale Development process, and secondly the difference in proportion of *don't know* answers between the two approaches. The items constructed using text analysis perform adequate in the first stage. However, they present an increased *don't know* proportion. The conclusion is that the approach can construct equally good items, but with an increased proportion of *don't know*, as the standard approach. This conclusion is limited to the choice of model and aspect of investigation, and not generalized for all text analysis.

Contents

1 Introduction

When conducting a survey, there is a lot of research within the area of survey quality to assist the entire process. Among these literature, there is some practical publications such as Biemer & Lyberg (2003) and de leeuw et al. (2008). What they provide, is hands on information on different survey designs. There are many important factors to consider for achieving the desired result of a study, such as population boundaries, sampling methods, questionnaire item formulation etc. This field of research play an important roles in societies, due to the importance for government, companies and other institutions to collect and analyse social behaviours and changes within societies.

When it comes to formulating questionnaire items, what is discussed in literature, mainly concerns grammar and formulation of words to construct understandable questions that provide clear answers in regards to the purpose. What question to ask is seldom, or ever, discussed due to its purpose often being quite clear from the researchers perspective. A political party, may be interested in their chances in an upcoming election, and survey a sample of voters asking, *What political party will you vote on, in the upcoming election?*. The purpose of the question is very clear from the party, and result in a simple single question. In other cases, such as when measuring customers opinion on their perceived sustainability of a company through a latent variable constructed of multiple questions, the individual questions explaining the latent variable may not be as straight forward to formulate.

Hinkin et al. (1997) provides a description of Scale Development, which is a common approach in construction of questionnaires. The first step is to consult with experts within the field of research, about what items best measure the actual latent variable. This is a complex process with focus groups, basis in the specific area of research and often requires external collaborations. The second step is validating the items created. This common approach has been applied by, e.g., Svenskt Kvalitetsindex (SKI), which is a market research company that evaluate and support companies work with customer satisfaction. Their main model to assess customer satisfaction is Partial Least Square Path Modeling (PLS-PM) with five latent drivers; image, expectation, product quality, service and value (see Tenenhaus et al. (2005) for more information about PLS-PM). These all point towards two latent result variables; customer satisfaction and loyalty. The paths are depicted in Figure 1. The study *Customer satisfaction, market share, and profitability* (Anderson et al. 1994) on the Swedish market constitute the foundation of this model, that has been developed and refined during the past 30 years at SKI.

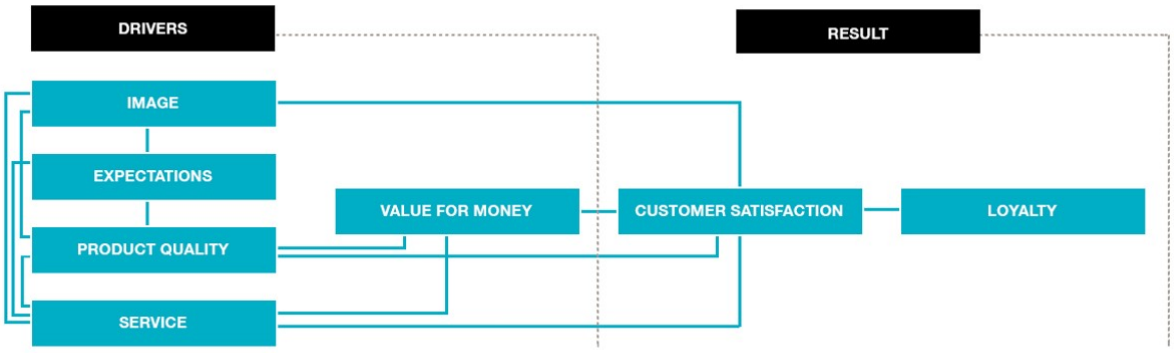

Figure 1: PLS-PM paths at SKI (Hallencreutz & Parmler 2019).

These latent drivers are respectively constructed of 2-5 items, that in theory should represent the latent variable. The process of formulating questionnaire items is an important step in understanding social behaviours. As an example, let us consider a company defining items regarding customer

opinions about a company's contribution to sustainability. Experts formulating these items typically work within the sustainability research. They might suggest items regarding very technical aspects, e.g., stakeholder's carbon oxide emissions. However, the customers, might not at all consider this aspect when deciding between different stakeholders, and there appears a conflict between these experts and the respondents regarding what is important for sustainability. The questions does not capture true customer preferences which is evidenced by high rates of *don't know* answers. In survey quality literature, this is considered as a type of specification error (Biemer & Lyberg 2003, p. 40-41). Throughout this thesis, we will refer to this situation as non-informative response, which indicates an answer that contribute no information on the item's scale. This is a common problem for items at SKI, and many other companies.

The experts are formulating items that will measure actual sustainability, however, the survey's objective is to measure perceived sustainability. There is a big difference between these two objectives. This process is depicted in the left flow chart in Figure 2.

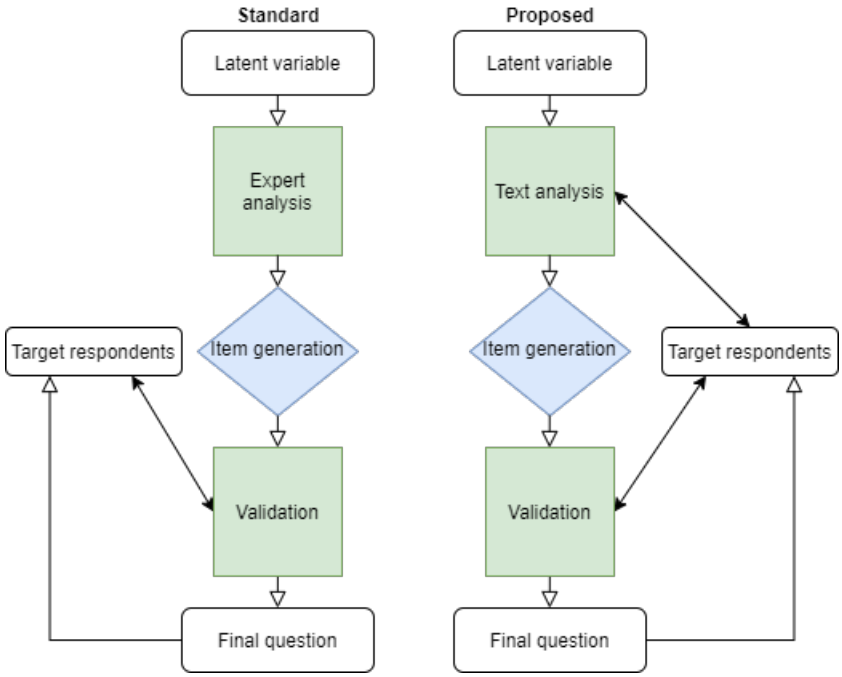

Figure 2: The left chart, shows the standard process of creating new questionnaire items. The right chart, shows the proposed approach that substitute experts with a quantitative step.

Due to the outdated, and tedious process of formulating new questionnaire items, developing a new approach is desired by SKI, but will serve a purpose for other stakeholders as well. The new approach, shown in the right flow chart in Figure 2, intends to substitute experts with a text analysis before the item generation. This is a way to include the target respondents early in the process, in order to capture what is important and tangible by the actual customer. This way, the perception of the respondent's beliefs are assumed to be measured more accurately, and indicate what questions to include in the survey. This new step, is what this thesis will implement and evaluate. More practically, the first step is to sample observations, where the respondents are asked to describe in an open answer, what they believe is important regarding the aspect. Then, with these independent open comments, topic analysis will be conducted in an attempt to categorize latent variables from the responses, which will serve as the foundation to formulate new survey questions. These newly constructed items will be compared, through the steps of Scale Development process, with items constructed using the standard approach of experts. With the business vision at SKI, *Actionable insights for a sustainable future*, the latent variable investigated in this thesis is the sustainable aspect.

As a baseline, the aim is to evaluate if this innovative approach, using the rapidly advancing field of text analysis, can substitute the old process of deciding what questions to represent a latent variable in a survey. Validation is performed based solely on the treatment of deciding the topic for questionnaire items. As a final step, the difference of non-informative response will be measured and evaluated between the two processes of developing new questionnaire items. This second step will serve as a more practical indication of improvement, however, the main evaluation of the approach is determined by the Scale Development process.

The content of this thesis is arranged as follows. In Section 2-4, the methods are described along with the results. LDA text analysis is described in Section 2, and validation using Scale Development in Section 3. Hypothesis testing with Bayes Factor is derived in section 4. Conclusions and discussions are presented in Section 5, and section 5.1 consists of an ethical discussion. All implementation of analysis has been conducted in Python. The code is publicly available at GitHub[1].

[1] https://github.com/phdah/LDA-for-scale-development

2 Text analysis

With an open-ended question, each respondent's answer represents a document, from which topics can be interpreted and constructed as new questionnaire items. For topic analysis on textual data, there are several models to consider; Non Negative Matrix Factorization (NNMF) (Pauca et al. 2006), Latent Semantic Analysis (LSA) (Dumais et al. 1988), Probabilistic Latent Semantic Indexing (pLSI) (Hofmann 2004), Pachinko Allocation Model (PAM) (Li & McCallum 2006), and more. A widely used topic model in Natural Language Processing (NLP) is Latent Dirichlet Allocation (LDA) (Blei et al. 2003). This method assign topics to each document of text, based on the words included in the set of documents. For this thesis, only one model will be applied, conditioning the results on the chosen model LDA. Hence, the results will not be general for all text analysis models.

The LDA model, is the Bayesian approach of the *frequentist* model pLSI. In the purpose of formulating items, LDA has an advantage of providing more generalized results than pLSI (Blei et al. 2003). The pLSI model has a linearly increasing number of parameters to estimate from the number of documents and topics, allowing the model to quickly overfit. LDA deals with these parameters as hidden *random variables*, preventing this linear increase of parameters. This is further evidenced by the comparison in Blei et al. (2003), where pLSI suffers from overfitting, while LDA performs relatively consistent for different dimensions of topics and documents.

Rehurek & Sojka (2011) pipeline Gensim is used for LDA modelling in Python. Gensim is a package with tools for *"topic modelling for humans"*.

2.1 Open-ended question

The first step of the process, is to generate an open-ended question to collect the documents for construction of the latent variables. The data collection was done through email survey with one single open-ended question, and a follow up question: *Sustainability is defined by three pillars: environmental, social and economic sustainability. Think of a company that you believe is sustainable based on these three pillars. What does this company do that makes you consider them sustainable? What company were you thinking of?*. All data collection has been conducted by PFM Research with a simple random sample from the Swedish population of age 18 or older.

For a more general analysis of the text, it has been transformed into a more generic structure. Firstly, all answers have been translated from Swedish to English, through Google translate. This is done due to the advances in pre-trained text cleaning with English words. Secondly, all special signs[2] were removed. All empty documents were removed, and stop words[3] were excluded. Finally, normalization of words with a pre-trained lemmatisation[4] model was applied.

With the cleaned data, each word within each document, are separated into tokens and mapped with an unique id. All documents are structured as a Bag of Words, i.e., the order of the tokens are lost, where only the token frequency in each documents are of importance. These documents, constructed as a Bag of Words, are then run through the LDA algorithm.

2.2 Latent Dirichlet Allocation

LDA is a generative probabilistic model used to capture global dependencies, in context of NLP, between words in form of topics. We define all answers of the open-ended question as the corpus, and each answer as a document. For the model, we assume exchangeability (Blei et al. 2003), i.e., that words are generated by topics and that topics are infinitely exchangeable within documents.

LDA make use of the Dirichlet distribution as a prior, and Multinomial distribution as the likelihood. For the sake of completeness, we introduce the Dirichlet distribution of order K,

[2]Special signs, such as *#, !, @*, are found in string.punctuation documentation.
[3]Stop words, such as *the, a, an, in*, are found in nltk English documentation.
[4]Lemmatisation, which is grouping together the inflected forms of a word, are found in spacy documentation.

$$\theta \sim Dir(\alpha),$$
$$p(\boldsymbol{\theta}|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K}(\alpha_k)\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where $\alpha_k > 0$ $(k = 1, \ldots, K)$ and θ_k lies on the $(K-1)$-simplex, i.e., $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k = 1$. The Dirichlet Distribution is defined for $K \geq 2$.

Further, we have the Multinomial distribution of order K,

$$z \sim Multinomial(\boldsymbol{\theta}),$$
$$p(z|\boldsymbol{\theta}) = N! \prod_{k=1}^{K} \frac{\theta_k^{z_k}}{z_k!}, \tag{1}$$

where $\boldsymbol{\theta}$ is a probability vector of K number of categorical outcomes, and N is number of independent trials. The Multinomial distribution belongs to the Bernoulli family of distributions. When $K > 1$ and $N = 1$, e.g. when sampling one observation from the Multinomial distribution, we obtain the Categorical distribution,

$$p(z|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{[z=k]}.$$

Categorical distribution is also part of the Bernoulli family of distributions. However, we refer to the multinomial distribution in the case of LDA when sampling one observation.

Following the notation from Blei et al. (2003), we have a corpus $\mathcal{D} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$ with M documents. Each document $\boldsymbol{w}_m = \{w_{m,1}, \ldots, w_{m,N_m}\}$ is composed of N_m number of words. Additionally we have a dictionary, i.e., the set of unique words indexed by the set $\mathcal{V} = \{w^1, \ldots, w^V\}$, where V is the number of unique words in the entire corpus $\mathcal{D}$. For our data, we have 1205 documents and 982 unique words after cleaning of text.

The LDA algorithm is then,

1) For each k topic in $\mathcal{K} = \{1, \ldots, K\}$, where K is chosen arbitrarily by the researcher, see section 2.4.
    a) Draw $\boldsymbol{\beta} \sim Dir(\eta)$, where η is the Dirichlet hyper-parameter for $\boldsymbol{\beta}$, of length V
2) For each document in $\mathcal{D}$
    a) Draw $\boldsymbol{\theta} \sim Dir(\alpha)$
3) For each word in $\boldsymbol{w}_m$
    a) Choose a topic $z_{m,n} \sim Multinomial(\boldsymbol{\theta})$
    b) Choose a word $w_{m,n}$ from $p(w_{m,n}|z_{m,n}, \boldsymbol{\beta})$, which is a multinomial probability such in equation (1), conditioned on the topic $z_{m,n}$.

With the chosen topic $z_{m,n}$, from step 3a, and the V unique words, we describe the word topic probabilities with the $K \times V$ dimension parameter matrix $\boldsymbol{\beta}$,

$$\beta_{v,k} = p(w^v = 1|z^k = 1),$$

where we denote the v^{th} word with superscript w^v, and the word's topic with superscript z^k.

A graphical representation of this algorithm is depicted in Figure 3. Given the hyper-parameter α and η, our posterior distribution is,

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{m=1}^{M} p(\theta_m|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\theta_m)\ p(w_{m,n}|z_{m,n}, \boldsymbol{\beta}), \tag{2}$$

where,

$$\begin{aligned}
p(\beta_k|\eta) &\sim Dir(\eta) && \text{(Dirichlet)} \\
p(\theta_m|\alpha) &\sim Dir(\alpha) && \text{(Dirichlet)} \\
p(z_{m,n}|\theta_m) &= \theta_{m,z_{m,n}} && \text{(Draw from Multinomial)} \\
p(w_{m,n}|z_{m,n}, \boldsymbol{\beta}) &= \beta_{z_{m,n},w_{m,n}}. && \text{(Draw from Multinomial)}
\end{aligned}$$

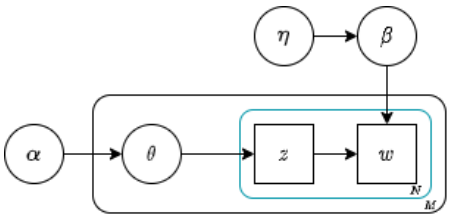

Figure 3: Graphical illustration of the LDA model. The outer box, represents the M documents, and the inner box is the N_m repeated word iterations. The M box represents step 1 and 2 in the LDA algorithm. The N box, represents step 3.

2.3 Online variational inference

For modeling of the LDA, analytical optimization of α and η is intractable of equation (2) proposed by Blei et al. (2003). Approximation can be done through, e.g., different types of Markov Chain Monte Carlo (MCMC) or, as Hoffman et al. (2010) proposes, using online variational inference. This approach is used as an optimization tool in this thesis. The online variational inference has a very fast convergence towards the true posterior, which is an advantage when working with large sets of data. The hyper-parameters are optimized by maximizing The Evidence Lower BOund (ELBO), which is defined as,

$$\log\ p(\boldsymbol{w}|\alpha, \eta) \geq \mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \triangleq \mathbb{E}_q[\log\ p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\alpha, \eta)] - \mathbb{E}_q[\log\ q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta})],$$

where the per-word posterior topic assignment z is parameterized by ϕ, θ is parameterized by γ on document level, and $\boldsymbol{\beta}$ is parameterized by $\boldsymbol{\lambda}$. We use a fully factorized entropy variational distribution family $\mathcal{Q}$ of the form,

$$\begin{aligned}
q(z_{m,n} = k) &= \phi_{m,w_{m,n},k} \\
q(\theta_m) &= Dir(\theta_m; \gamma_m) \\
q(\beta_k) &= Dir(\beta_k; \lambda_k).
\end{aligned}$$

Equivalently as maximizing the ELBO, we can optimize the hyper-parameters by minimizing the Kullback-Leibler (KL) divergence between the simpler function $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and the marginal posterior $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{w}, \alpha, \eta)$, which can be factorized as,

$$\mathcal{L} = \sum_{m=1}^{M} \{\mathbb{E}_q[\log\ p(\boldsymbol{w}_m|\theta_m, z_m, \boldsymbol{\beta})] + \mathbb{E}_q[\log\ p(z_m|\theta_m)] - \mathbb{E}_q[\log\ q(z_m)] \\ + \mathbb{E}_q[\log\ p(\theta_m|\alpha)] - \mathbb{E}_q[\log\ q(\theta_m)] + (\mathbb{E}_q[\log\ p(\boldsymbol{\beta}|\eta)] - \mathbb{E}_q[\log\ q(\boldsymbol{\beta})])/M\}.$$

By further extending the likelihood with its respective variational parameters, we realise how the variational object relies only on $n_{m,w}$, the number of times the word w appears in document m. For length V of the dictionary we derive,

$$\begin{aligned}
\mathcal{L} &= \sum_{m=1}^{M} \sum_{w \in \boldsymbol{w}_m} n_{m,w} \sum_{k=1}^{K} \phi_{m,w,k} (\mathbb{E}_q[\log\ \theta_{m,k}] + \mathbb{E}_q[\log\ \beta_{k,w}] - \log\ \phi_{m,w,k}) \\
&- \log\ \Gamma\left(\sum_{k=1}^{K} \gamma_{m,k}\right) + \sum_{k=1}^{K} (\alpha - \gamma_{m,k})\, \mathbb{E}_q[\log\ \theta_{m,k}] + \log\ \Gamma(\gamma_{m,k}) \\
&+ \left(\sum_{k=1}^{K} - \log\ \Gamma\left(\sum_{w \in \boldsymbol{w}_m} \lambda_{k,w}\right) + \sum_{w \in \boldsymbol{w}_m} (\eta - \lambda_{k,w})\, \mathbb{E}_q[\log\ \beta_{k,w}] + \log\ \Gamma(\lambda_{k,w})\right) /M \\
&+ \log\ \Gamma(K\alpha) - K \log\ \Gamma(\alpha) + (\log\ \Gamma(V\eta) - V \log\ \Gamma(\eta))/M \\
&\triangleq \sum_{m=1}^{M} \ell(n_m, \phi_m, \gamma_m, \boldsymbol{\lambda}),
\end{aligned}$$

where $\ell(n_m, \phi_m, \gamma_m, \boldsymbol{\lambda})$ denotes the m^{th} document's contribution to the ELBO. The variational distribution of a document, depicted in Figure 4, is defined as,

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n).$$

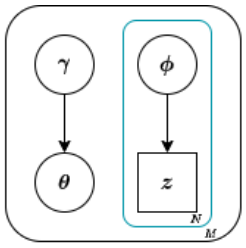

Figure 4: Graphical illustration of the online variational LDA model modification. The outer box, represents the M documents, and the inner box is the repeated N_m word iterations. Within the E step of Algorithm 1, we parameterize θ in the the outer box, and z in the inner box.

The LDA hyper-parameters, α and η, are approximated analogy with the Expectation-Maximization (EM) algorithm. The expectation under q in the E step, where both θ and β are Dirichlet distributed, is defined as,

$$\mathbb{E}_q[\log\ \theta_{m,k}] = \Psi(\gamma_{m,k}) - \Psi(\sum_{k=1}^{K} \gamma_{m,k}) \\ \mathbb{E}_q[\log\ \beta_{k,w}] = \Psi(\boldsymbol{\lambda}_{k,w}) - \Psi(\sum_{w \in \mathcal{V}} \boldsymbol{\lambda}_{k,w}),$$

where Ψ denotes the digamma function, which is the first derivative of the logarithm of the gamma function, During this step, $\boldsymbol{\lambda}$ is held constant. Further, in the M step, $\boldsymbol{\lambda}$ is updated given ϕ and the log likelihood of α and η, on the basis of $\boldsymbol{\lambda}$, γ and ϕ.

Hoffman et al. (2010) online variational Bayes for LDA is defined in Algorithm 1. The theoretical algorithm runs for $m = 0$ to ∞, but in practice we run for M iterations. We obtain the empirical Bayes estimation of $\boldsymbol{\beta}$, as $M \to \infty$. In the E step, we find the locally optimal values for γ_m and ϕ_m, holding $\boldsymbol{\lambda}$ fixed. In the M step, we then compute $\tilde{\boldsymbol{\lambda}}$, given the optimum ϕ_m. The weight given to $\tilde{\boldsymbol{\lambda}}$ is controlled by $\tau_0 \geq 0$, which slows down the early iterations, and $\kappa \in (0.5, 1]$ which controls the rate at which old values are forgotten. For machine computation, $\kappa > 0.5$ is arbitrary from $\kappa \geq 0.5$, allowing $\kappa \in [0.5, 1]$. We also compute α and η, where we use a linear-time Newton-Raphson method (Blei et al. 2003) which uses the second order Taylor expansion. Hence, $\tilde{\alpha}(\gamma_m)$ is defined as the inverse of the Hessian times the gradient $\nabla \ell(n_m, \gamma_m, \phi_m, \boldsymbol{\lambda})$, and $\tilde{\eta}(\boldsymbol{\lambda})$ is the inverse of the Hessian times the gradient $\nabla_\eta \mathcal{L}$, where $\mathcal{L}$ is $\sum_{m=1}^{M} \ell(n_m, \phi_m, \gamma_m, \boldsymbol{\lambda})$.

Algorithm 1: Online variational Bayes for LDA

Define $\rho_m \triangleq (\tau_0 + m)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $m = 0$ *to* ∞ **do**
  E step:
  Initialize $\gamma_{m,k} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{m,w,k} \propto \exp\left\{\mathbb{E}_q[\log\ \theta_{m,k}] + \mathbb{E}_q[\log\ \beta_{k,w}]\right\}$
    Set $\gamma_{m,k} = \alpha + \sum_{w \in \mathcal{V}} \phi_{m,w,k} n_{m,w}$
  until $\frac{1}{K} \sum_{k=1}^{K} |change\ in\ \gamma_{m,k}| < 0.001$;
  M step:
  Compute $\tilde{\lambda}_{k,w} = \eta + M n_{m,w} \phi_{m,w,k}$
  Set $\boldsymbol{\lambda} = (1 - \rho_m)\boldsymbol{\lambda} + \rho_m \tilde{\boldsymbol{\lambda}}$
  Set $\alpha = \alpha - \rho_m \tilde{\alpha}(\gamma_m)$
  Set $\eta = \eta - \rho_m \tilde{\eta}(\boldsymbol{\lambda})$
end

In our model, we use

$$\kappa = 0.5$$
$$\tau_0 = 1,$$

which is equivalent settings as the optimum findings in Hoffman et al. (2010).

Further, it is common in stochastic learning to use batches of observations per update in the algorithm (Hoffman et al. 2010). This means computing $\tilde{\boldsymbol{\lambda}}$ using subsets of the data, where $1 < S < M$ such that,

$$\tilde{\lambda}_{k,w} = \eta + \frac{M}{S} \sum_{s=1}^{S} n_{m,s,w} \phi_{m,s,w,k},$$

where we use $S = 100$. With the batches, we run each batch through the entire corpus 500 times and repeat the E step a maximum of 10000 times, or until convergence of γ.

2.4 Coherence score

In an attempt of evaluating a topic model, there is multiple measures that accomplices this. One common measurement is model perplexity (Chang et al. 2009). However, perplexity score will not

provide the best score based on number of topics, but improve as they increase. Common for the LDA model, Coherence score is used and implemented in the Gensim package. This is a measurement of the coherence, i.e., the semantic similarities, between words and the assigned topic. The advantage is that it can find the highest Coherence score among number of topics, given a fixed set of topics. This is an important property for our purpose of creating an automatic model, where we let the measurement decide the numbers of topics, i.e., number of questionnaire items to create. Röder et al. (2015) derives the Coherence score method used in this thesis. The process is divided in four steps; segmentation, probability estimation, confirmation and aggregation. There are several methods for calculating Coherence, which all have slightly different schemes in every step of the process. However, we only apply a single scheme in this thesis.

In the segmentation step, combinations of bigrams are constructed. We choose the n words with the highest probability for each topic. Then for each word ($w^{v=i}$) in the set $\{1, 2, \ldots, n\}$, we combine it with each of the remaining words in the set ($w^{v=j}$). This segmentation is denoted as S^{one}_{set}, where *set* defines that we make the calculations for all assigned topic. We denote the i^{th} word as w' and the word it is combined with as w^*. In our analysis, we use $n = 20$.

With the segmentation and combinations constructed, word estimation of the probabilities is computed with Boolean sliding window ($\mathcal{P}_{sw}$), which moves a window over the set of documents, each token at a time. Each window step, defines a virtual document. For these virtual documents, Boolean document ($\mathcal{P}_{bd}$) calculates the proportion of documents containing the words and joint words in each set of S^{one}_{set}. In our estimation, we use $\mathcal{P}_{sw(110)}$, i.e., the window size is 110 tokens.

Further, the confirmation measure takes a single pair $S_k = (w', w^*)$ of words and their respective probabilities, and measures how well the conditioning word w^* supports w', through indirect confirmation measure. We define $\boldsymbol{v}(w')$ and $\boldsymbol{v}(w^*)$ as the sum of the direct confirmations of the single words in the set of words, which is the Normalized Pointwise Mutual Information (NPMI),

$$NPMI(w', w^*) = \frac{PMI(w', w^*)}{-\log(p(w', w^*) + \epsilon)},$$

where ϵ is a constant to avoid taking the logarithm of zero. Pointwise Mutual Information (PMI) is defined as,

$$PMI(w', w^*) = \log \frac{p(w', w^*) + \epsilon}{p(w')p(w^*)}.$$

The indirect confirmation measure is then defined as the cosine similarity,

$$S_{cos}(\boldsymbol{v}(w'), \boldsymbol{v}(w^*)) = \frac{\boldsymbol{v}(w')\boldsymbol{v}(w^*)}{\|\boldsymbol{v}(w')\| \ \|\boldsymbol{v}(w^*)\|},$$

where $\|\cdot\|$ is the Euclidean norm. Finally, the mean of all sets, i.e. topic, are calculated as the final model Coherence score,

$$\text{Coherence Score} = \frac{1}{K} \sum_{k=1}^{K} S_{cos,k}. \tag{3}$$

The higher the score is, the more Coherence is between words and topics.

To assess a suitable number of topics, a model and its respective Coherence score is measured for the set $\mathcal{B} = \{K = 2, \ldots, K = T\}$, where T is the maximum number of topics to be estimated, and the number of topics which attains the highest Coherence score, is chosen as the final model. In our estimation we use $T = 10$. The choice of $\mathcal{B}$ is decided based on a minimum and maximum number of items representing a latent variable. Less than two would not fulfill the conditions of a Dirichlet distribution, and more than ten items would potentially overfit the model and not generate general topics suitable for item generation.

2.5 Results text analysis

The highest Coherence score using equation (3), of 0.5871, is attained for the model with eight topics, and is used as the final model. The Coherence score for all models are depicted in Figure 5.

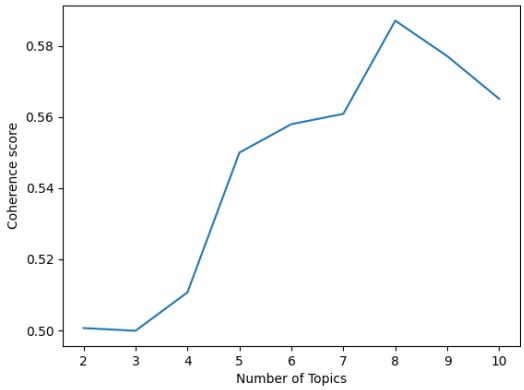

Figure 5: Coherence score for each model in the set $\mathcal{B} = \{K = 2, \ldots, K = 10\}$. Maximum Coherence is 0.5871, found for the model with eight topics.

From the final model and algorithm 1, we optimized α to be,

$$\alpha = \begin{bmatrix} 0.58 & 0.25 & 0.29 & 0.67 & 0.29 & 0.23 & 0.26 & 0.42 \end{bmatrix},$$

which is distributed with more probability mass towards topic four, one and eight in the Dirichlet distribution. The topic probability is determined by the size of the Dirichlet hyper-parameter in relation to all other topic hyper-parameters. For small values, the probability mass is located towards the edges of the simplex. A small value is considered when $\alpha_k \leq 1$. Further, η, which is the hyper-parameter for β, is a vector of length 982, hence, difficult to illustrate. However, we observe a relative symmetric distribution of η values between 0.11 and 0.14. The analysis was for the sake of comparison further conducted using symmetric priors, and provided similar results. Hence, optimized priors are used in the final model.

From the final model, we observe a distribution of documents over topics $\boldsymbol{\theta}$, depicted in Figure 6, where we have assigned each document to its respective highest density topic probability. The proportions follow our optimized topic prior knowledge α, as expected since we iteratively update the model prior hyper-parameters in algorithm 1. Hence, we can interpret topic four as more general, due to its higher α value and large proportion of document assignment. This knowledge is useful when interpreting the topics as items.

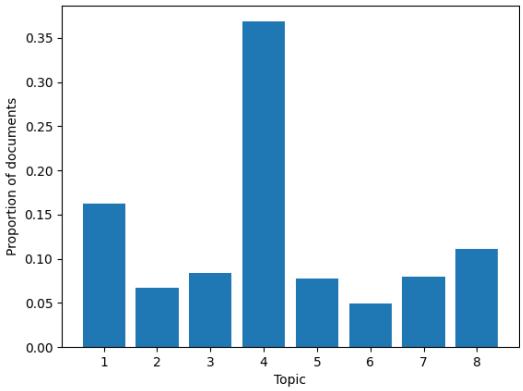

Figure 6: Distribution of documents over the eight topics.

Further, we observe the topics as distributions over words in the optimized $\boldsymbol{\theta}$ matrix, and obtain a topic probability from each unique word. However, we only define the top ten words as the topic keywords for each topic. These are depicted for the respective topics in Figures 7-14. From the keywords, we interpret the latent meaning of each topic in context of sustainability, which we use to construct questionnaire items. These keywords come from the normalized β matrix.

Looking at topic one, we see words such as *good, organization* and *time*, which we interpret as the company working with sustainability long-term. Topic two is represented by words such as *production, quality* and *advertising*, which we interpret as the company communicate how sustainable they are. Further in topic three, we observe words such as *environment, people* and *area*, which we interpret as caring about individuals and their well-being. Topic four contains words such as *environmental, sustainable* and *company*, which we interpret as the company being associated with sustainability. For topic five we observe words such as *sustainability* and *responsibility*, which we interpret as working towards a sustainable society. Topic six has keywords such as *profit, renewable* and *energy*, which we interpret as having a sustainable business model. Further, topic seven is represented by words such as *recycling, climate* and *electricity*, which we interpret as investing in environmentally conscious resources. Lastly, topic eight is explained by words such as *products, friendly, glass* and *jar*, which we interpret as promoting sustainable consumption.

Common for word distribution in text is that their frequency follow Pareto Rule. This means that roughly 80% of consequences comes from 20% of the causes. In relation to this rule we observe that the top 10 keywords for each respective topic, represents a large proportion of the word probabilities, which indicates their relevance for topic interpretation.

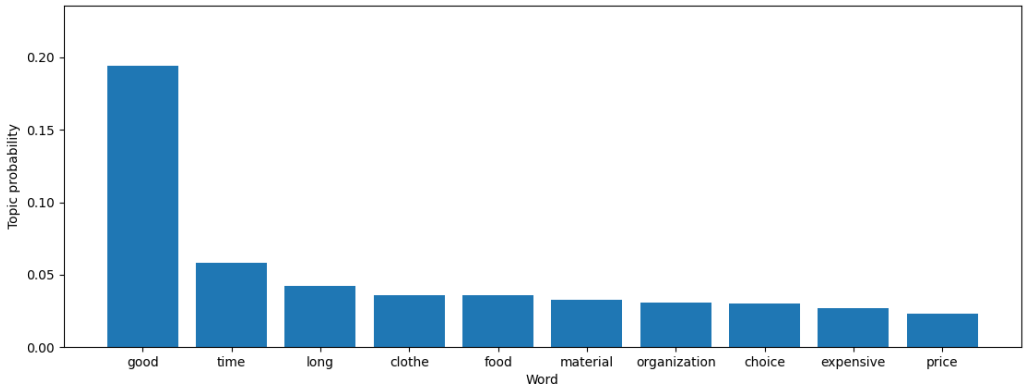

Figure 7: Distribution of topic one over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

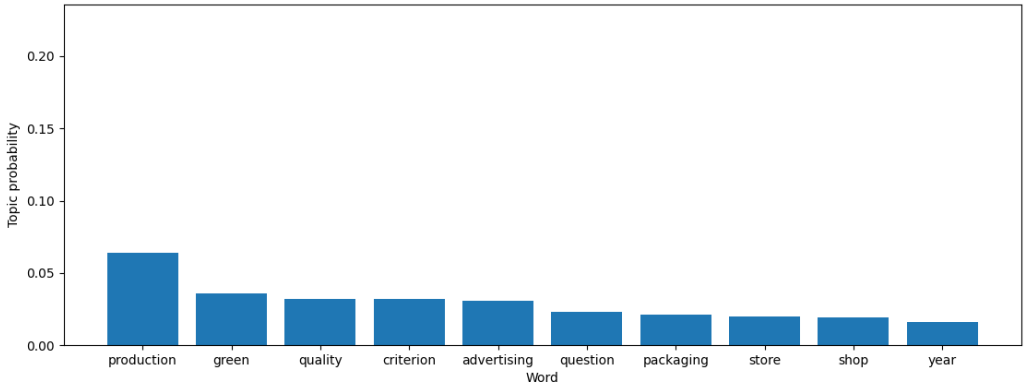

Figure 8: Distribution of topic two over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

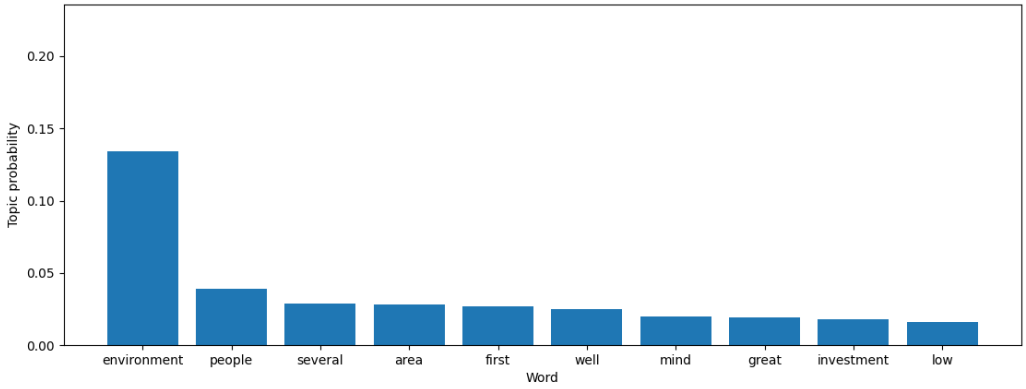

Figure 9: Distribution of topic three over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

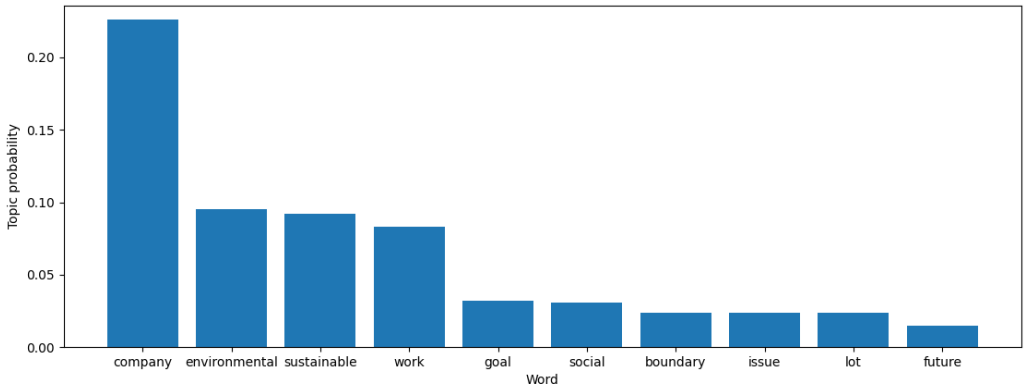

Figure 10: Distribution of topic four over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

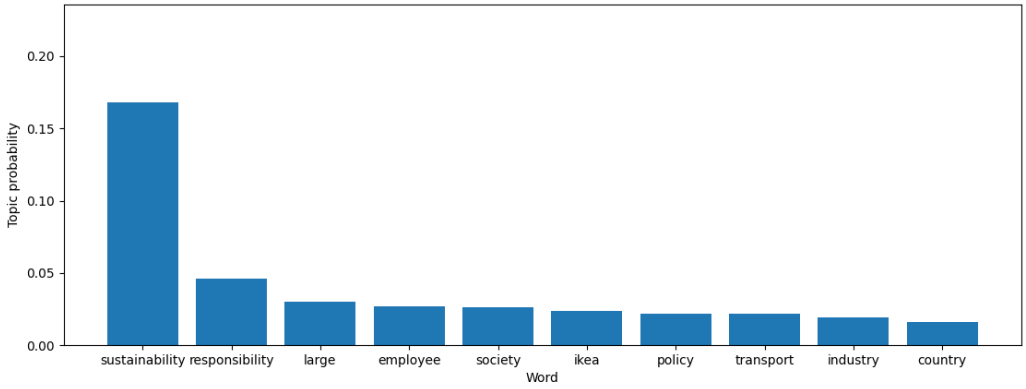

Figure 11: Distribution of topic five over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

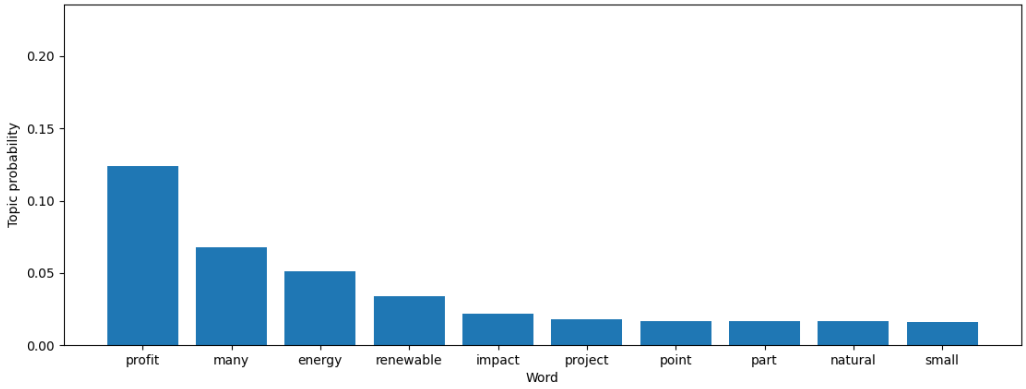

Figure 12: Distribution of topic six over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

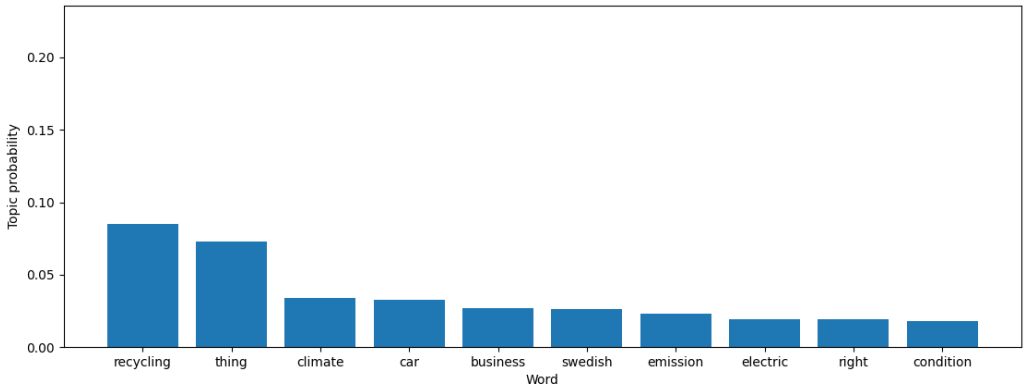

Figure 13: Distribution of topic seven over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

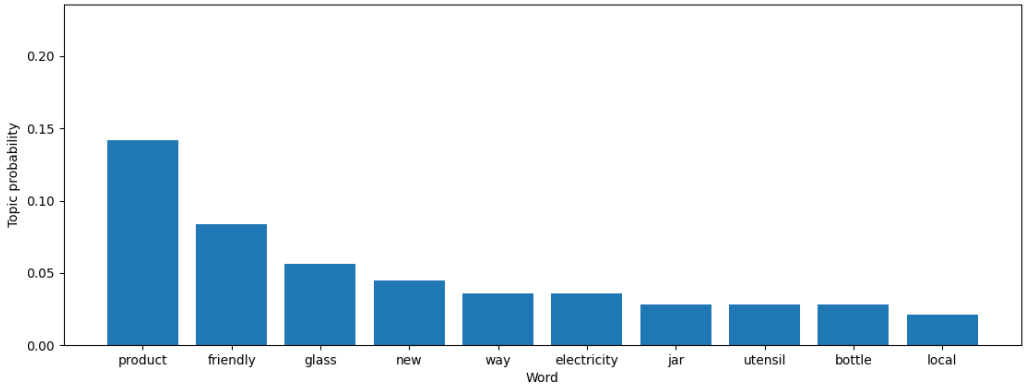

Figure 14: Distribution of topic eight over top ten keywords. This is a subset of word probabilities, hence, does not sum to one.

3 Scale Development

With the results from the text analysis, described in section 2, new questionnaire items are constructed and compared with the questionnaire items created from the standard approach.

Hinkin et al. (1997) describes a seven step Scale development process. This process can be found in Table 1, and is the standard process for companies such as SKI, in Sweden.

Table 1: Steps of the Scale development process (Hinkin et al. 1997).

Scale Development process
Step 1: Item Generation
Step 2: Content Adequacy Assessment
Step 3: Questionnaire Administration
Step 4: Factor Analysis
Step 5: Internal Consistency Assessment
Step 6: Construct Validity
Step 7: Replication

This process will be the foundation upon where the questionnaire items are evaluated, however, only step 1 and 3-5 will be applied. Step 1 is where results from the new approach is introduced, and the following steps are means of validating the items. This process will be assumed as a fixed means of validation for both methods.

3.1 Step 1: Item generation

The number of questions created by using text analysis, is dependent on the most appropriate number of topics, where Coherence score has been used to find the number of topics in the LDA model, see section 2.4.

From the text analysis, new questionnaire items have been generated from human interpretation of the topics and their respective keywords, see Table 2. All items have been constructed of a Likert scale from $1-10$, where 1 represents *disagree completely* and 10 *agree completely*. All questions also have an response option *don't know*. The reference items previously generated using the standard approach, also follow this structure and can be found in Table 3.

Table 2: Questionnaire items generated from the text analysis, proposed approach. "XX" is a placeholder for a company name.

No.	Questions
1	"XX" are long-term in their sustainable work.
2	"XX" uses sustainability in its marketing communication.
3	"XX" cares about the health and well-being of individuals.
4	"XX" is associated with sustainability.
5	"XX" works towards a sustainable society.
6	"XX" has a sustainable business model.
7	"XX" are environmentally conscious in their investments / production.
8	"XX" promotes sustainable consumption.

Table 3: Questionnaire items generated from experts, standard approach. "XX" is a placeholder for a company name.

No.	Questions
1	"XX" are long-term in their business
2	"XX" communicates to its customers about how to work with various issues around sustainability
3	"XX" contributes to the development of a better society
4	"XX" does everything it can to eliminate or reduce its negative impact on the environment
5	"XX" takes responsibility for the sustainability of their products / services
6	"XX" provides services that follow ethical and sustainable guidelines
7	"XX" is responsible for its services

3.2 Step 3: Questionnaire Administration

Both the newly constructed and the existing items have been distributed through email as blocks, to a random sample of banking customers. The choice of validating the questionnaire items on banking customers is due to it being the largest industry of clients at SKI. Further, evidenced from the open-ended question, a majority of respondents thought about retail companies when describing what they believe constitutes a sustainable company. Hence, to reduce bias in validating the items, we chose to test them on a different industry.

Hinkin et al. (1997) discusses different minimum requirements for sample sizes, where Bollen (1989) argues 100 observations to be sufficient for validation. The sample size needed is dependent on the number of items evaluated, where seven and eight is relatively low in this context. Both sets of items have been sampled with 411 complete answers, i.e., an answer on all items in the assigned block including *don't know* answers. The data collection has been conducted by PFM Research from the Swedish population of age 18 or older on the banking industry, from the four largest banks in Sweden.

3.3 Step 4: Factor Analysis

The objective of constructing these questions, is to find items that best represents a latent variable. In this case sustainability. To validate this condition, an exploratory factor analysis has been conducted (Hinkin et al. 1997) with Principal Component Analysis (PCA). Since we are only constructing a single factor, i.e. latent variable, only the first eigen vector is of interest. As a baseline, only items that clearly load on the first factor will be retained in the item block, where we use the 0.3 criterion (Morgado et al. 2018).

How to handle missing values in PCA analysis, is a well discussed problem. PCA uses correlation between continuous scales, and missing values are not on the continuous space. There are multiple methods to impute missing values, such as mean imputation (Donders et al. 2006) and imputation using Random Forest (Tang & Ishwaran 2017). The main idea is to impute a simulated value to replace the missing value with, instead of removing the entire observation. However, imputing the data may inflict error, specifically editing error (Biemer & Lyberg 2003), which is why no imputation has been conducted in our analysis.

In Table 4, the result of the factor analysis using PCA for the questionnaire items generate from text analysis, are shown. The first factor corresponds to 90% of the total variance of the data, consisting of 122 observations after removing observations with any missing values.

We observe that all loadings are pointing towards the same component, i.e., latent variable that is sustainability. Further, all eight items meets the 0.3 criterion, and all will be kept in the final block of items.

Table 4: Output of the first eigen vector from factor analysis, using PCA, on questionnaire items created using text analysis.

Loading	PC1 (90% variance explained)
1	-0.35
2	-0.358
3	-0.358
4	-0.354
5	-0.352
6	-0.354
7	-0.345
8	-0.357

In Table 5, the result of the factor analysis from questionnaire items created using the standard approach, are shown. The first factor corresponds to 81% of the total variance of the data, which consists of 117 observations after removing observations with any missing values. Similarly to the PCA on text analysis generated items, the standard approach items attain a clear latent variable where all items are kept.

Table 5: Output of the first eigen vector from factor analysis, using PCA, on questionnaire items created using standard approach.

Loading	PC1 (81% variance explained)
1	-0.364
2	-0.358
3	-0.396
4	-0.392
5	-0.388
6	-0.349
7	-0.397

From the result of both the standard approach, and the items generated using text analysis, we observe that all items in both approaches clearly load on the first latent variable, and all items are kept in the final block of questionnaire items.

3.4 Step 5: Internal Consistency Assessment

The aim is to construct questionnaire items that measure the intended concept. The most commonly used measure of internal consistency in field studies is Cronbach's alpha. It was first introduced by Cronbach (1951), and measures the reliability of respondents over items. Tavakol & Dennick (2011) describes Cronbach's alpha as an objective means of measuring the extent to which an instrument, i.e. our questionnaire items, measures what is intended to be measured. Cronbach's alpha relies on the assumption of unidimensionality to be interpreted as a single statistic for a group of items, which is fulfilled in our case of a single latent variable sustainability.

Cronbach's alpha lies on $[0, 1]$ and as a baseline a value larger or equal to 0.7 is assumed to provide adequate internal consistency. A too small value indicates either low number of items or poor interrelatedness between items. This step of the validation process, indicates if the block of items are *good*, in a measurable way.

Cronbach's alpha is defined as,

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right), \tag{4}$$

where k is number of items, $\sum_{i=1}^{k}\sigma_i^2$ is the sum of item wise variance, and σ_X^2 is the variance of respondents sum of item answers. Similarly as the PCA, Cronbach's alpha does not handle values not on the continuous scale. Hence, the data for the text analysis approach consists of 122 observations, and the standard approach of 117 observations. From equation (4), we calculate the alpha for items derived from the standard approach to be,

$$\alpha_1 = 0.96,$$

and,

$$\alpha_2 = 0.984,$$

from the text analysis. This indicates that both sets of questionnaire item blocks measures what is intended to be measured, and the items constructed using text analysis performs at least as good as those constructed from experts.

4 Bayes Factor

For the final step, hypothesis testing has been performed on the proportion of non-informative response for all items in both models. For this, Bayes Factor (BF) is used. Morey et al. (2016) provides a derivation of BF, where we have a vector $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ of length n observations. For a situation of comparing J different models, we have $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J\}$. For testing a simple null hypothesis, and a simple alternative hypothesis, i.e. when $J = 2$, the BF looks like,

$$BF_{1,2}(\boldsymbol{y}) = \left[\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}\right]\left[\frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_2)}\right].$$

The first fraction, is the comparison of prior believes about the two different models. The model prior believes will in our case cancel out due to equal properties of no prior knowledge of how the two different sets of questionnaire items will perform. For comparing proportions of non-informative response, we assume binary data and a conjugate prior,

$$\begin{aligned} p(\boldsymbol{y}|\theta) &\overset{iid}{\sim} Bin(n, \theta), \\ p(\theta) &\sim Beta(\alpha, \beta), \end{aligned}$$

where θ is the proportion of occurrences, i.e., proportion of non-informative response in our case. First we find the posterior distribution,

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &= \binom{n}{s}\theta^s(1-\theta)^{(n-s)}\frac{\theta^{\alpha-1}(1-\theta)^{(\beta-1)}}{B(\alpha,\beta)}, \end{aligned}$$

Where n is the number of observations in the data, s is the number of *don't know* answers and $B(\cdot,\cdot)$ is the beta function. The posterior has the kernel of a Beta distribution, with updated parameters. Then we calculate the marginal likelihood for $\mathcal{M}_j$ by integrating with respect to θ,

$$\begin{aligned} p(\boldsymbol{y}|\mathcal{M}_j) &= \int_0^1 \binom{n}{s}\theta^s(1-\theta)^{(n-s)}\frac{\theta^{\alpha-1}(1-\theta)^{(\beta-1)}}{B(\alpha,\beta)}d\theta, \\ p(\boldsymbol{y}|\mathcal{M}_j) &= \binom{n}{s}\frac{B(\alpha+s, \beta+n-s)}{B(\alpha,\beta)}. \end{aligned}$$

For marginalizing the likelihood, this requires a true probability distribution, which is not necessarily defined while using an improper prior. Hence, for this step, using a proper prior allows for a uniquely defined marginal likelihood (Morey et al. 2016).

Since we are comparing proportions of non-informative responses of all items, from two independent sets of data, we calculate the marginal likelihood for both models, and simply take the fraction of them.

$$BF_{1,2}(\boldsymbol{y}, \boldsymbol{x}) = \left[\frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{x}|\mathcal{M}_2)}\right] = \frac{\binom{n_1}{s_1}B(\alpha+s_1, \beta+n_1-s_1)}{\binom{n_2}{s_2}B(\alpha+s_2, \beta+n_2-s_2)}, \tag{5}$$

where we assume the standard approach to be $\mathcal{M}_1$ and the text analysis as $\mathcal{M}_2$. Further, we treat each answer to each item as an individual answer. Hence, the proposed approach contain more observations in this analysis, due to it consisting of an additional item than the standard approach.

Looking in Figure 15, three different Beta prior distributions are depicted. From prior knowledge, i.e. previous studies using questionnaire items regarding sustainability, the proportion of non-informative response is centered around 45% with a relatively small variance. From previous

knowledge, the proportion of non-informative response lies between 0.2 and 0.7. Hence the most suitable prior is when using $\alpha = 30$ and $\beta = \frac{110}{3} \approx 36.7$, that is the orange semi-dotted line.

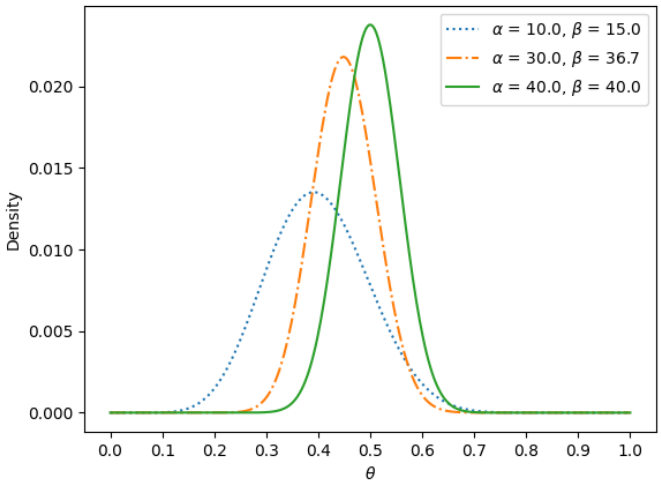

Figure 15: Illustration of prior Beta density distribution with different hyper-parameters for α and β.

Using our prior knowledge, the Beta prior, Binomial likelihood and Beta posterior for the standard approach data set is depicted in Figure 16. It is evident that the likelihood dominates the posterior distribution due to the high number of observations (2877). Further the same figure, but for the questionnaire items constructed using text analysis is depicted in Figure 17, where the same concept is evidenced by the number of observations (3288).

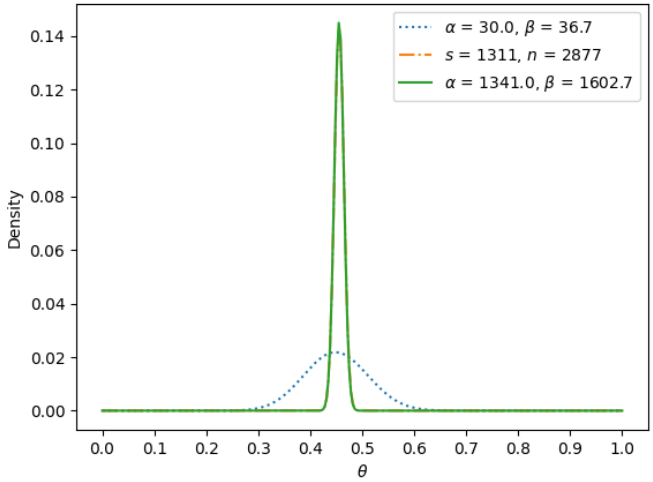

Figure 16: The standard approach's Beta prior, Binomial likelihood and Beta posterior distribution, using $\alpha = 30$ and $\beta = \frac{110}{3}$. s represents the number of missing values and n is the independent Binomial trials. It is clear, that the likelihood dominates the posterior. The blue dotted line is the prior, the yellow semi-dotted line is the likelihood and the green line is the posterior.

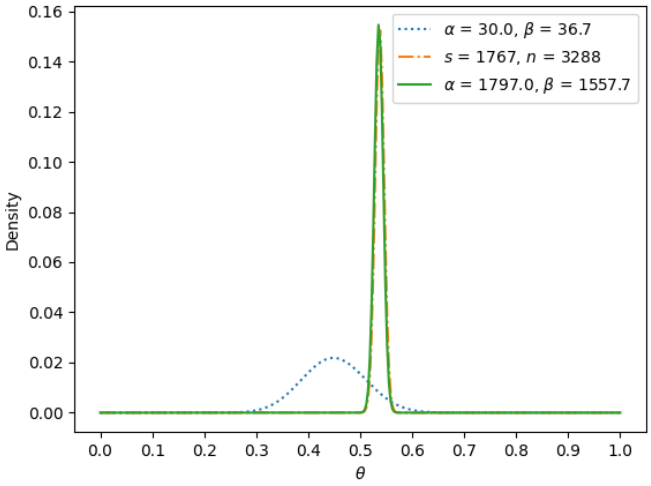

Figure 17: The text analysis's Beta prior, Binomial likelihood and Beta posterior distribution, using $\alpha = 30$ and $\beta = \frac{110}{3}$. s represents the number of missing values and n is the independent Binomial trials. It is clear, that the likelihood dominates the posterior. The blue dotted line is the prior, the yellow semi-dotted line is the likelihood and the green line is the posterior.

From the two sets of questions, we observe from the text analysis a proportion of non-informative response of 0.537, and 0.456 from the standard approach. Hence an increase when using the new approach of constructing questionnaire items. Following the described process of Bayes Factor, the significance of the difference is derived. From equation (5), we calculate the Bayes Factor to be 3.1 using prior hyper-parameters $\alpha = 30$ and $\beta = \frac{110}{3}$. Robert et al. (2008) have proposed a table to interpret the evidence provided by $\mathcal{M}_1$ in the BF, see Table 6. This implies moderate evidence for $\mathcal{M}_1$, and that there is a difference between the two sets of questions proportion of non-informative response.

Table 6: Bayes factor scale for $\mathcal{M}_1$, proposed by Robert et al. (2008). Analogues the inverse scale for $\mathcal{M}_2$.

$BF_{1,2}$	Evidence for $\mathcal{M}_1$
> 100	Extreme
$30 - 100$	Very strong
$10 - 30$	Strong
$3 - 10$	Moderate
$1 - 3$	Anecdotal
1	Inconclusive

However, the Bayes Factor is sensitive to prior knowledge. Looking in Figure 18, a subset of priors that all have a mean of 0.45, but with different variances are depicted. For a small variance, the density is larger than zero over a more narrow parameter interval, where a larger variance allow for a larger interval of θ.

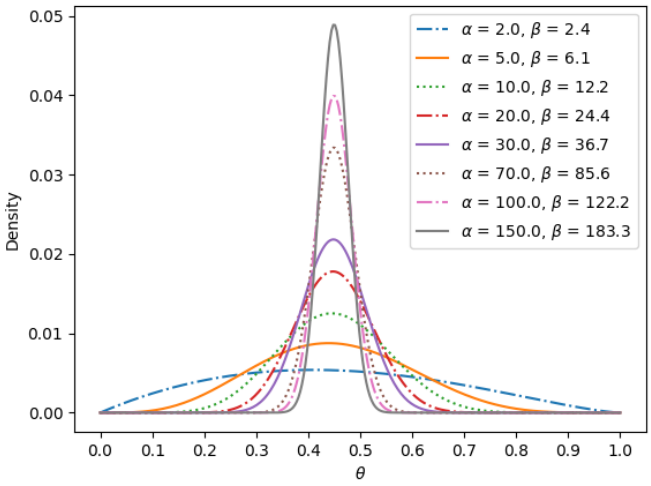

Figure 18: Beta distribution for eight sets of prior hyper-parameters with mean 45%, but different levels of variance.

The Bayes Factor for each individual prior depicted in Figure 18, are shown in Table 7. For strong evidence of a difference between the two sets of questionnaire items, a prior of $\alpha = 70$ and $\beta = \frac{770}{9} \approx 85.6$ yields this result. A Beta prior with these parameters, restricts θ to lie between 0.3 and 0.6. This interval is also reasonable from a practical perspective of previous knowledge. Hence, the evidence of a difference is strengthened based on the information in in Table 7.

Table 7: Bayes factor for difference in proportion of standard approach and text analysis, using Beta priors with mean 45%, but different levels of variance.

Priors (α, β)	$BF_{1,2}$	Evidence for $\mathcal{M}_1$
(2, 2.4)	1.2	Anecdotal
(5, 6.1)	1.4	Anecdotal
(10, 12.2)	1.6	Anecdotal
(20, 24.4)	2.2	Anecdotal
(30, 36.7)	3.1	Moderate
(70, 85.6)	11.0	Strong
(100, 122.2)	27.2	Strong
(150, 183.3)	114.8	Extreme

5 Conclusion and discussion

For construction of questionnaire items, LDA text analysis can substitute experts in the Scale Development process, and yield adequate good topics for construction of questions in regards to factor analysis and internal consistency.

The only requirements for the proposed approach, is that the conductor of the survey, e.g. SKI, have a latent aspect of interest and a general approach of asking questions. In our case, the sustainability block has historically had problems with high rates of *don't know* answers and was in need of improvement. Further, the use of Likert scale where the respondent answer on what level they agree on a statement, provides a simple and clear approach on how to formulate the questions with regards to the topics presented from the text analysis.

Due to the inclusion of non-informative responses in the validation step, a majority of observations had to be excluded in the PCA and Cronbach's Alpha calculations for both approaches. The exclusion of these observations inflicts bias since most respondents answered either *don't know* on all questions, or none. However, the results provide strong evidence of high correlation and internal consistency.

Additionally, the proportion of non-informative responses, proved with moderate to strong evidence to be higher for the questionnaire items constructed using text analysis. From the perspective of SKI, this was unfortunate since the reason for investigating the sustainability aspect was the high rates of non-informative responses. However, the proposed approach is conditioned on this specific aspect and the chosen model LDA, and may perform different in this sense for either other aspects or models. The most probable reason for the increased non-informative response, is due to the fact that the new questionnaire items were constructed from the respondents perspective of a retail company. However, in an attempt of reducing treatment error, the validation was done on banking customers. The items constructed using the standard approach, were more adapted for this industry, which may have provided a lower non-informative response rate.

As stated in section 2, there are many text analysis models to choose from, and advances are constantly made for better results of predicting, clustering and extracting information from open text. The Bag of Words, that LDA uses, may loose meaning of words when negations are lost before or after a specific word. However, since a manual interpretation of the topic is done, the main feature is keywords when identifying a specific topic. The formulation, i.e., the meaning of the answers to the open ended question, is also dependent on the open-ended question formulation. What we have done, is formulate an open-ended question where the respondent is asked to provide a positively loaded answer, in an attempt of reducing the miss interpretation of negations.

One of the positive feature of this proposed approach, is the simplicity of deploying the method. For construction of new questionnaire items, the stakeholders are required to sample the open-ended answers, interpret them as topics, and formulate questions from the topics.

For further exploration of the proposed approach, evaluation of different factors such as choice of model, aspect of investigation and an increased data set during the text analysis part would provide more general arguments for the approach. A specific model, that could potentially perform good at the task investigated in this thesis, is TopicRNN proposed by Dieng et al. (2017). This model combines the unsupervised aspect and interpretation of the LDA, but uses word ordering for obtaining the word probabilities with a Recurrent Neural Network. With analysis of other aspects, more known and well defined areas could derive more convincing results, especially for the proportion of non-informative responses. Such an aspect would be product quality, service quality or digital quality. The issue with sustainability is that most respondents, and stakeholders, have conflicting definitions of its meaning. The product quality is much more tangible by most individuals, and could from this proposed approach, also result in a reduction of non-informative response.

5.1 Ethical discussion

The main idea of the proposed approach of using text analysis to find topics for construction of questionnaire items, rely on using the respondents to decide themselves on what questions that best represent a latent variable. By introducing the respondents in the construction of items, this

inflicts bias since we use their opinion, to measure their opinion about an aspect. It implies an internal feedback loop, which could potentially inflict a conflict of reliability where the results are boosted due to question dependence. However, since the latent model at SKI measures perceived aspects, the constructed items need a contextual concept centered around what the respondents believe represents an aspect.

In a practical situation, and given the results provided from this thesis, a combination of text analysis and expert opinion could provide objective construction of items and merely bridge the perception gap between technical experts and the respondents beliefs. However, the performance typically increase when more information is introduced and combined. What this proposed approach investigate, is a method taking into account scarce resources, and finding items more suitable for perceived rather than actual aspect measures.

References

Anderson, E. W., Fornell, C. & Lehmann, D. R. (1994), 'Customer satisfaction, market share, and profitability: Findings from sweden', *Journal of Marketing* **58**(3), 53–66.
URL: *https://doi.org/10.1177/002224299405800304*

Biemer, P. & Lyberg, L. (2003), 'Introduction to survey quality', *Introduction to Survey Quality* .

Blei, D., Ng, A. & Jordan, M. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**, 993–1022.

Bollen, K. A. (1989), 'Structural equations with latent variables, new york', *Wiley Interscience* .

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. (2009), Reading tea leaves: How humans interpret topic models, Vol. 32, pp. 288–296.

Cronbach, L. (1951), 'Coefficient alpha and the internal structure of tests', *Psychometrika* **16**(3), 297–334.
URL: *https://EconPapers.repec.org/RePEc:spr:psycho:v:16:y:1951:i:3:p:297-334*

de leeuw, E., Hox, J. & Dillman, D. (2008), 'International handbook of survey methodology (2008)'.

Dieng, A. B., Wang, C., Gao, J. & Paisley, J. (2017), 'Topicrnn: A recurrent neural network with long-range semantic dependency'.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), 'Review: A gentle introduction to imputation of missing values', *Journal of Clinical Epidemiology* **59**(10), 1087–1091.
URL: *https://www.sciencedirect.com/science/article/pii/S0895435606001971*

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. & Harshman, R. (1988), Using latent semantic analysis to improve access to textual information, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', CHI '88, Association for Computing Machinery, New York, NY, USA, p. 281–285.
URL: *https://doi.org/10.1145/57167.57214*

Hallencreutz, J. & Parmler, J. (2019), 'Important drivers for customer satisfaction – from product focus to image and service quality', *Total Quality Management & Business Excellence* **32**(5-6), 501–510.
URL: *https://doi.org/10.1080/14783363.2019.1594756*

Hinkin, T. R., Tracey, J. B. & Enz, C. A. (1997), 'Scale construction: Developing reliable and valid measurement instruments', *Journal of Hospitality & Tourism Research* **21**(1), 100–120.
URL: *https://doi.org/10.1177/109634809702100108*

Hoffman, M., Blei, D. & Bach, F. (2010), Online learning for latent dirichlet allocation, Vol. 23, pp. 856–864.

Hofmann, T. (2004), 'Probabilistic latent semantic indexing', *the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99):1999* .

Li, W. & McCallum, A. (2006), Pachinko allocation: Dag-structured mixture models of topic correlations, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, Association for Computing Machinery, New York, NY, USA, p. 577–584.
URL: *https://doi.org/10.1145/1143844.1143917*

Morey, R. D., Romeijn, J.-W. & Rouder, J. N. (2016), 'The philosophy of bayes factors and the quantification of statistical evidence', *Journal of Mathematical Psychology* **72**, 6 – 18. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments.
URL: *http://www.sciencedirect.com/science/article/pii/S0022249615000723*

Morgado, F., Meireles, J., Neves, C., Amaral, A. & Ferreira, M. (2018), 'Scale development: Ten main limitations and recommendations to improve future research practices', *Psicologia: Reflexão e Crítica* **30**.

Pauca, V. P., Piper, J. & Plemmons, R. J. (2006), 'Nonnegative matrix factorization for spectral data analysis', *Linear Algebra and its Applications* **416**(1), 29–47. Special Issue devoted to the Haifa 2005 conference on matrix theory.
URL: *https://www.sciencedirect.com/science/article/pii/S002437950500340X*

Rehurek, R. & Sojka, P. (2011), 'Gensim–python framework for vector space modelling', *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* **3**(2).

Robert, C., Chopin, N., Rousseau, J., Bernardo, J., Gelman, A., Kass, R., Lindley, D., Senn, S. & Zellner, A. (2008), 'Harold jeffreys's theory of probability revisited'.

Röder, M., Both, A. & Hinneburg, A. (2015), Exploring the space of topic coherence measures, *in* 'Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6'.
URL: *http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf*

Tang, F. & Ishwaran, H. (2017), 'Random forest missing data algorithms: Tang and ishwaran', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10**.

Tavakol, M. & Dennick, R. (2011), 'Making sense of cronbach's alpha', *International Journal of Medical Education* **2**, 53–55.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. & Lauro, C. (2005), 'Pls path modeling', *Computational Statistics Data Analysis* **48**(1), 159–205. Partial Least Squares.
URL: *https://www.sciencedirect.com/science/article/pii/S0167947304000519*