Topic Modeling

Emre Kavak
Technical University Munich
emre.kavak@tum.de

Abstract

Topic Modeling (TM) is an unsupervised approach to analyze corpora of documents by correlating word occurencies, and, thereby, to find topics. This paper shows the evolution of Topic Models and particularly points to the fact how the probabilistic view on this task unifies different topic models, making each of them extensions/simplifications of one another. Especially, one of the main findings in this paper, is that Topic Models are flexible and relatively simple to extend depending on the context one wants to employ them. Finally, by presenting common applications, this paper shows that Topic Models are a main tool in data exploration and explanation, above all, in any domain that faces huge amounts of data that need to be understood by humans.

1 Introduction

We, as humans, prefer structured and organized information. Our newspapers have different sections such as sports, finance, and local news to locate facts quickly that are thematically related. Libraries have distinct areas where books are separated by the topics they cover. This exact same notion applies to nearly all areas of our lives. Categories and clusters of things, that share specific aspects of interest with eachother, help us in finding and understanding the given objects.

Obviously, in phyiscal settings such as in libraries, manual human effort needs to be performed in order to formulate common book topics, and to assign each book to the set of constructed *topics*. Borrowing this library analogy, one could view a clollection of unlabeled documents as books without tags in a library that has no sections yet. Documents in this sense can be Wikipedia articles, scientific paper abstracts, social media comments or product reviews - the definition is flexible as long as a given text is coherent/cohesive.

The task of *Topic Modelling (TM)*, in metaphorical terms, is to set up this library with its thematical sections without manual effort. In more formal terms, the goal is, given a set of raw textual documents, to find topics that are treated in the corpus and to show in what extends the topics are touched in each document automatically. The presented methods in this paper define topics as weights/probabilities of distinct words in the corpus (sports could be 10% ball, 7% referee, ...) and documents as a (probabilistically) weighted set of topics (document x could be 65% finance, 10% foreign policy, ...) (Deerwester et al., 1990; Blei et al., 2002; Hofmann, 2001). TM is thereby a fully *unsupervised* method.

This paper elaborates on the most common and basic Topic Modeling tools in section 2. The emphasis is on the fact that the presented models are introduced in increasing model complexity. Particularly noteworthy is that all these models are related to each other - the complex ones extending their predecessors. Section 3 shows high level applications in different fields and gives illustrative examples from different papers to support the arguments. The final section 4 concludes this paper with a summary.

2 Methods

All of the following methods require that documents are in their standard representation: word vectors. Thereby an entire corpus will be represented as a $D \times n$ matrix $X = [x_1\ x_2\ ...\ x_n]$ the columns x_i being

document *bag of word (BoW)* vectors, D the number of vocabulary in the corpus, and n the number of documents.

2.1 Latent Semantic Analysis (LSA)

2.1.1 Overview and Introduction

The easiest, non-probabilistic method was introduced by (Deerwester et al., 1990). The idea is to decompose the document-term matrix by the *Singular Value Decomposition (SVD)* into three components, and as a consequence to introduce a latent space representation of the original matrix. It is important to note here that the original paper proposed the method in the field of *information retrieval* with regards to documents and their representations. The idea was to find new representations for documents that are more efficient and complex compared to simple BoWs. Nevertheless, a by-product was the finding of an implicit topic model that can be formulated explicitly in a probabilistic fashion. An analysis of the link between the simple LSA here and the Probabilistic LSA (pLSA) is given in section 2.2.

SVD is a common and known mathematical finding/tool that enables a matrix decomposition that reads

$$X_{D\times n} = U_{D\times m}\Sigma_{m\times m}{V_{n\times m}}^T \tag{1}$$

where $U_{D\times m}$ and $V_{n\times m}$ have orthonormal columns and $\Sigma_{m\times m}$ is a diagonal matrix with non-negative entries $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_n \geq 0$. Note that the dimensions of V are before transposing the matrix. In many different domains, e.g. signal processing, biomedical computing, or scientific computing in general, the SVD is used as a comparably simple method to denoise, compress, or to reduce the dimensions of a given dataset (Brunton and Kutz, 2019). The approach is to keep only the largest k singular values to have a new, approximate representation $\hat{X}$ of the original matrix X

$$\hat{X}_{D\times n} = \hat{U}_{D\times k}\hat{\Sigma}_{k\times k}{\hat{V}_{n\times k}}^T. \tag{2}$$

A common way of interpreting this new decomposition is to view $\hat{U}_{D\times k}$ as the new rotated (orthonormal columns) and reduced dimension coordinate system. Then the columns of $\hat{\Sigma}_{k\times k}{\hat{V}_{n\times k}}^T$ will be the coordinates in this new coordinate system.

Tailoring the interpretation to the context of the document-term matrix, the new reduced dimension of k can be seen as the number of latent topics that is freely chosen by the user. To call this latent space topics may seem completely arbitrary. But the following interpretation should give some intuition on why this latent space can be seen as an abstract, non-observed topic space.

The column i of $\hat{U}_{D\times k}$ consists of numerical values that relate the D different words to the selected latent topic i as can be seen in

$$col_i(\hat{U}_{D\times k}) = \begin{pmatrix} w_{1,i} & w_{2,i} & \cdots & w_{D,i} \end{pmatrix}^T. \tag{3}$$

For every latent dimension, the factorization reveals that each of them is a vector of weights (numerical values) for each vocabulary in the corpus. Latent dimension i could be, for example, a vector like $(3.5\,car, -4.0\,transportation, 0.2\,cat, 2.0\,train, ...)$. As a human evaluator, one can see that this latent dimension gives high absolute values to terms that correspond to transportation and less to other words. Now it gets obvious that these novel structures can be seen as topics, since the decomposition weighs co-occuring words highly in certain dimensions. The premise is that frequently co-occuring words also share semantical aspects. Whilst this is an over-simplification, it often seems to be a reasonable approach (Deerwester et al., 1990).

Looking at the other matrix $\hat{\Sigma}_{k\times k}{\hat{V}_{n\times k}}^T$, one can see that each column relates the different topics to a document. Taking column j from this matrix directly shows

$$col_j(\hat{\Sigma}_{k\times k}{\hat{V}_{n\times k}}^T) = \begin{pmatrix} t_{1,j} & t_{2,j} & \cdots & t_{k,j} \end{pmatrix}^T \tag{4}$$

where $t_{i,j}$ are numerical values and the vector $col_j(\hat{\Sigma}_{k\times k}{\hat{V}_{n\times k}}^T)$ contains the new coordinates of document j in the new topic space. The interpretation here is that a document can be represented as a vector that assigns weights to topics.

Although the LSA method did not provide any modelling assumptions regarding topics and their representations, we implicitly attained the view:

- A topic is a collection of words
- A document is a collection of different topics.

2.1.2 Analysis

The initial goal of LSA was to handle document queries more efficiently. Instead of handling sparse BoW documents, one could use the shown decomposition to describe documents as weighted topics as in equation 4. Since the number of topics, k, is a hyperparameter chosen by the user, the new dimensions of documents are greatly reduced, since $k \approx 50\ to\ 300 \ll D$ (cf. (Deerwester et al., 1990)), where D is the number of vocabulary in a corpus (usually tens of thousands). Thereby the problem of *sparsity* is handled very efficiently. Furthermore, we have attained a novel description of words that partially alleviates problems of *synonymy*, words having similar/the same meaning whilst being written differently (cat, feline), and *polysemy/homonymy*, words with same orthography but differing meanings (bank as institution vs. bank as building). Due to the non-observed latent space, words with different orthography can now have high values in the new latent dimension, e.g. 'cat' and 'feline' could equally contribute to the same topics that highly correlate vocabulary concerning animals. Additionally, the same word, e.g. 'bank' can have high values in different latent dimensions (topics), enabling it to cover both the meaning of an institution and building.

Thus, the use case for efficient document representations is clear. For the use case for Topic Modelling, however, the pure LSA is not frequently used. Firstly, negative values in the vectors are often irritating and hard to interpret compared to normalized values. Especially beneficial would be an intuitively interpretable representation that would replace the real valued numerical weights by probabilities. Secondly, the introduction to LSA in terms of TM seems to be somehow arbitrary. No explicit assumptions are made, which, conversely, means that many things are decided without any control. Section 2.2 will illuminate what underlying probabilistic nature the simple LSA has and shows how to overcome the implied issues.

2.2 Probabilistic Latent Semantic Analysis (pLSA)

2.2.1 Overview and Introduction

We derived in the previous section the intuition that topics may be a collection of words and that documents can be seen as collections of topics.

Probabilistic LSA makes these assumptions explicit and assigns probabilities to a document d regarding a set of topics z_i by introducing $p(z|d)$, a probability vector given a document for each existing topic z_i (Hofmann, 2001). Moreover, pLSA models the probability vector $p(v|z_i)$, the probability for each vocabulary v_j given the topic z_i (Hofmann, 2001).

The *generative* framework is depicted and described in *plate notation* in figure 1. The prosa formulation of the process is (cf. (Hofmann, 2001))

1. Choose Document d from corpus $\mathcal{C} = \{d_1, d_2, \ldots, d_n\}$ with probability $p(d)$
2. For each of the N_d words in document d
   (a) choose a latent topic $z \in \{z_1, z_2, \ldots, z_k\}$ where k is the predefined parameter representing the number of topics. This can be chosen freely depending on the task (Mahmood, 2009; Boyd-Graber et al., 2017; Jelodar et al., 2019; Asmussen and Møller, 2019).
   (b) choose a word $v \in \{v_1, \ldots, v_D\}$ where D denotes the number of distinct vocabulary in the corpus.

Having attained a generative process that explains the intuition behind the modeling, we can estimate the parameters we are interested in. The goal is to recreate the training data which is the document-term

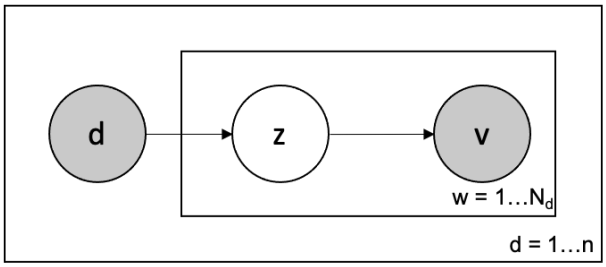

Figure 1: Plate notation of pLSA adapted from (Hofmann, 2001). The boxes depict the repetition of the generative process whilst the number of the iterations is written in the corners. The nodes are standard representations of random, observed and latent variables as known from *Bayesian Networks*. The gray nodes depict observed variables, in this case document d and word v. The white nodes represent random variables - here we have the explicitly introduced latent random variable z: the topic.

corpus as in every generative approach. The exact formulation reads

$$p(v|d) = \sum_z p(v|z)p(z|d). \tag{5}$$

The estimation procedure aims to recreate the observed corpus whilst optimizing our parameters $p(d), p(z|d)$ and $p(v|z)$. We basically want to perform a maximum likelihood estimation of the paramerters given the term-document matrix. The final goal reads

$$\max_{p(d),p(z|d),p(v|z)} \prod_{v,d} p(v,d) \tag{6}$$

There also exist different equivalent parametrizations of the same model given the Bayes Theorem that one may encounter in literature. To estimate this term, one takes the log-likelihood of it and employs an *Expectation Maximization (EM)* scheme that iteratively estimates the probabilities (Hofmann, 2001). The reason for employing this estimation scheme is that we introduced latent variables that are not observed in the data. In the esimation process, these need to be cancelled out requiring one to perform EM (Bishop, 2006).

2.2.2 Analysis

LSA introduced the latent space by using the SVD whilst pLSA poses the Topic Modelling task explcitly in a clear probabilistic manner. Although the methods seem to be unrelated, there is a strong link between these two methods.

One can link LSA and pLSA by formulating pLSA also as a matrix factorization task. The unifying formulation is that the document-term matrix can be factorized as $X = U\Sigma V^T$. For LSA we elaborted on this. For pLSA, one can do an anlogue factorization by formulating the matrix as $p(v,d) = p(v|z)p(z)p(d|z)$ (Hofmann, 2001; Christophe Dupuy, 2017; Papadimitriou et al., 2000). Just renaming the probabilty matrices, one can see that LSA and pLSA indeed have a common structure.

The main difference here, however, is that LSA minimizes the L_2- or Frobenius $norm$ (Hofmann, 2001; Christophe Dupuy, 2017; Papadimitriou et al., 2000)

$$min \left\| X - \hat{X} \right\|^2, \tag{7}$$

pLSA minimizes analogously the $KL - Divergence$

$$min\ KL(X||\hat{X}) \tag{8}$$

with regard to to the decomposition $\hat{X} = U\Sigma V^T$. KL() denotes the *Kullback-Leibler divergence*. Minimizing this term is equivalent to a maximum likelihood estimation (Bishop, 2006) given the probability matrices.

On the one hand, the LSA method, using a $L_2 - norm$ minimization, assumes Gaussian noise on the term-document counts. It is obvious that this assumption is odd and arbitrary since we handle discrete word counts. A result is that we obtain new representations that may include negative numbers that are not normalized and hard to interpret/compare. Lastly, the topic vectors are orthogonal, implying that they do not share any semantical aspects. This consequence is obviously flawed since in a realistic Topic Model, topics should be allowed to share semantical information (e.g. topics with high weights on words regarding animals could have semantical links to topics regarding humans, depending on the context).

The pLSA approach, on the other hand, realistically assumes multinomial topic-word and topic-document distributions capturing the discrete nature of word counts. Moreover, the entire process is fully probabilistic resulting in easily interpretable topic and document vectors (1.0 being the max number and 0.0 the lowest). Finally, the topics in this model are not restricted to be orthogonal in any geometric sense, allowing the model to relate topics to a certain degree. (Hofmann, 2001; Christophe Dupuy, 2017; Papadimitriou et al., 2000).

The pLSA approach improves on most of the short-comings of LSA, and, given advanced optimization schemes for the EM algorithm, the disadvantages of possibly poor estimations and computation time is also alleviated (Hofmann, 2001; Bassiou and Kotropoulos, 2011).

Nevertheless, pLSA also has its own limitations. The major one is that the maximum likelihood scheme is overfitting a given corpus since the number of estimated parameters increases linearly with the corpus size. Furthermore, there is no clear scheme to estimate probabilities of unseen documents $p(d)$ weakening its probabilistic nature (Blei et al., 2002). The posed issues are handled by the Latent Dirichlet Allocation (LDA) method that is presented next.

2.3 LDA

2.3.1 Overview and Introduction

The aim in LDA - as it was in pLSA - is still to assign probabilities to words being in a topic and topics being in a document (Blei et al., 2002). The graphical model is depicted in figure 2. The β_z in LDA

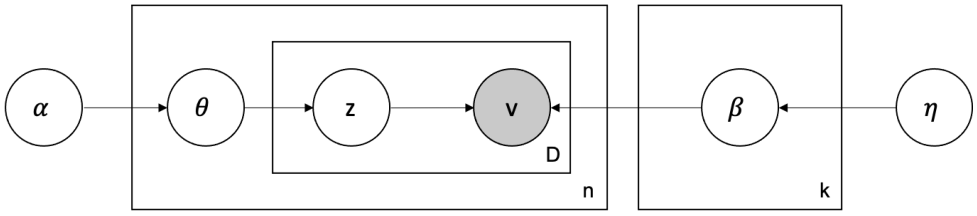

Figure 2: LDA plate notation adapted from (Blei et al., 2002). Analogue to pLSA model. θ and β are new Dirichlet random variables with their respective symmetric parameters α and η.

is related to the $p(v|z)$ of pLSA, which models the probability vector of words given a topic. The difference is that β is a random variable here that is Dirichlet distributed (Bishop, 2006) and controlled by the parameter η, assinging prior probabilities to the corpus wide topics in advance. The $p(z|d)$, the proportions of topics given a document, is in pLSA a model parameter and fixes thereby the corpus making the trained model not (easily) extendible to unseen documents. In LDA, θ is also the per-document topic distribution, but in this case, modeled as a random variable that is again Dirichlet distributed with hyperparameter α, relaxing the tight coupling to the observed documents.

The effects of the Dirichlet priors are depicted in figure 3. In pLSA, we did not provide any prior distribution but solely performed a maximum likelihood estimation. This implies uniform priors on the parameters, thus, assumes that given a document, each topic is equally likely to occur, and, given a topic, each word is equally likely to be observed (in advance). Given figure 3, it would equal the case with hyperparameter 1. LDA, in contrast, allows to control the prior assumption. It can favor only a few topics given a document, or, prefer a handful of words given a topic by varying the Dirichlet parameters.

The introduced Dirichlet distributed prior variables θ and β on the Multionomials z and v, however, introduce a growing complexity since we have now integrals in our inference process that cannot be

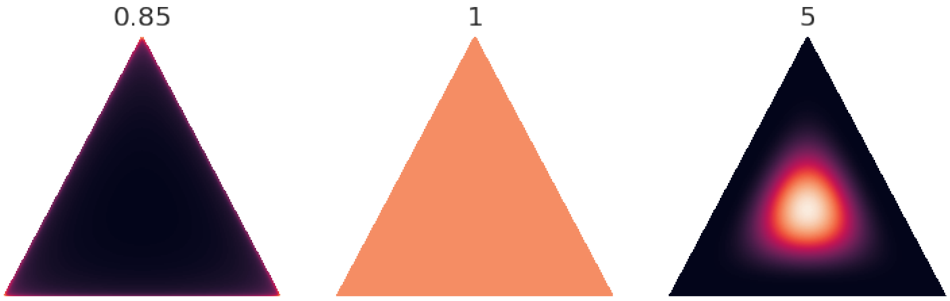

Figure 3: Dirichlet distribution on a probability simplex (here triangle) for β with different hyperparameters η. In this example, the vocabulary has size 3 and the different words are the corners of the triangles. High probability density is marked with bright color. Value of $\eta = 1$ equals a uniform distribution: given a topic, all words are equally likely to appear. Same would apply for θ with α, where corners would be different topics given a document.

handled trivially. The detailed expression is left out here since it is rigorous mathematics that does not provide any useful insights regarding the aim of the paper.

The essence, however, is that the standard Expectation Maximization algortihm cannot be applied anymore. Alternatives are using *variational inference* by approximating the fully bayesian inference or by using sampling methods, i.e. *Gibbs-Sampling* (Blei et al., 2002).

2.3.2 Analysis

LDA builds upon the pLSA algorithm by introducing prior distributions to word-topic and document-topic distributions. Although the priors are conjugate priors (Blei et al., 2002; Bishop, 2006) to the multinomial distributions, the inference process becomes much more complicated and computationally intensive.

The main question that arises is whether this price is worth paying. From a pure estimation view, the new LDA process is more Bayesian as it considers priors and also integrates them out in a fully-bayesian fashion, whilst the pLSA is more of a point estimate, and thus, more prone to over-fitting (Blei et al., 2002). Indeed, the pLSA parameter size grows linearly with the document numbers, whilst the LDA model fixes the parameter size by the number of topics and the vocabulary. The definition of the latent variable as a random variable helps significantly thereby, and results show that LDA outperforms pLSA in many areas due to its better generalizability (Blei et al., 2002).

Moreover, the major contributions of the LDA assumptions are the following ones (Blei et al., 2002):

- When a document is written, usually, only very few topics are touched and most of the remaining topics are unimportant. This is reflected in the usage of the Dirichlet prior distribution θ with parameter α. When $\alpha < 1$, only few topics are weighted higher, and the rest gets lower probabilities (sparse probability vector). The most left distribution in figure 3 shows that having the hyperparameter < 1 leads to high densities at the borders (favoring only few topics).

- Topics are focused on a few words, and most of the remaining vocabulary is usually less relevant. This is analogously reflected in the Dirichlet prior β that is parametrized by η.

In contrast to LDA, pLSA defines the given distributions as parameters. The implication is that pLSA starts its estimation by assigning some weight to all words and topics. But it is more likely that a text/document focuses on few topics and, additionally, topics are mostly condensed only on a handful of words. This makes the LDA method more robust and often reflects the reality better (Blei et al., 2002). This is also the reason why in most research applications LDA is the standard method (Mahmood, 2009; Boyd-Graber et al., 2017; Jelodar et al., 2019; Asmussen and Møller, 2019; Christophe Dupuy, 2017).

2.4 Other Extensions

Although LDA is powerful and possibly the most used TM method, it also has further extensions. This is also a major benefit of Bayesian modeling, since one is allowed to introduce as many random variables as one wishes, as long as they have a proper purpose.

One can, for example, extend the LDA approach by introducing time dependency in the topic variable (amongst others) to model topic shifts over time. This model is called the *Dynamic Topic Model* (Blei and Lafferty, 2006). (Blei and Lafferty, 2006) developed and directly used that model to analyze a topic drift in scientific publications over 100 years.

Another possible extension arrises when one wants to model topic correlations explicitly. (Blei and Lafferty, 2005) have therefore introduced *Correlated Topic Models* that are capable of coupling topics that are discovered. They show that this model gives better results in certain domains where more overlapping topics exist, or, in cases where one wants to visualize topic correlations and connections.

Given different requirements, the LDA model can always be tailored and adjusted to new tasks making it a flexible tool in Topic Modeling (Wang and McCallum, 2006; Blei and Lafferty, 2005; Liu et al., 2016; Blei and Lafferty, 2006).

One needs to keep in mind that by introducing more variables and different complex distributions, the inference process will be more difficult and less efficient, and, with a high likelihood, less accurate.

3 Applications

3.1 Overview

Topic models can be used in different ways. Beginning with unsupervised data exploration and explanation tasks, one can also extend the usage of them in terms of dimensionality reduction for further down-stream tasks (classification, clustering, document retrieval, ...) (Mahmood, 2009; Jelodar et al., 2019; Christophe Dupuy, 2017; Boyd-Graber et al., 2017).

Although TMs can be used for down-stream tasks, novel methods that are mainly based on *deep learning* are dominating the field of NLP in these (Young et al., 2018; Zhou et al., 2020).

The work in (Asmussen and Møller, 2019), in contrast, clearly shows the demand and efficiency of TMs in review, exploration and explanation tasks.

Especially in the fields of *bioinformatics, medical science, social science, political science, history*, etc. TMs are a main method to analyze large corpora of data to make them human-understandable (Jelodar et al., 2019; Boyd-Graber et al., 2017). TMs are extraordinarily useful in these cases since these fields deal with unlabeled data in huge amounts. Being able to analyze the prevalent topics, or doing trend-analyses how the fields have changed over time, are seen very often and provide valuable insights.

This section aims to give a unified view on the fields in which TMs are employed. The following subsection will thereby briefly describe the standard manual labeling process and give examples of applications from different papers.

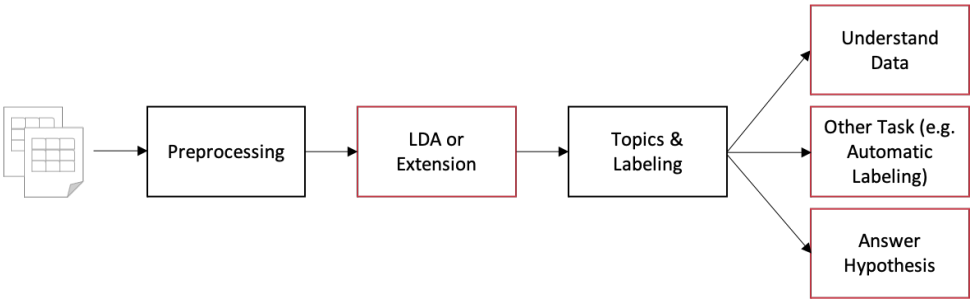

Figure 4: TM pipeline in Most Applications. The red parts indicate where differences are incorporated. In general, TM topics are used by hand-labeling the abstract topics that are probabilities of words. (Jelodar et al., 2019).

3.2 Domain Independent Procedure

Nearly all of the literature review or corpus analysis methods require to label the found topics. The reason is that topics in TMs are probability vectors of different words. Although it is already useful to select and show the most probable n words in these topics, usually the human way of looking at topics is by describing them by one term that catches the overall semantics.

Figure 5 shows how the LDA variant used in the work of (Yau et al., 2014) outputs a topic and its most likely 5 words. Afterwards, a human annotator labels this topic according to the given task that suits it most. This kind of labeling is a standard procedure in the corpus exploration/explanation field.

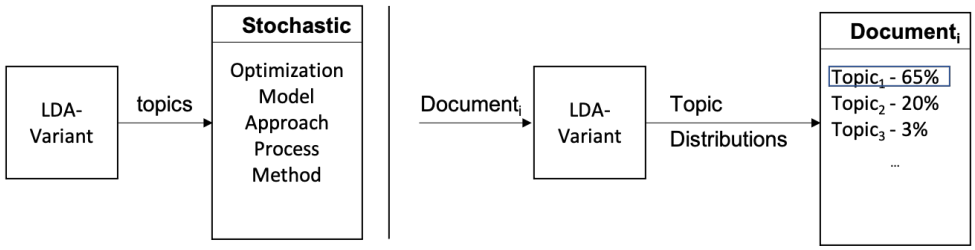

Figure 5: Topic Labeling example taken and freely adapted from (Yau et al., 2014). Left part shows topic output of the LDA system. A human annotator labeled this topic based on its most probable words. Right side shows output of the LDA algorithm given a document. Highest probable topic is assigned as label to document.

The same figure 5 further shows that (Yau et al., 2014) assigned the topic with $p > 50\%$ as a label to a document. This should give an idea how TM can also be used as a hybrid-approach to cluster/classify documents with no labels. An interested reader is referred to the work by (Xia et al., 2019; Miller et al., 2016) where additional methods are proposed that can label or directly classify documents automatically.

(Elgesem et al., 2016), for example, extracted textual data from blog posts that treat topics around surveillance with an emphasis on the Snowden affair. The paper shows how research questions/hypotheses can be posed and supported by topic models to find answers. Therefore, they analyzed the blog posts regarding the question of trustworthiness and sentiment towards Snowden in the context of PRISM. TM was used for a previous Topic trend analysis over time as part of the big picture research question. Their findings with the help of TM was that surveillance has always been a highly discussed topic. The Snowden Affair, however, made these topics more popular, shifting them to a broader audience. Additionally, they could observe a rise in topics such as data protection laws since then. The rest of their research questions is answered by other methods and thereby left out here.

A marketing focused research work from (Mathaisel and Comm, 2021) analyzed past political campaigns by Clinton, Trump, and Obama. Their main point is to present how different tools in NLP can be used to analyze past campaigns to possibly adjust future marketing strategies. Their main use case for TMs is of pure explorative and demonstrative nature where they primarily emphasize the practical use of visualization tools such as *LDAvis*, a tool that visualizes topics and their possible overlappings, and simple word clouds. Concluding their research, they finally incorporate a sentiment analysis to evaluate the audience's stance towards each candidate and their overall presence.

Further applications can be found in (Mahmood, 2009; Jelodar et al., 2019; Christophe Dupuy, 2017; Boyd-Graber et al., 2017). The point this section makes is that Topic Models are mainly an exploratory approach that put the human understanding in the focus. TMs help us to understand large corpora of unlabeled data. In scientific research, for instance, their use cases are mainly in the fields of humanities.

4 Conclusion

Topic Modelling is an unsupervised approach to analyze large collections of documents. The modelling follows the idea that certain words contribute more to certain topics. Furthermore, the second important assumption is, that documents cover different topics to different extents.

This paper elaborated on the most important TMs. Beginning with LSA, it was shown that the presented methods have an evolutionary character. Through pLSA, the shortcomings of LSA could be alleviated, and, even further, the limitations of pLSA can be mitigated by the LDA model with the help of prior probabilities. Particularly interesting is the fact that TMs can (easily) be extended when one knows how to model properly with Bayesian networks. Suitable extensions are mainly defined to solve real life data exploration tasks in different fields. This paper showed some 'paradigmatic' examples whilst highlighting the commonalities in different applications. Eventhough TMs can be used in any domain that handles large volumes of unlabeled textual data, the basic processing pipeline and approach is always similar. In particular, having extracted topics from a corpus, in most cases single term labels are assigned to them. The remaining steps differ depending on the research question.

References

Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):1–18.

N. Bassiou and C. Kotropoulos. 2011. RPLSA: A novel updating scheme for Probabilistic Latent Semantic Analysis. *Computer Speech and Language*, 25(4):741–760.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer New York LLC.

David M Blei and John D Lafferty. 2005. Correlated topic models. In *Advances in Neural Information Processing Systems*, pages 147–154.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *ACM International Conference Proceeding Series*, volume 148, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael T. Jordan. 2002. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 3, pages 993–1022.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296.

Steven L. Brunton and J. Nathan Kutz. 2019. Singular Value Decomposition (SVD). In *Data-Driven Science and Engineering*, pages 3–46. Cambridge University Press.

Christophe Dupuy. 2017. Inference and applications for topic models.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dag Elgesem, Ingo Feinerer, and Lubos Steskal. 2016. Bloggers' Responses to the Snowden Affair: Combining Automated and Manual Methods in the Analysis of News Blogging. *Computer Supported Cooperative Work: CSCW: An International Journal*, 25(2-3):167–191.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Lin Liu, Lin Tang, Libo He, Wei Zhou, and Shaowen Yao. 2016. An overview of hierarchical topic modeling. In *Proceedings - 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2016*, volume 1, pages 391–394. Institute of Electrical and Electronics Engineers Inc.

Asma Mahmood. 2009. Literature survey on topic modeling. *Technical report, Dept. of CIS, University of Delaware Newark, Delaware*.

Dennis F.X. Mathaisel and Clare L. Comm. 2021. Political marketing with data analytics. *Journal of Marketing Analytics*, 9(1):56–64.

Timothy Miller, Dmitriy Dligach, and Guergana Savova. 2016. Unsupervised document classification with informed topic models. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, page 83–91, Berlin, Germany. Association for Computational Linguistics, Association for Computational Linguistics.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. 2000. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235.

Xuerui Wang and Andrew McCallum. 2006. Topics over Time: A non-markov continuous-time model of topical trends. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2006, pages 424–433.

Linzhong Xia, Dean Luo, Chunxiao Zhang, and Zhou Wu. 2019. A Survey of Topic Models in Text Classification. In *2019 2nd International Conference on Artificial Intelligence and Big Data, ICAIBD 2019*, pages 244–250. Institute of Electrical and Electronics Engineers Inc.

Chyi Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. 2014. Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing [Review Article].

Ming Zhou, Nan Duan, Shujie Liu, and Heung Yeung Shum. 2020. Progress in Neural NLP: Modeling, Learning, and Reasoning.