Evaluating the Robustness of Embedding-based Topic Models to OCR Noise

Elaine Zosa[1 [0000−0003−2482−0663]], Stephen Mutuvi[2,3 [0000−0002−3067−9806]], Mark Granroth-Wilding[1,4], and Antoine Doucet[2 [0000−0001−6160−3356]]

[1] University of Helsinki
[2] University of La Rochelle, L3i laboratory, France
[3] Multimedia University of Kenya, Nairobi, Kenya
[4] Silo AI

Abstract. Unsupervised topic models such as Latent Dirichlet Allocation (LDA) are popular tools to analyse digitised corpora. However, the performance of these tools have been shown to degrade with OCR noise. Topic models that incorporate word embeddings during inference have been proposed to address the limitations of LDA, but these models have not seen much use in historical text analysis. In this paper we explore the impact of OCR noise on two embedding-based models, Gaussian LDA and the Embedded Topic Model (ETM) and compare their performance to LDA. Our results show that these models, especially ETM, are slightly more resilient than LDA in the presence of noise in terms of topic quality and classification accuracy.

Keywords: Topic modelling · Word embeddings · OCR noise

1 Introduction

Large-scale collections of historical documents are becoming more accessible to researchers due to the efforts made to digitize these materials. Digitization pipelines commonly involve passing the material through an optical character recognition (OCR) engine which outputs text that can be used for downstream tasks. Due to various factors such as the printing quality of the original material, font, and layout styles, the output of OCR engines varies in quality. OCR errors stemming from this process can have a significant impact when downstream natural language processing (NLP) tools are used to analyse this data.

Topic modelling is a method to extract latent topics in a collection of documents. It is a popular approach in Digital Humanities and data-driven historical research to analyse large historical collections such as newspaper archives [18, 16, 9], academic journals [11] and handwritten diaries [3]. Probabilistic topic models such as the Latent Dirichlet Allocation (LDA) [2] model a topic as a distribution over a vocabulary and a document as a mixture of topics. Prior research quantifying the impact of OCR noise on topic modelling shows that the topics and topic mixtures deteriorate in quality as the level of noise increases [17, 12].

Word embeddings are distributed representations of words in a dense vector space that encode their usage in a corpus [10, 14]. They can capture both syntactic and semantic attributes of words such that words that typically occur in similar contexts are

in close proximity to each other in the embedding space. Approaches that combine topic modelling with word embeddings to improve the semantic coherence of topics and address the challenge of scaling topic models to large vocabularies include Gaussian LDA [4], spherical Hierarchical Dirichlet Process (sHDP) [1], and the Embedded Topic Model (ETM) [5]. GLDA and ETM are LDA-like models that use word embeddings and have shown improved topic quality over LDA on clean datasets.

Non-embedded topic models like LDA use word co-occurence statistics to discover latent topics in a corpus and the negative impact of OCR noise on topic modelling is due to the distortion of the word distributions when words are misspelled [17]. In embedding-based models, word identities are replaced with word *embeddings* that, in principle, can be more resilient to OCR noise, provided misspellings of the same word cluster together in the embedding space. There is, however, no existing work that investigates the robustness of these models on data with OCR noise and whether they show any improvement over LDA.

In this paper we conduct a quantitative assessment of the performance of two embedding-based models, GLDA and ETM, on datasets with OCR noise. Our aim is to test whether embedding-based models can be used to improve the analysis of digitised historical documents.

2 Related Work

Latent Dirichlet Allocation (LDA) [2] is a probabilistic topic modelling method for extracting topics from a document collection. It models a topic as a probability distribution over a fixed vocabulary and a document as a mixture of topics. LDA relies on the co-occurrence of the words in the documents to infer the latent topics and topic mixtures of the documents.

Models that use word embeddings have been proposed to improve topic quality and handle out-of-vocabulary words. Gaussian LDA (GLDA) [4] is the first LDA-based topic model that directly incorporates word embeddings during topic inference. Instead of treating topics as categorical distributions over the vocabulary, GLDA characterizes topics as multivariate Gaussian distributions over the word embedding space whose mean and variance are estimated during topic inference. Words are ranked according to their probability density under the posterior-predictive distribution given the training corpus.

In the Embedded Topic Model (ETM) [5], topics and words share the same embedding space and a topic is a point in the embedding space called a topic embedding. Words are generated from a categorical distribution whose natural parameter is the inner product of the word embeddings associated with a topic and the respective topic embedding. The most probable words in the topic are those with embeddings that are close to the topic embedding.

Various studies have evaluated the impact of OCR errors on unsupervised topic modelling. A comparative study of document clustering and topic modelling on OCRed text indicated that OCR noise had a greater performance impact on topic modelling than on document clustering [17]. Another evaluation revealed that while OCR noise resulted in lower topic coherence, it had little impact on model stability [12]. A more

general study on the impact of noisy OCR on historical text analysis using a corpus of eighteenth-century texts found that topics extracted from OCRed texts aligned well with topics from the gold standard texts although the authors hinted that the topic model had trouble with poetry-adjacent topics [7]. These previous evaluations, however, focused on well-established topic models based on word co-occurrence and as far as we are aware embedding-based models have not been tested to analyse OCR-ed data.

3 Methodology

Following [17], we first evaluate the topic models on a corpus of historical documents with real OCR noise that have aligned gold standard (GS) texts. Then we evaluate the models on a larger corpus where synthetic noise has been introduced at increasing levels.

3.1 Datasets

Real noise The Overproof dataset consists of 30,301 digitised news articles from the Sydney Morning Herald 1842–1954, from the archives of the National Library of Australia [6][5]. The articles were processed using the ABBYY FineReader OCR tool and additional corrections were done using crowd-sourced annotations. The OCRed articles have a word error rate (WER) of 25% [13]. The OCR and GS articles are aligned on a character level.

Synthetic noise To generate data with synthetic noise, we start with a clean dataset and gradually corrupt the data by introducing noise at increasing levels. We use the Reuters RCV1 dataset as the clean dataset. This consists of over 800K English news wire articles with assigned categorical labels [8]. We use a reduced dataset of 50K articles sampled from the largest categories.

We follow a procedure that generates synthetic noise based on a noise model constructed from a dataset with real noise [17]. To build a noise model, we construct a matrix $\mathbf{M}$ where $\mathbf{M}_{x,y}$ is the number of times character x in a GS article is confused with character y in the corresponding OCR article.

To generate parameterised noise, we interpolate the matrix $\mathbf{M}$ such that $\mathbf{M}_\gamma = \gamma\mathbf{M} + (1-\gamma)\mathbf{I}$ where γ is the interpolation parameter. When $\gamma = 0$, no noise is introduced, while at $\gamma = 1.0$, the interpolated matrix is equivalent to $\mathbf{M}$. We generate corrupted datasets from the Reuters corpus with γ ranging from 0 to 1 in increments of 0.2. This resulted in datasets with character error rates (CER) of 0%, 7%, 14%, 21%, 28% and 35%. Table 1 summarizes the datasets used in our experiments.

3.2 Model training and word embeddings

We use LDA as our baseline model. We trained LDA models using the Gensim library[6], leaving the prior parameters to be inferred during training. For ETM, we used the authors'

[5] http://overproof.projectcomputing.com/datasets/
[6] https://radimrehurek.com/gensim

	#types	#tokens	#art.
Overproof OCR	1.3M	10M	30K
Overproof GS	414K	9.8M	30K
Reuters	414K	12.4M	50K

Table 1: Datasets used in the experiments.

implementation[7] with default hyperparameters. For GLDA, we used the `gaussianlda` package, which implements the algorithm in Python [8]. We ran the sampler for 20 iterations, based on initial experiments with the clean *20-Newsgroups* dataset.

In our experiments with real noise data, we experimented with two different types of word embeddings: (1) pre-trained GloVe embeddings trained on English Wikipedia and Gigaword [14][9]; and (2) word2vec embeddings [10] trained on the Overproof data (we trained separate embeddings for the OCR and GS portions of the data). This is to investigate whether word embeddings trained on a large amount of clean data result in better topic models than embeddings trained on more limited and noisier data. On experiments with synthetic data, we used word2vec embeddings trained on the corrupted Reuters data. We trained separate embeddings for each noise level.

We trained topic models with 50 topics on the OCR and GS portions of the Overproof data and 100 topics for each noise setting of the synthetic Reuters data. To account for the randomness inherent in the models we repeated each experiment ten times and report the averaged results.

3.3 Evaluation measures

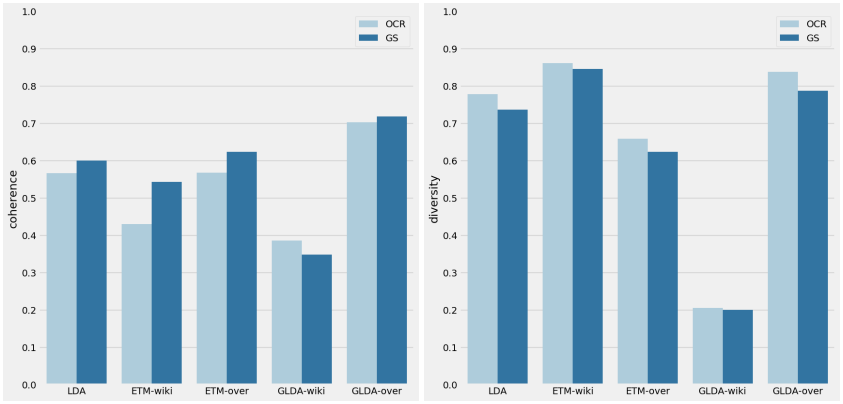
Topic coherence Topic coherence quantifies the interpretability of a topic as represented by its most probable terms. We use the C_v coherence measure [15] implemented in the Gensim package [10].

Topic diversity Models that learn more diverse topics are preferable to models with redundant topics. We measure topic diversity as the proportion of unique words out of all the top words representing all the topics in the model [5]. For topic coherence and diversity, we evaluate on the top 20 terms of each topic.

Classification accuracy We evaluate the quality of the per-document topic proportions inferred by the models through a supervised document classification task. We train a classifier on a portion of the data using the inferred topic proportions as features and pre-assigned categories as labels, then test the classifier on the unseen portion. As this evaluation requires gold standard labels, we only run this evaluation on the Reuters dataset with synthetic noise. We used a logistic regression classifier with ten-fold cross-validation in our evaluation.

[7] https://github.com/adjidieng/ETM
[8] https://pypi.org/project/gaussianlda/
[9] https://nlp.stanford.edu/projects/glove/
[10] https://radimrehurek.com/gensim/models/coherencemodel.html

4 Results and Discussion



(a) Topic coherence (b) Topic diversity

Fig. 1: Performance on the Overproof dataset averaged over 10 runs. *wiki* models use word embeddings trained on Wikipedia while *over* models use embeddings trained on the Overproof data.

4.1 Performance on Real Noise

Figure 1 shows the results of our experiments on real noise data. In terms of topic coherence, almost all the models perform better on the GS documents than the OCR documents, as would be expected (Figure 1a). GLDA with Overproof embeddings is the best performing model while GLDA with Wikipedia embeddings is the worst-performing. ETM with Overproof embeddings has similar topic coherences to LDA–both have a coherence of 0.57 for OCR while for GS, ETM is only a little better with a mean coherence of 0.62 and LDA has 0.6. ETM with Wikipedia embeddings performs worse than LDA with a coherence of 0.43 for OCR and 0.54 for GS.

These results indicate that for embedding-based topic models, it is preferable to use embeddings trained on the target corpus rather than on a general-knowledge dataset like Wikipedia, despite the latter being larger in size and cleaner, especially when the target corpus is a specialized document collection, such as historical documents. One reason for this could be that Wikipedia is a modern dataset while the Overproof corpus is made up of articles from the mid-nineteenth to the mid-twentieth century.

Now we take a closer look at the characteristics of the topics produced by one run of each of the models (Table 2). We focus on ETM and GLDA with Overproof embeddings. We see that the most coherent ETM and LDA topics are more coherent than the GLDA topics despite GLDA having the best mean topic coherence overall. GLDA is known to produce qualitatively different topics from LDA [4] and we notice that it also produces qualitatively different topics from ETM. Another difference is that topics produced by ETM on the OCR documents show a high degree of correspondence with

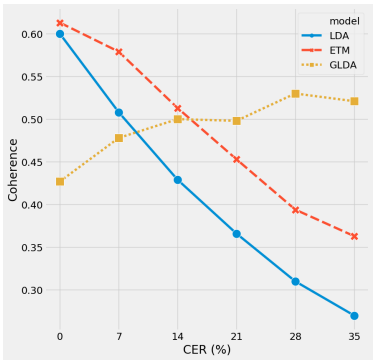
Topic No.	*Top Words*	*Coh*
LDA-OCR		
33	petitioner, respondent, nisi, decree, honor, formerly, appeared, ground, marriage, granted	0.95
8	club, match, team, cricket, played, play, runs, first, association, matches	0.83
LDA-GS		
11	petitioner, marriage, decree, respondent, formerly, nisi, appeared, married, ground, granted	0.95
9	accused, prisoner, charged, guilty, charge, court, trial, stealing, months, sessions	0.82
ETM-OCR		
31	respondent, petitione, nisi, appeared, honor, formerly, decree, ground, issue, foi	0.91
9	charged, court, fined, john, police, prisoner, two, sentenced, months, guilty	0.81
ETM-GS		
21	petitioner, marriage, appeared, formerly, respondent, decree, ground, nisi, married, granted	0.95
41	match, cricket, team, played, wickets, runs, play, second, first, club	0.88
GLDA-OCR		
12	managers, woiking, administrator, guidance, servlco, goneral, publicity, lenders, bown	0.73
38	accompanying, pipers, recoived, governors, alio, transmitted, photographs, btato, lag	0.73
GLDA-GS		
47	parent, outset, sult, cardiff, terror, dawn, tha, alley, biggest, sweepin	0.72
1	discontinued, livered, forcibly, blacksmith, extracted, interrupted, reopened, sampson, tempted	0.72

Table 2: Most coherent topics from LDA, ETM, and Gaussian LDA on the Overproof dataset.

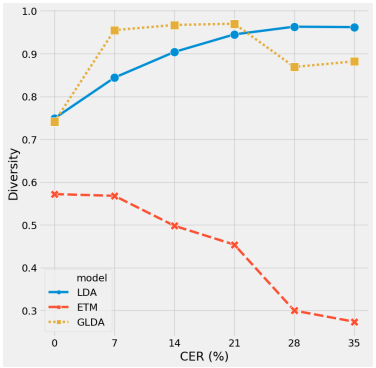
topics from the GS data, while the same cannot not be said of the GLDA topics. For instance, Topic 31 of ETM-OCR and Topic 21 of ETM-GS are topics on legal matters and show many overlapping terms (they share 17 of their top 20 terms). We found no such correspondences with the GLDA topics.

In terms of topic diversity, OCR topics are more diverse than GS topics for all models (Figure 1b). We hypothesize that this is primarily due to the higher vocabulary size of the OCR documents resulting from misspellings. While the training data used for word embeddings has a high impact on the coherence of the embedding-based models, it does not seem have a significant influence on topic diversity. ETM with Wikipedia embeddings has the most diverse topics while GLDA with the same Wikipedia embeddings has the most redundant topics.

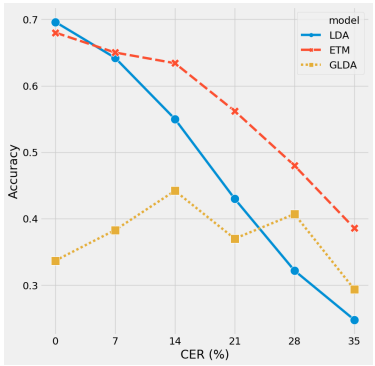
4.2 Performance on synthetic noise



(a) Topic coherence



(b) Topic diversity



(c) Mean classification accuracy

Fig. 2: Performance on the Reuters data with synthetic noise averaged across 10 runs.

Experimental results on the synthetic data are shown in Figure 2. ETM and LDA degrade linearly in coherence as noise increases, but ETM is more resilient than LDA (Figure 2a). Interestingly, GLDA improves in coherence as noise increases. We think one reason for this is that as an effect of the nature of GLDA topics, which are unimodal distributions in the embedding space, GLDA topics have the tendency to cluster misspelled words together. This leads to topics that, while having a high coherence score, are not qualitatively meaningful.

With regards to topic diversity, our results show that ETM produces less diverse topics than LDA or GLDA at all noise levels (Figure 2b), corroborating our results in the real noise data (Figure 1b). As noise increases, ETM topics become even less diverse (at 35% CER, diversity is at 0.27, 0.88, and 0.96 for ETM, GLDA and LDA, respectively).

It is surprising therefore to find that even though ETM has the lowest topic diversity, it performs better than LDA and GLDA in the document classification task (Figure 2c). On further investigation we found that for ETM and LDA, the topics that are most relevant in document classification are diverse and, for the most part, are preserved across noise levels while the redundant topics tend to have smaller weights that do not impact the classification performance significantly.

5 Conclusions

In this paper we assessed the impact of real and synthetic OCR noise on two embedding-based topic models, Gaussian LDA and ETM, with LDA as our baseline. We also experimented with different word embeddings for GLDA and ETM.

On real noise, GLDA is the best-performing model in terms of topic coherence while. ETM peforms as well as LDA. ETM and GLDA produce more diverse topics than LDA. We note, however, that GLDA produces qualitatively different topics than ETM and LDA. Our experiments on synthetic data revealed that ETM performed better than LDA in terms of topic coherence and classification accuracy across noise levels. On the other hand, GLDA improved in topic coherence with increased noise and produced more varied topics but performed worse in document classification because its topics do not correlate with the gold standard labels in the dataset.

LDA is a popular method for analysing digitised historical collections but it is not without its shortcomings, especially when applied to documents with OCR errors. In our experiments, we have shown that topic models that incorporate information from word embeddings improve slightly over LDA in the presence of OCR noise in terms of coherence, diversity, and document classification.

Acknowledgements

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

1. Batmanghelich, K., Saeedi, A., Narasimhan, K., Gershman, S.: Nonparametric spherical topic modeling with word embeddings. In: Proceedings of the conference. Association for Computational Linguistics. Meeting. vol. 2016, p. 537. NIH Public Access (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Blevins, C.: Topic modeling martha ballard's diary. Cameron Blevins (2010)
4. Das, R., Zaheer, M., Dyer, C.: Gaussian lda for topic models with word embeddings. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 795–804 (2015)
5. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020)

6. Evershed, J., Fitch, K.: Correcting noisy ocr: Context beats confusion. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. pp. 45–51 (2014)
7. Hill, M.J., Hengchen, S.: Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. Digital Scholarship in the Humanities **34**(4), 825–843 (2019)
8. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research **5**(Apr), 361–397 (2004)
9. Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., Tolonen, M.: Topic modelling discourse dynamics in historical newspapers. arXiv preprint arXiv:2011.10428 (2020)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Mimno, D.: Computational historiography: Data mining in a century of classics journals. Journal on Computing and Cultural Heritage (JOCCH) **5**(1), 1–19 (2012)
12. Mutuvi, S., Doucet, A., Odeo, M., Jatowt, A.: Evaluating the impact of ocr errors on topic modeling. In: International Conference on Asian Digital Libraries. pp. 3–14. Springer (2018)
13. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Nguyen, N.V., Doucet, A.: Deep statistical analysis of ocr errors for effective post-ocr processing. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 29–38. IEEE (2019)
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
15. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015)
16. Viola, L., Verheul, J.: Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. Digital Scholarship in the Humanities (2019)
17. Walker, D., Lund, W.B., Ringger, E.: Evaluating models of latent document semantics in the presence of ocr errors. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 240–250 (2010)
18. Yang, T.I., Torget, A., Mihalcea, R.: Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 96–104 (2011)