Topic Modelling

> *... reviewing machine learning literature abut topic modelling*

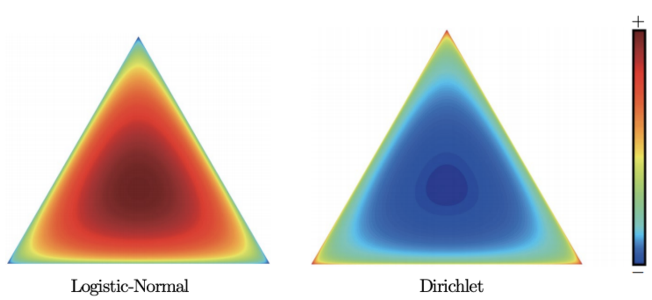
topic modelling	topic classification
• unsupervised machine learning • automatically analyses text data to determine cluster words for a set of documents • If you don't have a lot of time to analyse texts, or you're not looking for a fine-grained analysis and just want to figure out what topics a bunch of texts are talking about	• supervised machine learning techniques • automatically label a review with predefined topic tags rather than inferring what similarity cluster the review belongs to • if you have a list of predefined topics for a set of texts and want to label them automatically without having to read each one, as well as gain accurate insights
example) *"The nice thing about Eventbrite is that it's **free to use** as long as you're not **charging** for the event. There is a **fee** if you are **charging** for the event – **2.5% plus a $0.99 transaction fee.**"* By identifying words and expressions such as *free to use*, *fee*, *charging*, *2.5% plus 99 cents transaction fee*, topic modelling can group this review with other reviews that talk about similar things (these may or may not be about pricing).	*example)* *"We have the **gold level plan** and use it for everything, **love the features**! It is one of the **best bang for buck** possible."* A topic classification model that's been trained to understand these expressions (gold level plan, love the features, and best bang for buck) would be able to tag this review as topics *Features* and *Price.*
topic modelling algorithms churn out collections of expressions and words that it thinks are related, leaving you to figure out what these relations mean, while topic classification delivers neatly packaged topics, with labels such as *Price,* and *Features,* eliminating any guesswork.	

topic modelling method

1. Latent Semantic Analysis (LSA)
    - one of the NLP techniques for analysis of semantics and introduced in 2005
    - trying to dig out some meaning out of a corpus of text
    - unsupervised approach
    - helpful technique in the reduction of dimensions of the topic modelling
    - main concept is to group together all the words that have a similar meaning
    - tf-idf
        - term-frequency (TF) is a number of times keyword appears in a single document divided by the total number of words in that document $\mathrm{TF(t,d)} = \frac{\text{Number of times term "t" appears in a document}}{\text{Total Number of terms in a document "d"}}$
        - inverse-document-frequency (IDF) shows how important the term is to be in the collection of documents. calculates the weight of rare term of the text in a collection of documents $\mathrm{IDF} = \log \varepsilon \left(\frac{\text{Total number of documents}}{\text{Number of documents that contains term "t"}} \right)$
        - provide each word count and the frequency of rare words in order to provide them weights on the basis of their rarity
        - better than conventional counting of occurrence of the word as it only counts the frequency without classification
2. Latent Dirichlet Allocation (LDA)
    - uses dirichlet distribution
    - effective technique in the next word prediction
    - unsupervised approach
    - assumption is that each document mix with various topics and every topic mix with various words
    - commonly used in topic modelling
    - https://www.analyticssteps.com/blogs/introduction-latent-semantic-analysis-lsa-and-latent-dirichlet-allocation-lda

dirichlet distribution

- probability distribution
- sum of probabilities unlike the normal distribution
- continuous multivariate probability distribution
- commonly used for extraction task

	Logistic-Normal distribution does not exhibit multiple peaks at the vertices of the simplex as that in the Dirichlet distribution and as such, it is less capable to capture the multi-modality which is crucial in topic modelling

Neural Topic Modelling with Bidirectional Adversarial Training

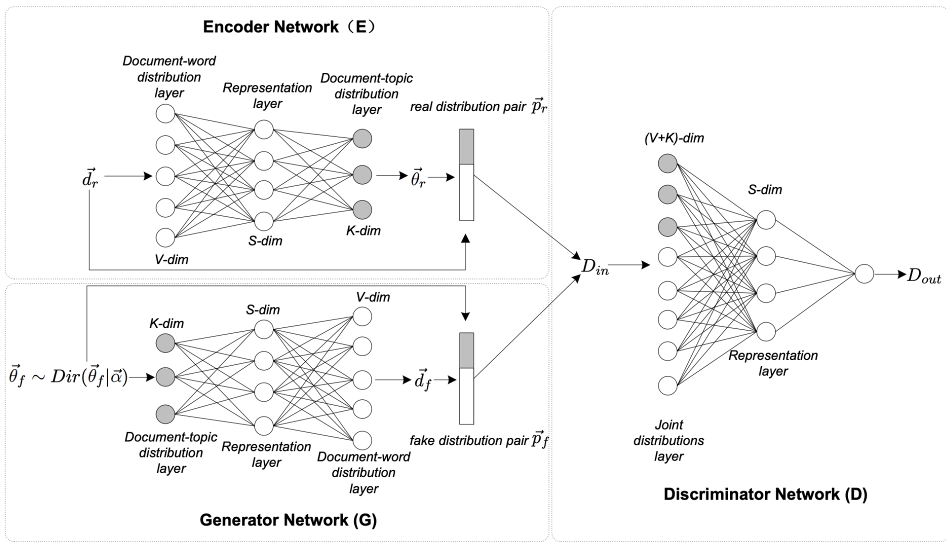
- Goal : propose a neural topic modelling approach, called Bidirectional Adversarial Topic (BAT) model, which represents the first attempt of applying bidirectional adversarial training for neural topic modelling
- Link : https://aclanthology.org/2020.acl-main.32.pdf (07/2020)
- Code : https://github.com/zll17/Neural_Topic_Models
- Credible source: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

Model

<table>
<tr><th>approaches</th><th>description</th></tr>
<tr><td>

due to the difficulty of exact inference, LDA variants require approximate inference methods - small changes to the modelling assumptions result in a re-derivation of the inference algorithm
Bidirectional Adversarial Topic model (BAT)

employs a generator network to learn the projection function from randomly-sampled document-topic distribution to document-word distribution
encoder network is used to learn the inverse projection, transforming a document-word distribution into a document-topic distribution
employs a discriminator which aims to discriminate between real distribution pair and fake distribution pair, thereby helps the networks (generator and encoder) to learn the two-way projections better
during the adversarial training phase, the supervision signal provided by the discriminator will guide the generator to construct a more realistic document and thus better capture the semantic patterns in text
the encoder network is also guided to generate a more reasonable topic distribution conditioned on specific document-word distributions
to incorporate the word relatedness information captured by word embeddings, we extend the BAT by modelling each topic with a multivariate Gaussian in the generator and propose the Bidirectional Adversarial Topic model with Gaussian (Gaussian-BAT)

</td><td>


BAT consists of three components

encoder

learns a mapping function to transform document-word distribution to document-topic distribution
takes V dimensional document representation d sampled from text corpus C as input
transform it into the corresponding K dimensional topic distribution

generator
<ol type="a">
provides an inverse projection from document-topic distribution to document-word distribution
takes a random topic distribution drawn from a dirichlet prior as input
generates a V dimensional fake word distribution

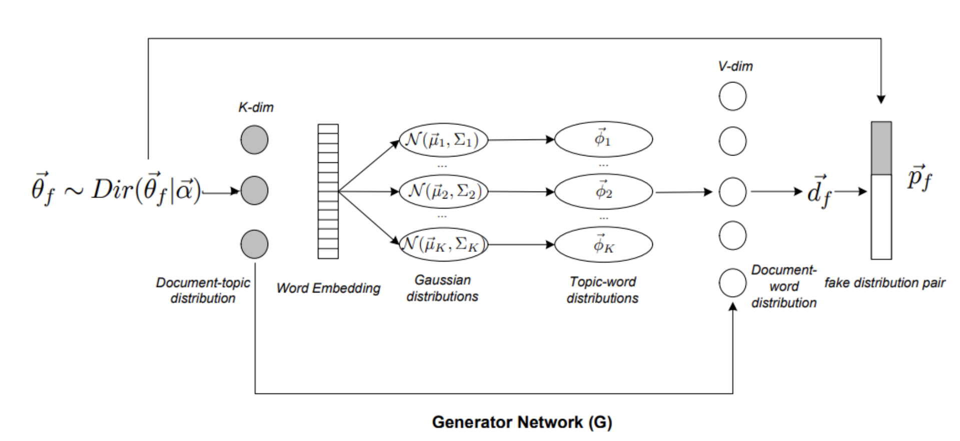
discriminator
<ol type="a">
takes the real distribution pair and fake distribution pair as input
discriminates the real distribution pairs from the fake ones
the outputs of the discriminator are used as supervision signals to learn during adversarial training

</td></tr>
</table>

<table>
<tr>
<td>

BAT with Gaussian

in BAT, the generator models topics based on the bag-of-words assumption
to incorporate the word relatedness information captured in word embeddings into the inference process, propose Gaussian-BAT
models each topic with a multivariate Gaussian

</td>
<td>


</td>
</tr>
</table>

Experiment

<table>
<tr>
<th>training data</th>
<th>experiment</th>
<th>evaluation</th>
<th>output</th>
</tr>
<tr>
<td>
three datasets used

20Newsgroups

collection of approximately 20,000 news group articles

Grolier

built from Grolier Multimedia Encycopedia
covers various fields

NYTimes

collection of news articles published between 1987 and 2007
contains a wide range of topics

use the full datasets of 20Newsgroups and Grolier
for the NYTimes dataset, randomly select 100,000 articles and remove the low frequency words

</td>
<td>
baseline model

LDA - extracts topics based on word co-occurrence patterns from documents
NVDM - an unsupervised text modelling approach based on VAE
GSM - an enhanced topic model based on NVDM
NVLDA - also built on VAE but with the logistic-normal prior
ProdLDA - a variant of NVLDA, in which the distribution over individual words is a product of experts
ATM - a neural topic modelling approach based on adversarial training

advanced model

BAT and Gaussian-BAT
Gaussian-BAT generator : pre-trained 3000-dimensional Glove embedding is used

</td>
<td>
topic coherence measures - average or median of pairwise word similarities formed by top words of a given topic

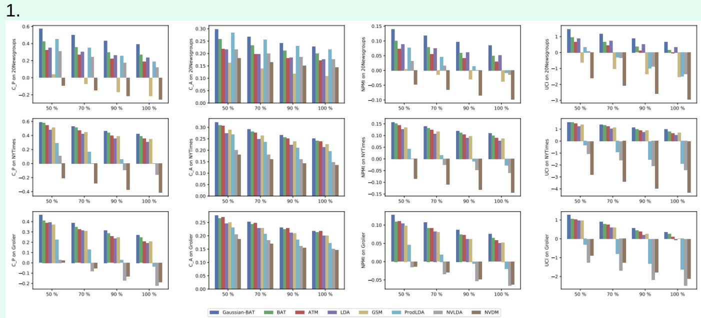
employ four topic coherence metrics (C_P, C_A, NPMI and UCI) to evaluate the topics generated by various models (Roder et al. , 2015)
each topic is represented by the top 10 words according to the topic-word probabilities
all the topic coherence values are calculated using the Palmetto library (https://github.com/dice-group/Palmetto)

evaluation

compare topic coherence vs. different topic proportions

calculate the average topic coherence values among topics whose coherence values are ranked at the top 50%, 70%, 90%, 100% position

compare the average topic coherence values numerically to show the effectiveness of proposed BAT and Gaussian-BAT

</td>
<td>
1.


Gaussian-BAT outperforms all the baselines except for Grolier dataset on C_A when considering 100% topics

2.
<table>
<tr><th>Dataset</th><th>Model</th><th>C_P</th><th>C_A</th><th>NPMI</th><th>UCI</th></tr>
<tr><td rowspan="8">20Newsgroups</td><td>NVDM</td><td>-0.2558</td><td>0.1286</td><td>-0.0984</td><td>-2.9496</td></tr>
<tr><td>GSM</td><td>-0.2318</td><td>0.1067</td><td>-0.0400</td><td>-1.6083</td></tr>
<tr><td>NVLDA</td><td>0.1205</td><td>0.1763</td><td>-0.0207</td><td>-1.3466</td></tr>
<tr><td>ProdLDA</td><td>0.1858</td><td>0.2155</td><td>-0.0083</td><td>-1.5044</td></tr>
<tr><td>LDA</td><td>0.2361</td><td>0.1769</td><td>0.0523</td><td>0.3399</td></tr>
<tr><td>ATM</td><td>0.1914</td><td>0.1720</td><td>0.0207</td><td>-0.3871</td></tr>
<tr><td>BAT</td><td>0.2597</td><td>0.1976</td><td>0.0472</td><td>0.0969</td></tr>
<tr><td>Gaussian-BAT</td><td>0.3758</td><td>0.2251</td><td>0.0819</td><td>0.5925</td></tr>
<tr><td rowspan="8">Grolier</td><td>NVDM</td><td>-0.1877</td><td>0.1456</td><td>-0.0619</td><td>-2.1149</td></tr>
<tr><td>GSM</td><td>0.1974</td><td>0.1966</td><td>0.0491</td><td>-0.0410</td></tr>
<tr><td>NVLDA</td><td>-0.2205</td><td>0.1504</td><td>-0.0653</td><td>-2.4797</td></tr>
<tr><td>ProdLDA</td><td>-0.0374</td><td>0.1733</td><td>-0.0193</td><td>-1.6398</td></tr>
<tr><td>LDA</td><td>0.1908</td><td>0.2009</td><td>0.0497</td><td>-0.0503</td></tr>
<tr><td>ATM</td><td>0.2105</td><td>0.2188</td><td>0.0582</td><td>0.1051</td></tr>
<tr><td>BAT</td><td>0.2312</td><td>0.2108</td><td>0.0608</td><td>0.1709</td></tr>
<tr><td>Gaussian-BAT</td><td>0.2606</td><td>0.2142</td><td>0.0724</td><td>0.2836</td></tr>
<tr><td rowspan="8">NYtimes</td><td>NVDM</td><td>-0.4130</td><td>0.1341</td><td>-0.1437</td><td>-4.3072</td></tr>
<tr><td>GSM</td><td>0.3426</td><td>0.2232</td><td>0.0848</td><td>0.6224</td></tr>
<tr><td>NVLDA</td><td>-0.1575</td><td>0.1482</td><td>-0.0614</td><td>-2.4208</td></tr>
<tr><td>ProdLDA</td><td>-0.0034</td><td>0.1963</td><td>-0.0282</td><td>-1.9173</td></tr>
<tr><td>LDA</td><td>0.3083</td><td>0.2127</td><td>0.0772</td><td>0.5165</td></tr>
<tr><td>ATM</td><td>0.3568</td><td>0.2375</td><td>0.0899</td><td>0.6582</td></tr>
<tr><td>BAT</td><td>0.3749</td><td>0.2355</td><td>0.0951</td><td>0.7073</td></tr>
<tr><td>Gaussian-BAT</td><td>0.4163</td><td>0.2479</td><td>0.1079</td><td>0.9215</td></tr>
</table>

Gaussian-BAT gives the best overall results across all metrics and on all the datasets except for Grolier dataset on C_A

</td>
</tr>
</table>

Repository

Topic subject creation using unsupervised learning for topic modelling

- Goal : compare Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) algorithms in the topic mining performance and propose methods to assign topic subject labels in an automated way
- Link : https://arxiv.org/abs/1912.08868 (02/2019)
- Code : -
- Credible source: -

Topic Modelling Meets Deep Neural Networks: A Survey

- Goal : provide a focused yet comprehensive overview of neural topic models for interested researchers in the AI community, so as to facilitate them to navigate and innovate in this fast-growing research area
- Link : https://arxiv.org/abs/2103.00498 (02/2021)
- Code : -
- Credible source: -