Apples to Apples: A Systematic Evaluation of Topic Models

Ismail Harrando, Pasquale Lisena and Raphaël Troncy
EURECOM, Sophia Antipolis, France
{firstname.lastname}@eurecom.fr

Abstract

From statistical to neural models, a wide variety of topic modelling algorithms have been proposed in the literature. However, because of the diversity of datasets and metrics, there have not been many efforts to systematically compare their performance on the same benchmarks and under the same conditions. In this paper, we present a selection of 9 topic modelling techniques from the state of the art reflecting a diversity of approaches to the task, an overview of the different metrics used to compare their performance, and the challenges of conducting such a comparison. We empirically evaluate the performance of these models on different settings reflecting a variety of real-life conditions in terms of dataset size, number of topics, and distribution of topics, following identical preprocessing and evaluation processes. Using both metrics that rely on the intrinsic characteristics of the dataset (different coherence metrics), as well as external knowledge (word embeddings and ground-truth topic labels), our experiments reveal several shortcomings regarding the common practices in topic models evaluation.

1 Introduction

The automatic analysis of textual data has gained increasing levels of attention over the last few decades. The cost of manually analysing and annotating the ever-growing quantity of content created and shared on the Web continues to be prohibitively expensive. Topic modelling is an NLP task where, given a corpus of documents, the objective is to find the underlying meaningful *clusters* of documents (or *topics*) that are thematically coherent (use consistent and related vocabulary) and assign each document to one or more of these topics. As a text mining technique, it allows the analysis of big volumes of textual documents through clustering them into coherent sets addressing similar subjects (or topics), and labeling them using keywords that are understandable by the end-user. It has the advantage of not relying on any labeled data to achieve good results, as the training of topic models is done in an unsupervised matter. Moreover, the resulting topics and representations can then be used to perform other NLP tasks such as trend prediction (Lau et al., 2012), text summarization (Lin and Hovy, 2000), improving named entity recognition (Newman et al., 2006), and content recommendation (Papneja et al., 2021).

Because of the unsupervised nature of the task, the evaluation of the quality of topic modelling techniques relies usually on metrics that do not require human annotation or ground-truth labels. Most of the used "coherence" metrics – further detailed in Section 3.1 – attempt to measure how much the resulting topics reflect some statistical characteristics of the original dataset and its word co-occurrences distribution. These metrics utilise different definitions of what a "coherent topic" is, and they only contingently agree with humans judgement (Chang et al., 2009). Coupled with the different approaches for document preprocessing and the variety of used evaluation datasets, this complexity leads to several nuances in the evaluation process that are not widely acknowledged in the literature at large. Thus, comparisons can be inconsistent and sometimes misleading.

In this work, we selected a diverse array of topic modelling algorithms (probabilistic, algebraic, embedding-based and neural) from the literature and we provide a thorough comparison using a unified evaluation protocol. This protocol evaluates each topic model on several datasets, using a variety of metrics that range from intrinsic evaluation of the clustering quality to ones that assess the alignment between the extracted topics and the human-assigned labels. With this strategy, we aim to illustrate the inconsistency of these metrics when

Proceedings of Recent Advances in Natural Language Processing, pages 483–493
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_055

varying several subtle evaluation conditions. We analyse the results and we discuss the differences in performances across the different algorithms, datasets and parameters.

The remainder of this paper is organised as follows. In Section 2, we describe some related work, detailing some state-of-the-art topic modelling techniques. Different metrics for evaluating topic models are introduced in Section 3, while Section 4 describes the datasets we use for this purpose. In Section 5, we extensively analyse 9 topic models using coherence and ground truth related metrics. Finally, we provide some conclusions in Section 6.

2 Related Work

2.1 Topic Modelling Techniques

One of the first yet still widely used techniques is **Latent Dirichlet Allocation (LDA)** (Blei et al., 2003), an unsupervised statistical modelling approach that considers each document as a *bag of words* and creates a randomly assigned document-topic and word-topic distributions. Iterating over words in each document, the distributions are updated according to the probability that a document or a word belongs to a certain topic. The **Hierarchical Dirichlet Process (HDP)** model (Teh et al., 2006) considers instead each document as a group of words belonging with a certain probability to one or multiple components of a mixture model, i.e. the topics. Both the probability measure for each document (distribution over the topics) and the base probability measure – which allows the sharing of clusters across documents – are drawn from Dirichlet Processes (Ferguson, 1973). Unlike most other topic models, HDP infers the number of topics automatically. **Gibbs Sampling for a DMM (GSDMM)** applies the Dirichlet Multinomial Mixture model for short text clustering (Yin and Wang, 2014). This algorithm works by computing iteratively the probability that a document join a specific one of the N available clusters. This probability consists of two parts: 1) a part that promotes the clusters with more documents; 2) a part that advantages the movement of a document towards similar clusters, i.e. which contains a similar word-set.

Recently, pre-trained Word vectors such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) have been used to help to enhance topic-word representations, as achieved by the **Latent Feature Topic Models (LFTM)** (Nguyen et al., 2015). One of the LFTM algorithms is *Latent Feature LDA (LF-LDA)*, which extends the original LDA algorithm by enriching the topic-word distribution with a latent feature component composed of pre-trained word vectors. In the same vein, the **Paragraph Vector Topic Model (PVTM)** (Lenz and Winker, 2020) uses doc2vec (Le and Mikolov, 2014) to generate document-level representations in a common embedding space. Then, it fits a Gaussian Mixture Model to cluster all the similar documents into a predetermined number of topics – i.e. the number of GMM components.

Topic modelling can also be performed via linear algebraic methods. Starting from the high-dimensional term-document matrix, multiple approaches can be used to lower its dimensions. Then, we consider every dimension in the lower-rank matrix as a latent topic. A straightforward application of this principle is the **Latent Semantic Indexing model (LSI)** (Deerwester et al., 1990), which uses Singular Value Decomposition as a means to approximate the term-document matrix (potentially mediated by TF-IDF) into one with fewer rows – each one representing a latent semantic dimension in the data – and preserving the similarity structure among columns (terms). **Non-negative Matrix Factorisation (NMF)** (Paatero and Tapper, 1994) exploits the fact that the term-document matrix is non-negative, thus producing not only a denser representation of the term-document distribution through the matrix factorisation but guaranteeing that the membership of a document to each topic is represented by a positive coefficient.

In recent years, neural network approaches for topic modelling have gained popularity giving birth to a family of **Neural Topic Models (NTM)** (Cao et al., 2015). Among those, **doc2topic (D2T)**[1] uses a neural network which separately computes N-dimensional embedding vectors for words and documents (with N = number of topics) before computing the final output using a sigmoid activation. The distributions topic-word and document-topic are obtained by getting the final weights on the two embedding layers. The **Contextualized Topic Model (CTM)** (Bianchi et al., 2020) uses Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) – a neural transformer language model designed to compute sentences representations efficiently – to generate

[1] `https://github.com/sronnqvist/doc2topic`

a fixed-size embedding for each document to contextualise the usual Bag of Words representation. CTM enhances the *Neural-ProdLDA* (Srivastava and Sutton, 2017) architecture with this contextual representation to significantly improve the coherence of the generated topics.

2.2 Topic Models Comparison

To the best of our knowledge, no extensive comparison of recent topic models – covering multiple metrics and datasets under the same preprocessing condition – has been made. Some previous works have tried to compare different topic models on certain datasets and metrics. A review of statistical topic modelling techniques is included in Newman et al. (2006). Schofield and Mimno (2016) provide a comparison resulting from the effect of preprocessing on the performance of LDA on multiple corpora. Jelodar et al. (2017) offer a survey of topic modelling techniques based on LDA, as well as their different applications in recent literature. Yi and Allan (2009) and Alexander and Gleicher (2016) compare several topic models, evaluated as tools for performing Information Retrieval downstream tasks such as *Topic Alignment*, *Change Comparison*, *Document Retrieval* and *Query Expansion*. Several evaluation metrics based on top-words analysis was suggested by Newman et al. (2010). Alghamdi and Alfalqi (2015) compare 4 topic models (LDA, LSI, PLSA and CTM): this survey studied both their capability in modelling static topics, as well as in detecting topic change over time, highlighting the strengths and weaknesses of each. Burkhardt and Kramer (2019) provide a survey for the adjacent task of multi-label topic models, underlining its challenges and promising directions. Qiang et al. (2020) give an extensive performance evaluation of multiple topic models in the context of the *Short Text Topic modelling* sub-task (e.g. tweets). Finally, Doogan and Buntine (2021) studied several topic model coherence measures to assess how informative they are in several applied settings revolved around interpretability as an objective. They showed how standard coherence measures may not inform the most appropriate topic model or the optimal number of topics when measured up against human evaluation, thus challenging their utility as quality metrics in the absence of ground truth data.

2.3 Metrics

While our work utilises multiple comparison metrics (detailed in Section 3.1), it is worth highlighting that many other evaluation metrics were proposed in the literature to expose different characteristics of the studied topic models such as Classification Accuracy and Perplexity (Qiang et al., 2020), Entropy and Held-out Likelihood (Schofield and Mimno, 2016), Stability (Alexander and Gleicher, 2016), and Top-word Ranking (Greene et al., 2014), whereas finding a universally useful metric for topic modelling evaluation is still an open problem (Blei, 2012; Doogan and Buntine, 2021; Hoyle et al., 2021).

3 Metrics

The evaluation of machine learning techniques often relies on accuracy scores computed comparing predicted results against a ground truth. In the case of unsupervised techniques like topic modelling, the ground truth is not always available. For this reason, in the literature, we can find:

- metrics which enable to evaluate a topic model independently from a ground-truth, among which, coherence measures are the most popular ones (Röder et al., 2015; O'Callaghan et al., 2015; Qiang et al., 2020);
- metrics that measure the quality of a model's predictions by comparing its resulting clusters against ground truth labels, in this case a topic label for each document.

3.1 Coherence Metrics

The coherence metrics rely on the joint probability $P(w_i, w_j)$ of two words w_i and w_j that is computed by counting the number of documents in which those words occur together divided by the total number of documents in the corpus. The documents are fragmented using sliding windows of a given length, and the probability is given by the number of fragments including both w_i and w_j divided by the total number of fragments. This probability can be expressed through the *Pointwise Mutual Information (PMI)*, defined as:

$$PMI(w_i, w_j) = log\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \tag{1}$$

A small value is chosen for ϵ, in order to avoid computing the logarithm of 0. Different metrics based on PMI have been introduced in the literature, differing in the strategies applied for token segmentation, probability estimation, confirmation measure, and aggregation. The **UCI coherence** (Röder et al., 2015) averages the PMI computed between pairs of topics, according to:

$$C_{UCI} = \frac{2}{N\cdot(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j) \quad (2)$$

The **UMASS coherence** (Röder et al., 2015) relies instead on a different joint probability:

$$C_{UMASS} = \frac{2}{N\cdot(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} log \frac{P(w_i,w_j)+\epsilon}{P(w_j)} \quad (3)$$

The **Normalized Pointwise Mutual Information (NPMI)** (Chiarcos et al., 2009) applies the PMI in a confirmation measure for defining the association between two words:

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-log(P(w_i, w_j) + \epsilon)} \quad (4)$$

NPMI values go from -1 (never co-occurring words) to +1 (always co-occurring), while the value of 0 suggests complete independence. The most common implementation of C_{NPMI} applies NPMI as in Eqn (4) to couples of words, computing their joint probabilities using sliding windows.

This measure can be applied also to word sets. This is made possible using a vector representation in which each feature consists in the NPMI computed between w_i and a word in the corpus W, according to the formula:

$$\vec{v}(w_i) = \left\{ NPMI(w_i, w_j) | w_j \in W \right\} \quad (5)$$

The vectors related to each word of the topic are then compared using the cosine similarity C_V.

Fang et al. (2016) introduce **Word Embeddings-based Coherence**. This metric relies on pre-trained word embeddings such as GloVe or word2vec and evaluates the topic quality using a similarity metric between its top words. In other words, a high mutual embedding similarity between a model's top words reflects its underlying semantic coherence. In this paper, we will use the sum of mutual cosine similarity computed on the Glove vectors[2] of the top 10 words of each topic.

$$C_{WE} = \frac{2}{N\cdot(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} cos(v_i, v_j) \quad (6)$$

where v_i and v_j are the GloVe vectors of the words w_i and w_j.

In practice, these metrics are computed at the topic level and then aggregated using the arithmetic mean, in order to provide a coherence value for the whole model.

3.2 Metrics Which Relies on a Ground Truth

The most used metric that relies on a ground truth is the **Purity**, defined as the fraction of documents in each cluster with a correct prediction (Hajjem and Latiri, 2017). A prediction is considered correct if the original label coincides with the original label of the majority of documents falling in the same topic prediction. Given L the set of original labels and T the set of predictions:

$$Purity(T, L) = \frac{1}{|T|} \sum_{i \in T} \max_{j \in L} |T_j \cap L_j| \quad (7)$$

Other metrics are used in the literature for evaluating the quality of classification or clustering algorithms, applied to the topic modelling task:

1. **Homogeneity**: a topic model output is considered homogeneous if all documents assigned to each topic belong to the same ground-truth label (Rosenberg and Hirschberg, 2007);

2. **Completeness**: a topic model output is considered complete if all documents from one ground-truth label fall into the same topic (Rosenberg and Hirschberg, 2007);

3. **V-Measure**: the harmonic mean of Homogeneity and Completeness. A V-Measure of 1.0 corresponds to a perfect alignment between topic model outputs and ground truth labels (Rosenberg and Hirschberg, 2007);

4. **Normalized Mutual Information (NMI)** is the ratio between the mutual information between two distributions – in our case, the prediction set and the ground truth – normalised through an aggregation of those distributions' entropies (Lancichinetti et al., 2009). The aggregation can be realised by selecting the minimum/maximum or applying the geometric/arithmetic mean. In the case of arithmetic mean, NMI is equivalent to the V-Measure.

In this work, we use their implementations as provided by scikit-learn (Pedregosa et al., 2011).

[2] We use a Glove model pre-trained on Wikipedia 2014 + Gigaword 5, available at https://nlp.stanford.edu/projects/glove/

4 Datasets

In this section, we introduce the datasets that we use in our experiments. The features of each dataset are reported in Table 1.

A common pre-processing is performed on the datasets before training, consisting of:

- Removing numbers, which, in general, do not contribute to the broad semantics of the document;
- Removing the punctuation and lower-casing the text;
- Removing the standard English stop words;
- Lemmatisation using Wordnet, to deal with inflected forms as they are a single semantic item;
- Ignoring words with 2 letters or less. In facts, they are mainly residuals from removing punctuation – e.g. stripping punctuation from *people's* produces *people* and *s*.

The same pre-processing is also applied to the text before topic prediction.

4.1 20 NewsGroups

The 20 NewsGroups collection (20NG) (Lang, 1995) is a popular dataset used for text classification and clustering. It is composed of English news documents, distributed fairly equally across 20 different categories according to the subject of the text. We use a reduced version of this dataset[3], which excludes all the documents composed by the sole header while preserving an even partition over the 20 categories. This reduced dataset contains 11,314 documents. We pre-process the dataset to remove irrelevant metadata – consisting of email addresses and news feed identifiers – keeping just the textual content.

4.2 Agence France Presse

The Agence France Presse (AFP) publishes daily up to 2000 news articles in 5 different languages[4], together with some metadata represented in the NewsML XML-based format. Each document is categorised using one or more subject codes, taken from the IPTC NewsCode Concept vocabulary[5]. In the case of multiple subjects, they are ordered by relevance. In this work, we only consider the first level of the hierarchy of the IPTC subject codes. We extracted a subset containing 125,516 news documents in English released in 2019.

4.3 Yahoo! Answers Comprehensive Q&A

The Yahoo! Answers Comprehensive Q&A (later simply *Yahoo*) contains over 4 million questions and their answers, as extracted from the Yahoo! Answers website[6]. Each question comes with metadata such as title, date, and category, as well as a list of user-submitted answers. We construct documents by concatenating the title, body and best answer for each question – following Zhang et al. (2015) – and preprocess the documents in the same way as mentioned above. Then we create 2 subsets:

- ***Yahoo balanced***, in which each category is represented by the same number of documents (1000) for a total of 26,000 documents;
- ***Yahoo unbalanced***, in which the number of documents sampled from each category is proportional to its presence in the overall dataset, for a total of 22,121 documents.

These two subsets have been realised having a number of documents of the same order of magnitude. This allows to compare the differences in performance with balanced and unbalanced sets.

Table 1 summarises the properties of these datasets. The datasets present multiple differences, namely the size, the length of the documents and the distribution of documents per topic (i.e. ground truth label).

5 Experiment and Results

Evaluating an unsupervised task such as Topic Modelling is inherently challenging, and despite the variety of metrics, it is still an open problem (Hoyle et al., 2021). While intrinsic metrics (coherence) try to measure the underlying quality of the topical clusters generated by each model, they do not always match with human judgement. Two very coherent topics (according to the metric) can still fall under the same topic label for a human, and vice-versa. Topic models aim to maximise the posterior probability of a document belonging to a coherent topic, regardless of how it maps to human-perceived categories. For instance, *Christianity* and *Atheism* can be both filed as two independent topics or one topic (*religion*) by a human annotator, and while neither arbitrary option is wrong, it constitutes a big difference to how we would evaluate the topic modelling algorithms. They have no means

[3] https://github.com/selva86/datasets/

[4] http://medialab.afp.com/afp4w/

[5] http://cv.iptc.org/newscodes/subjectcode/

[6] https://answers.yahoo.com

Dataset	# Documents	# Labels	# Documents/label (std)	Document Length (std)
20 NewsGroups	11314	20	565 (56)	122 (241)
AFP	125516	17	4932 (8920)	242 (234)
Yahoo! Answers (balanced)	26000	26	1000 (0)	43 (47)
Yahoo! Answers (unbalanced)	22121	26	850 (726)	43 (46)

Table 1: Characteristics of the datasets being studied: number of documents per dataset, number of ground-truth labels, average number (and standard deviation) of documents per label and the average (and standard deviation) length of documents per dataset.

of inferring what humans find to be *topically distinct* beyond co-occurrence statistics, making the comparison to human-annotated labels (as a "gold standard") quite insufficient. Because of these challenges, few works in the literature (O'Callaghan et al., 2015; Alexander and Gleicher, 2016; Alghamdi and Alfalqi, 2015; Qiang et al., 2020) go beyond simple comparisons that only use one metric or dataset, eclipsing merits and shortcomings of the other methods. We attempt to provide a more thorough comparison using multiple evaluation datasets – varying in size, document length, number of topics, and label distribution – and metrics from the literature as a step towards a better understanding of the available options and their usability for different potential use-cases.

5.1 Varying the datasets

This section reports a comparison between 9 topic modelling algorithms described in Section 2. Our experimental setup goes as follows:

- For each dataset, we pre-process every document using the process described in Section 4;
- We train each topic model on each dataset, selecting the hyper-parameters through an optimisation process based on grid search, in order to maximise the C_{NPMI} score. The use of a coherence metric as an optimisation objective is justified by the common use-case scenario, in which ground-truth labels are not present. The full set of parameters is documented in the repository[7];
- For each trained model, we compute all the intrinsic (coherence) metrics and the ground-truth-based ones.

For the experiment, we rely on *ToModAPI* (Lisena et al., 2020), an open-source topic modelling API that is built to easily train, evaluate and compare several topic models. This framework provides a common interface for training, performing topic inference, and evaluating using coherence and ground truth. It includes all the metrics described above.

The number of topics – which must be provided in input to the algorithm for training – has been set to 20, 17 and 26 respectively when training on 20NG, AFP, and Yahoo, to mimic the original number of labels in each corpus. HDP has not been concerned with the choice of the number of topics, because it automatically infers it. For the first two datasets, we perform another training using the same hyper-parameters but increasing the number of topics to 50, to study its effect on the performance on the various metrics.

While all the obtained results are available in the appendix[8], we will report in Figure 1 a selection of the most noticeable scores, namely C_{NPMI}, Word Embeddings coherence and V-Measure.

C_{NPMI} values are in line with all the other coherence metrics in terms of ranking (listed in the appendix for brevity), i.e. LDA shows consistently good coherence scores across all datasets, followed by NMF and PVTM.

For the CTM model, we obtained a significantly lower coherence value than the one reported by Bianchi et al. (2020). Further investigation and experiments revealed the impact of an additional preprocessing step which reduces the vocabulary to the 2000 most frequent words. This further preprocessing improves the NPMI score of CTM from -0.028 to 0.116, while lowering the one of LDA from 0.133 to 0.126. This confirms the limits of topic modelling comparison and enforces the call for a standard procedure.

Word embeddings coherence demonstrated a better correlation with human judgement (Fang et al., 2016). Unsurprisingly, the two models that rely on word embeddings (LFTM, PVTM) tend to perform notably better (Figure 1).

The V-measure results included in Figure 1 are particularly relevant for understanding the correlation between the predicted topics and the ground

[7] https://github.com/D2KLab/ToModAPI/blob/master/params.md

[8] https://github.com/D2KLab/ToModAPI/blob/master/appendix.pdf

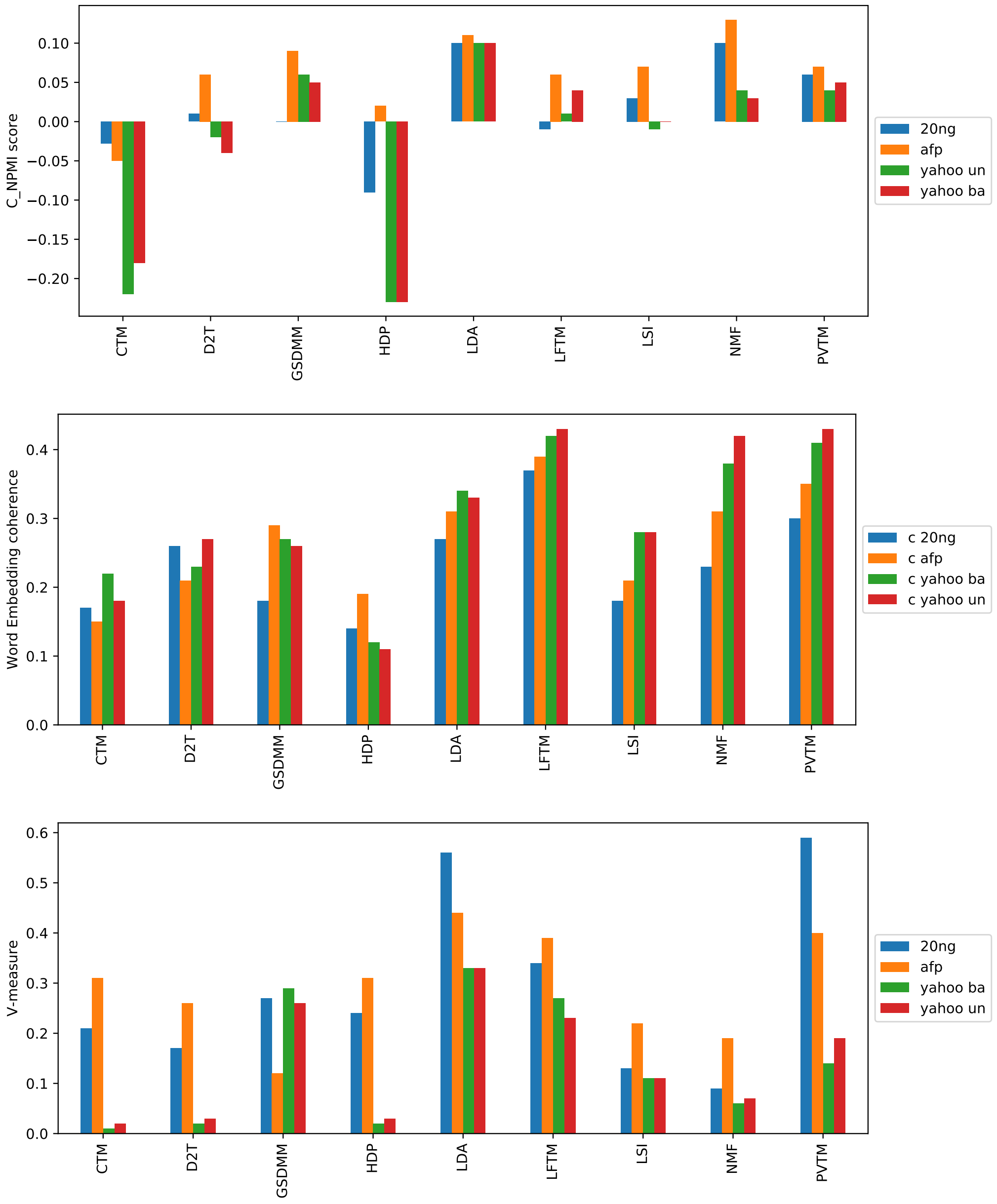

Figure 1: NPMI, Word embedding coherence and V-measure across the models trained on the different datasets.

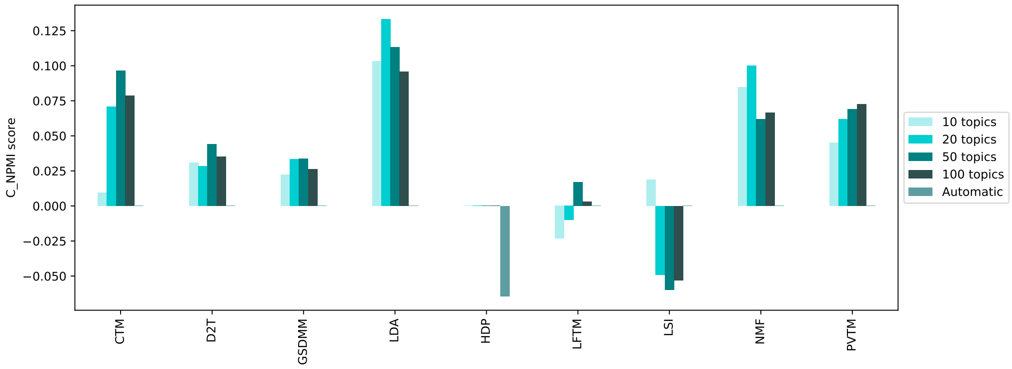

Figure 2: NPMI of each model on the 20NG dataset when varying the number of topics.

truth, as it summarises three metrics – homogeneity, completeness and purity. This metric relies on human choices – made either by the editors for AFP or the website users for 20NG and Yahoo – and so it approximates the correlation between the topics as decided by the algorithms and the human (subjective) judgement on the same matter. Again, LDA is leading in overall performances, while other models – LFTM, PVTM, GSDMM – have good scores on particular datasets. The Yahoo dataset is particularly challenging for all models (the maximum value for V-measure is 0.33 for LDA), as compared to AFP (0.55 for LDA) or 20NG (0.59 for PVTM). This is probably due to a combination of document length, noise and errors in user-submitted content, and the potential overlap in topics[9]. Increasing the number of topics systematically improves the results on AFP, raising the Homogeneity and Purity scores. This happens because the more a topic is granular, the highest is the chance that it maps correctly to the human label is correct. However, this is not observed on 20NG. Given the difference in size between 20NG and AFP, we conclude that the dimension of the former is not allowing it to extract smaller coherent topics, but rather causes an over-specialisation of them.

In summary, LDA still achieves the best scores overall, being often the first (or among the firsts) in ranking for every metric, whereas the other algorithms excel in particular contexts and can be specifically suitable for a given dataset. Increasing the number of topics is particularly helpful on bigger datasets, as it allows the topic models to find smaller yet more coherent subtopics within the collection, avoiding the drawback effect of being too specific. About label balance as tested through the Yahoo dataset, it appears that the balancing in the dataset has not a large impact in final results. On the contrary, training on the unbalanced version is often producing better coherence and V-measure. The reason can be found in the complete dropping of smaller categories, thus reducing the number of classes and achieving a higher-scoring topic/label mapping.

5.2 Varying the number of topics

To evaluate the effect of the choice of the number of topics (usually unknown beforehand), we train our models – except HDP, which infers the number of topics automatically – on 20NG using the same hyperparameters and varying only the number of topics. The results are shown in Table 2.

NPMI	Mean (std)	Max	Min
HDP	-0.176 (0.09)	-0.06	-0.28
LDA	0.120 (0.01)	0.133	0.101
NMF	0.083 (0.01)	0.102	0.063
PVTM	0.054 (0.01)	0.061	0.046

Table 2: The effect of random seeds on the NPMI for some models trained on 20NG

While there is a slight yet consistent improvement in the NPMI score for PVTM, we observe that increasing the number of topics does not consistently improve or hurt the coherence of the produced models. The fact that the score for 20 topics is usually the highest is probably due to the model finetuning, applied on this configuration. Finetuning every model for every number of topics requires a study of the co-optimisation of hyperparameters, which is out of the scope of this paper.

5.3 Varying the seed

For the models which allows to configure the random seed, we perform the evaluation on 20NG using the same hyperparameters except the seed (which we varied to have the values from 1 to 5). Even among 5 runs, we observe quite some variance in the metrics that is purely due to randomness which can be quite substantial. We report these results in Figure 2. While the effect is not very pronounced, it can be misguiding. We thus recommend for topic models relying on random initialization to evaluate their models using different seeds, to guarantee a statistically significant comparison.

6 Conclusions and Future Work

In this work, we empirically evaluated 9 topic modelling algorithms using different coherence and ground-truth-based metrics on 3 text corpora reflecting a variety of properties, using a common evaluation framework. The results reveal several differences between the trained models, which obtain better or worse performances depending on the evaluation setting. Among these, LDA proves to be the most consistent performer overall, while embedding-based models prove to be less prone to generating meaningless topics.

The task of evaluating topic models remains a challenging one because of the inherent lack of a

[9] Some examples are “News & Events”/“Politics and Government”, “Dining Out”/“Food & Drink”, and “Business and Finance”/“Local Businesses”

ground-truth, the subjectivity of what constitutes a "coherent topic", and the variety of settings wherein it is used. While every newly proposed topic model claims to improve on the existing state-of-the-art under some specific conditions, it is a worthwhile effort to revisit those claims and review them on a broader set of challenges and a unified pipeline, revealing their strengths and shortcomings. We also hope that by showing that no single metric can reflect the overall performance of any given topic model, we join a growing number of words drawing attention to the brittleness of most automatic metrics for topic models and the need of re-evaluating the standard practices of evaluation in the topic modelling literature.

As an extension to this work, we intend to study how other factors such as language, preprocessing and dataset characteristics can influence the performance on the metrics, as well as develop a unified protocol for evaluation that can allow us to draw more interesting insights into how the different topic modelling approaches fare in real use cases and downstream applications.

Acknowledgement This work has been partially supported by the French National Research Agency (ANR) within the ASRAEL project (grant number ANR-15-CE23-0018), by the European Union's Horizon 2020 research and innovation program within the Odeuropa project (grant agreement No. 101004469), and by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

References

Eric Alexander and Michael Gleicher. 2016. Task-Driven Comparison of Topic Models. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):320–329.

Rubayyi Alghamdi and Khalid Alfalqi. 2015. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *arXiv preprint arXiv:2004.03974*.

David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Sophie Burkhardt and Stefan Kramer. 2019. A Survey of Multi-Label Topic Models. *SIGKDD Explor. Newsl.*, 21(2):61–79.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A Novel Neural Topic Model and Its Supervised Extension. In *AAAI Conference on Artificial Intelligence*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 288–296.

Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede. 2009. *Von der Form zur Bedeutung: Texte automatisch verarbeiten - From Form to Meaning: Processing Texts Automatically*. Narr Francke Attempto Verlag GmbH + Co. KG.

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.

Caitlin Doogan and Wray Buntine. 2021. Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data. In *39^{th} International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1057–1060, New York, NY, USA. Association for Computing Machinery.

Thomas S. Ferguson. 1973. A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.*, 1(2):209–230.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How Many Topics? Stability Analysis for Topic Models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Berlin, Heidelberg. Springer Berlin Heidelberg.

Malek Hajjem and Chiraz Latiri. 2017. Combining IR and LDA Topic Modeling for Filtering Microblogs. In *21^{st} International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)*, volume 112, pages 761 – 770, Marseille, France.

Alexander Miserlis Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. *CoRR*, abs/2107.02173.

Hamed Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. 2017. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *CoRR*, abs/1711.04305.

Andrea Lancichinetti, Santo Fortunato, and János Kertész. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Ken Lang. 1995. NewsWeeder: Learning to Filter Netnews . In *20th International Conference on Machine Learning*, pages 331 – 339, San Francisco, USA. Morgan Kaufmann.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online. In *International Conference on Computational Linguistics (COLING)*, pages 1519–1534, Mumbai, India. The COLING 2012 Organizing Committee.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *31st International Conference on Machine Learning Research*, volume 32, pages 1188–1196, Bejing, China. PMLR.

David Lenz and Peter Winker. 2020. Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE*, 15(1):1–18.

Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *18th Conference on Computational Linguistics - Volume 1*, COLING '00, page 495–501, USA. Association for Computational Linguistics.

Pasquale Lisena, Ismail Harrando, Oussama Kandakji, and Raphael Troncy. 2020. ToModAPI: A Topic Modeling API to Train, Use and Compare Topic Models. In *2nd International Workshop for Natural Language Processing Open Source Software (NLP-OSS)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *26th International Conference on Neural Information Processing Systems (NIPS)*, volume 2, pages 3111–3119, Lake Tahoe, NV, USA. Curran Associates Inc.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. In *Intelligence and Security Informatics*, pages 93–104, Berlin, Heidelberg. Springer.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 100–108, USA. Association for Computational Linguistics.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645 – 5657.

Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Sachin Papneja, Kapil Sharma, and Nitesh Khilwani. 2021. Content Recommendation Based on Topic Modeling. In *Computational Methods and Data Engineering*, pages 1–10, Singapore. Springer Singapore.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *8th ACM International Conference on Web Search and Data Mining (WSDM)*, page 399–408, New York, USA. ACM.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. In *ICLR*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Xing Yi and James Allan. 2009. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Advances in Information Retrieval*, pages 29–41, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jianhua Yin and Jianyong Wang. 2014. A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering. In *20^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 233–242, New York, USA. ACM.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 649–657.