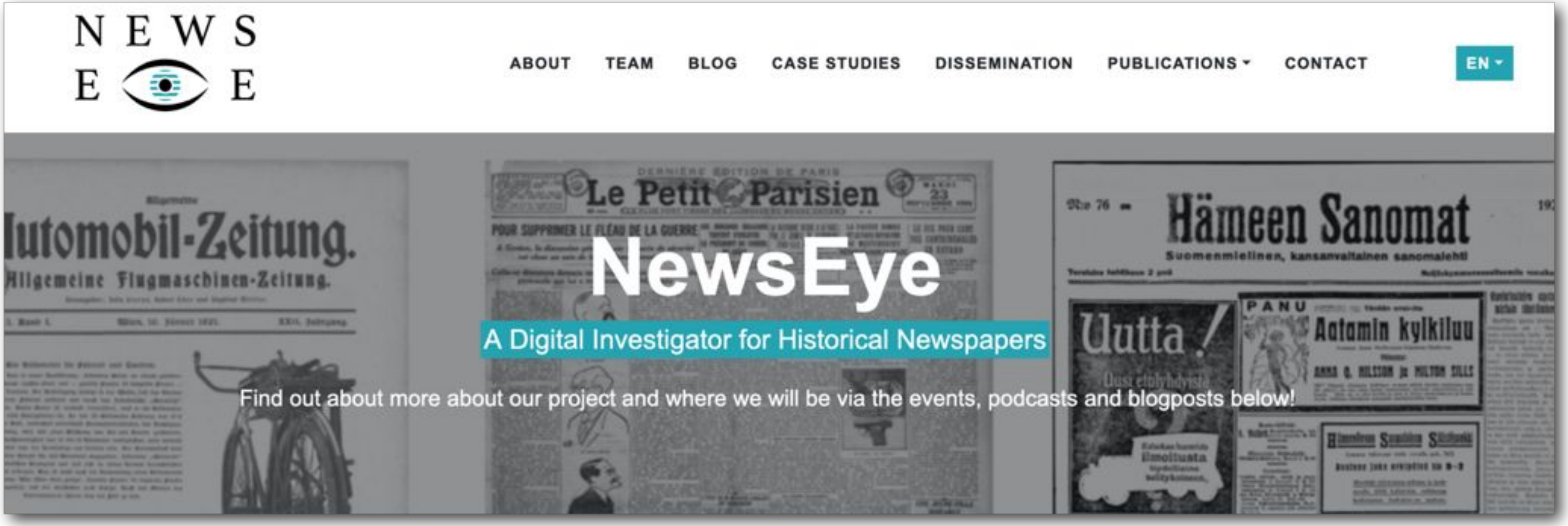
LT Seminar: Topic Modelling and Historical Newspapers

22 October 2020

Elaine Zosa
University of Helsinki

newseye.eu



NewsEye

A Digital Investigator for Historical Newspapers

Find out about more about our project and where we will be via the events, podcasts and blogposts below!

Which historical newspapers?

- Many newspapers across a long periods of time and in different languages:
  - National Library of Finland (NLF), 1859-1918
  - National Library of France (BnF), 1915-1945
  - National Library of Austria (ÖNB), 1895-1937

Why topic modelling?

- Extract themes in a collection or subset of the collection depending on the research question (refugees in WWI, women’s suffrage movement)
- Link articles that share similar topics in the same language or across different languages (PLTM)
- Trace evolution of topics across time (DTM)
- Track prominence of a topic across time (Topics Over Time, DTM, LDA, …)
- And many others...

Two recent works

- Disappearing Discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections
- Investigating the robustness of embedding-based topic models to OCR Noise

Disappearing discourses:

Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections

Elaine Zosa, Lidia Pivovarova, Simon Hengchen,
Jani Marjanen & Mikko Tolonen

The NewsEye project has received funding from the
European Union's Horizon 2020 research and innovation
programme under grant agreement no. 770299.

Historian’s dilemma:

How to be true to the past and relevant for the present?

The NewsEye project has received funding from the
European Union’s Horizon 2020 research and innovation
programme under grant agreement no. 770299.

- Data-driven text-mining methods allow – *more than ever before* – a historian to disregard the present perspective
- With textual data, we study discourses, not things.

⇨ Can we study discourses in data without predefining them? What is the process of interpreting discourse dynamics?

The NewsEye project has received funding from the
European Union’s Horizon 2020 research and innovation
programme under grant agreement no. 770299.

Data

- Finnish-language digitized newspaper collection from the National Library of Finland (NLF)
- Focus on time period with more data (1854-1917)
- We subsampled the collection to get equal data for each year for training
- Then used trained models to infer topic proportions of documents for the whole data set
- Topic proportions are used to determine the prevalence of a topic in a specific time slice

NEWS EYE

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Methods

- Idea: We can trace things that disappear by topic modelling
- Topics ≠ Discourses
- Testing two methods: LDA and DTM

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

What is LDA?

Latent Dirichlet Allocation (LDA) is a probabilistic model that extracts prevalent themes (“topics”) from a set of documents. If trained on different time slices, there is no guarantee it is possible to match topics across time.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation.
Journal of machine Learning research, 3(Jan), pp.993-1022

NEWS EYE

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

What is DTM?

Dynamic topic model (DTM) is another type of topic model that has the advantage of taking time into account: the data is divided into discrete time slices and the method infers topics aligned across these time slices to capture *topics evolving over time*.

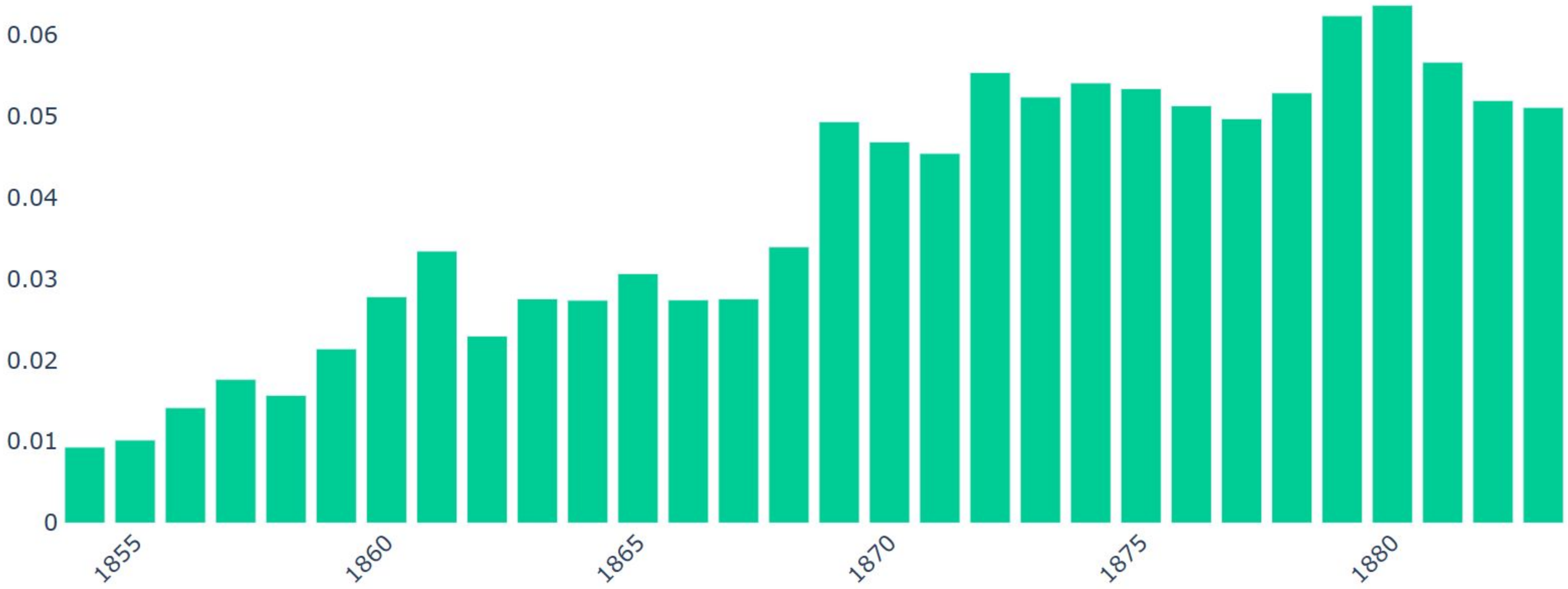
Blei, D.M. and Lafferty, J.D., 2006, June. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

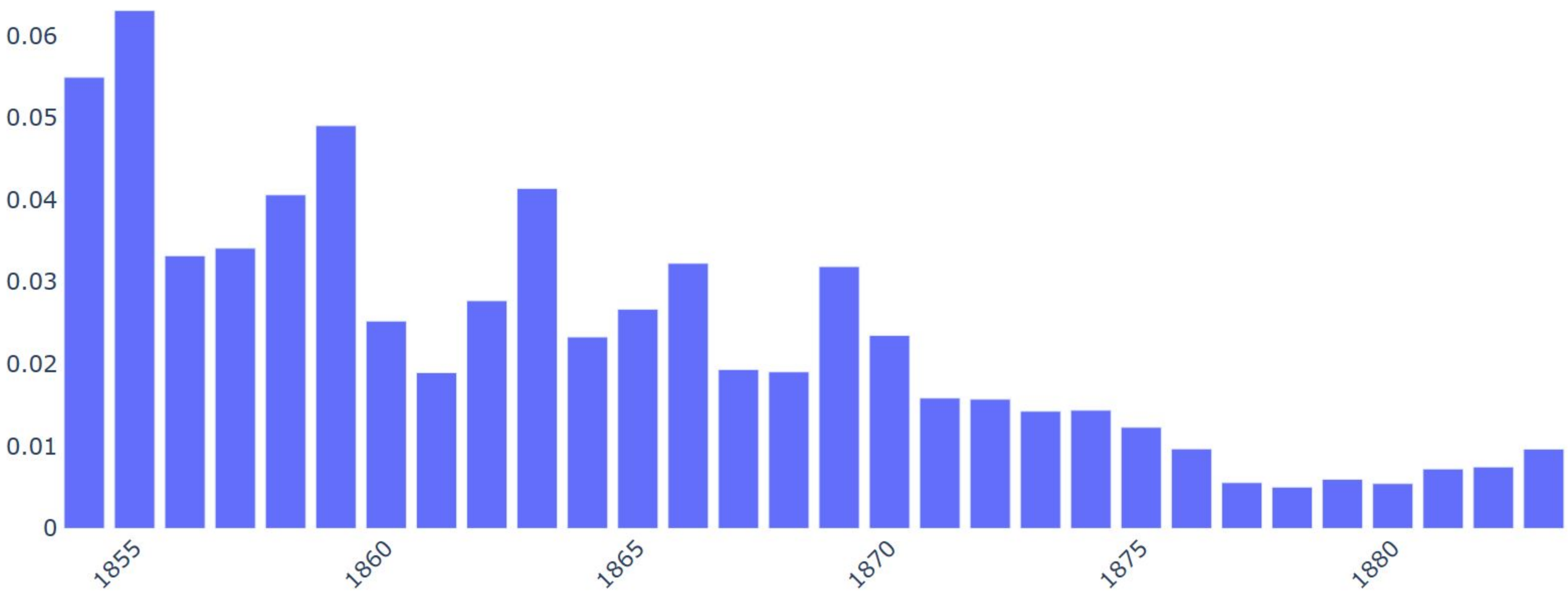
DTM Results

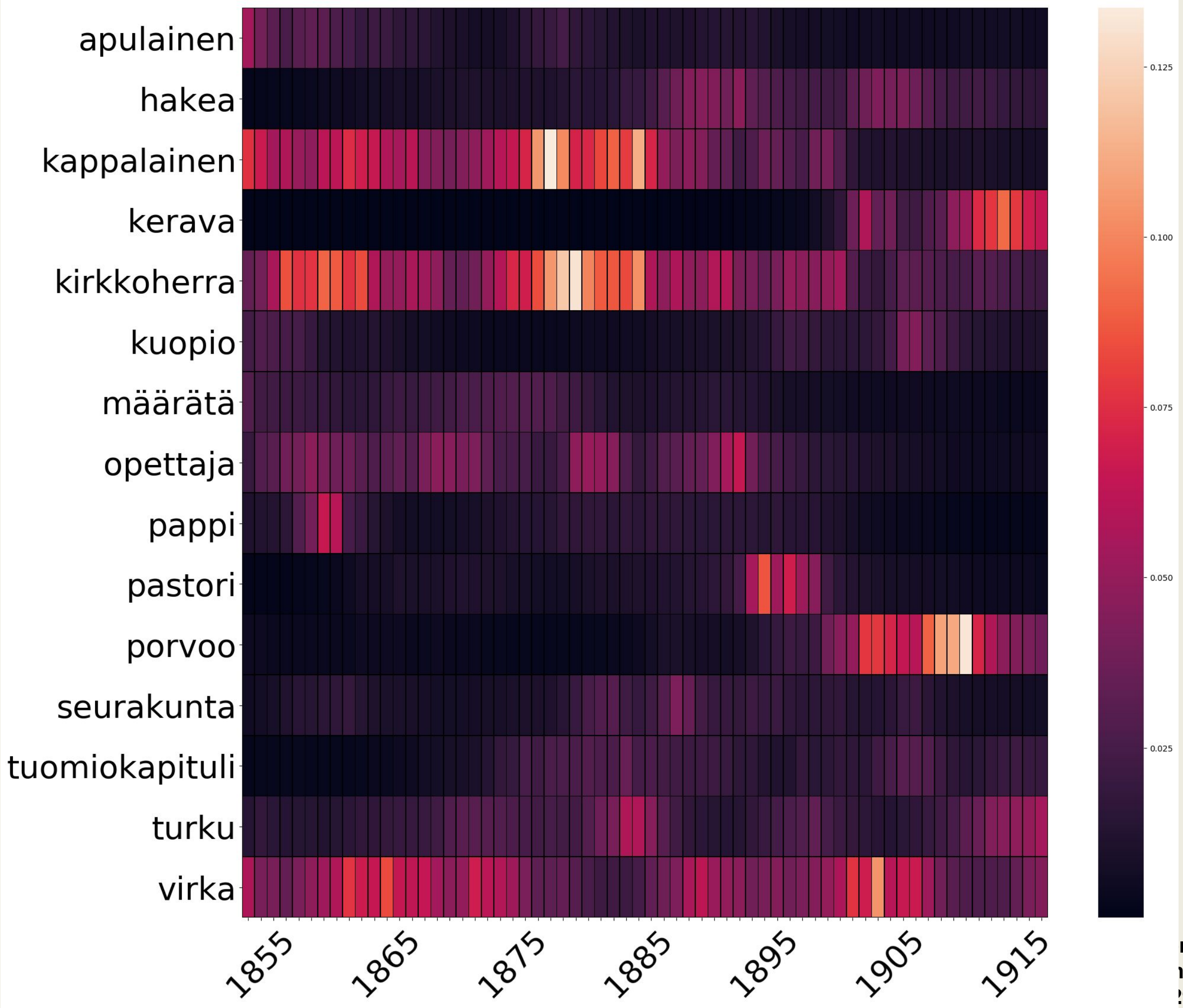
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Topic 13: Education topic on the rise

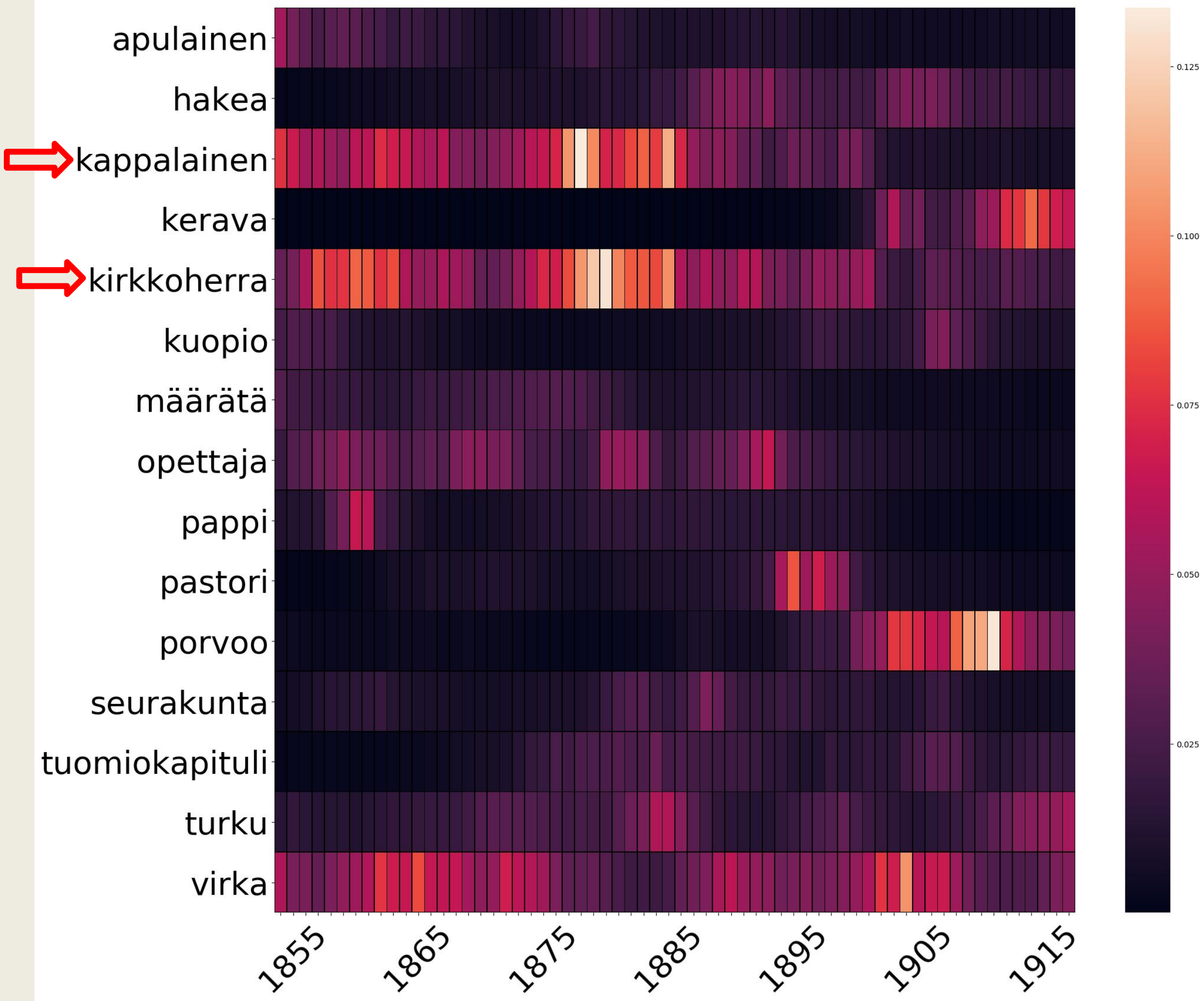


Topic 11: A religious topic: chaplain, priest and office

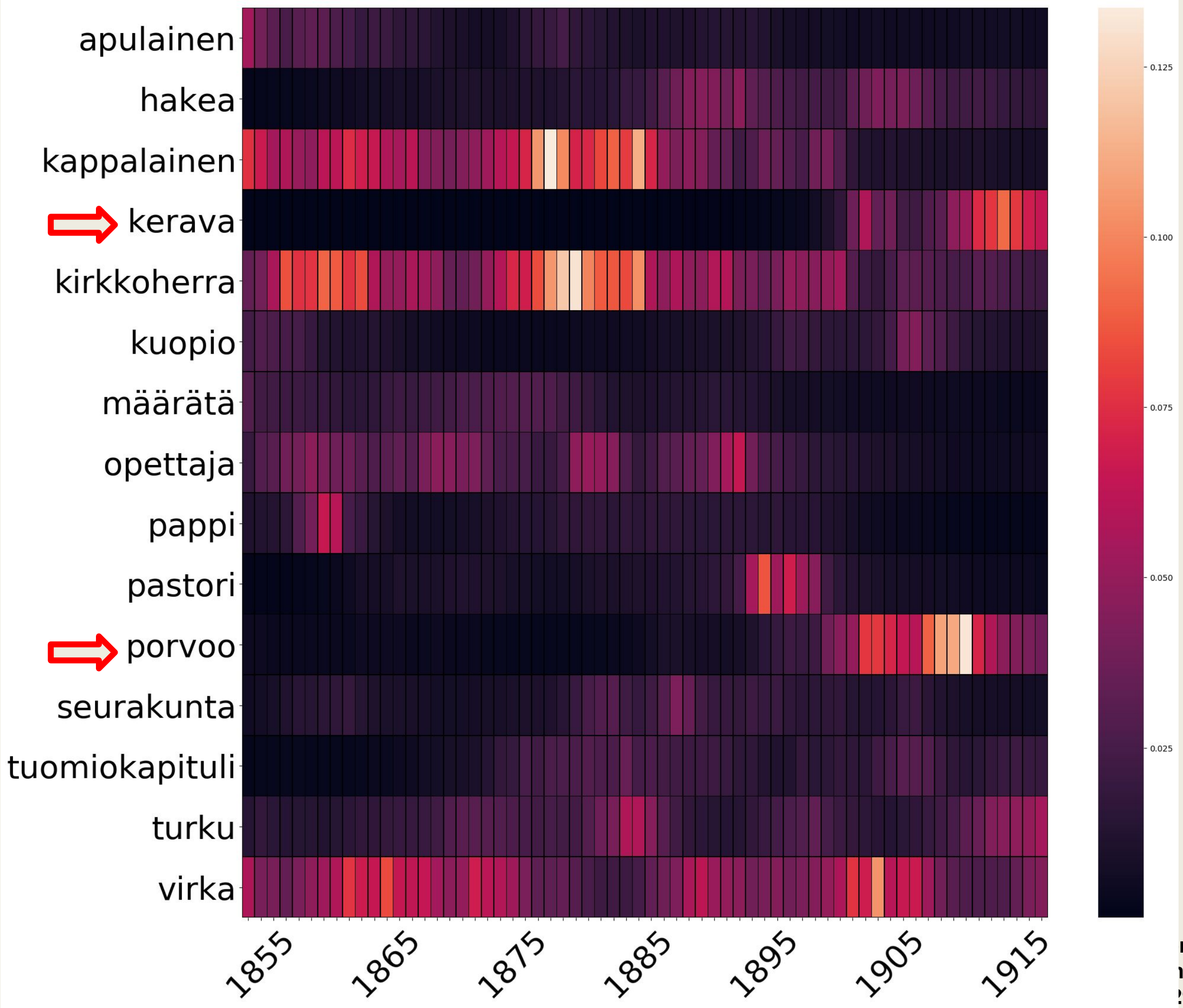




m the
novation
299.

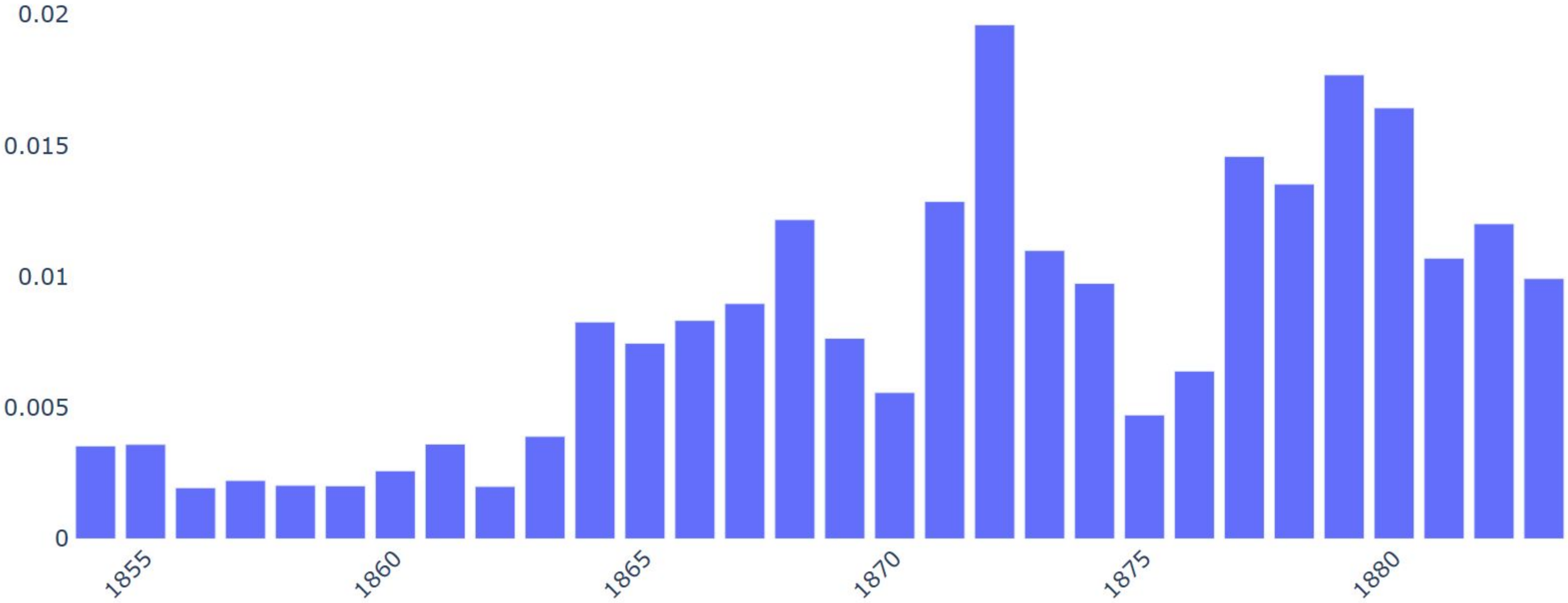


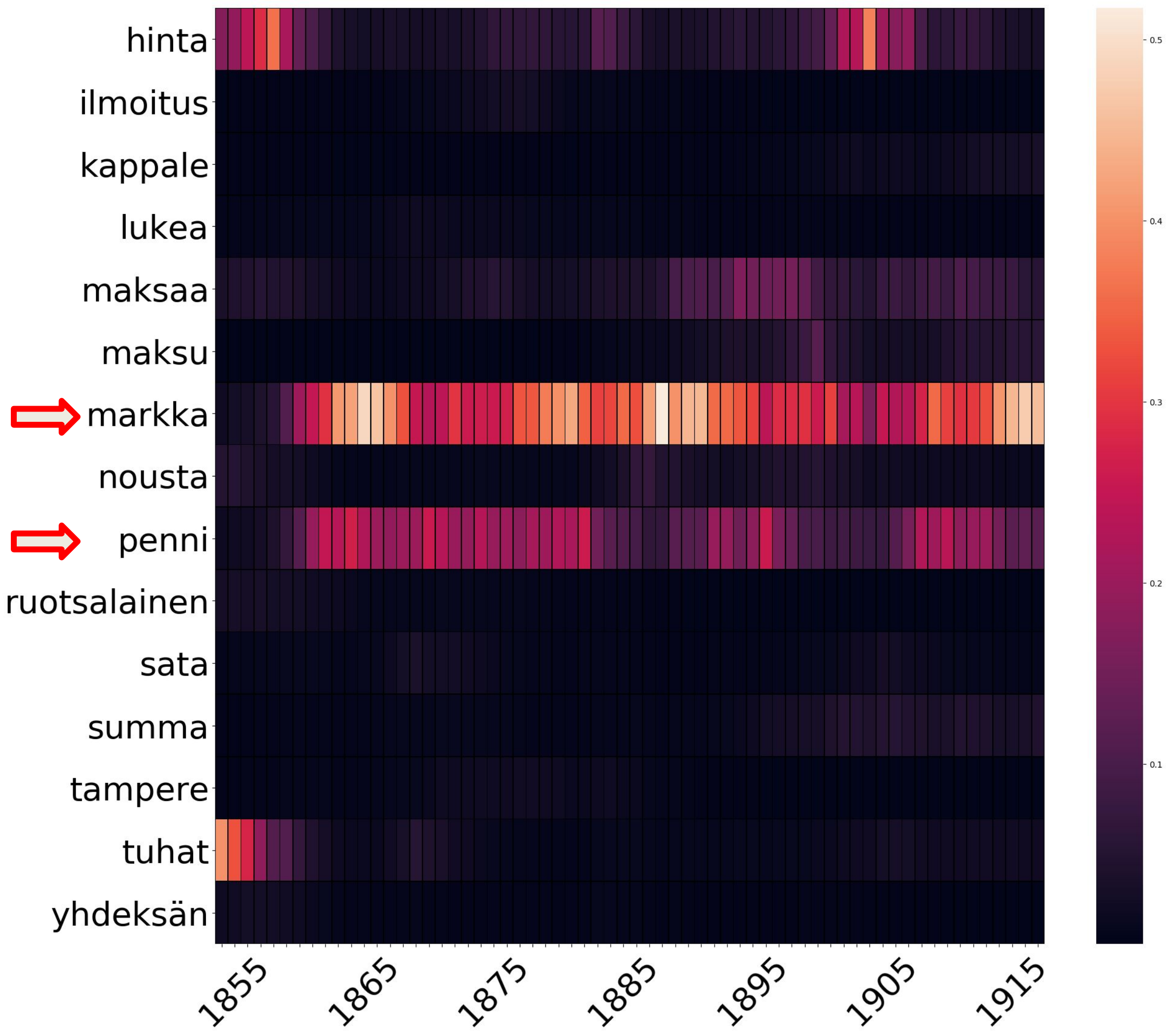
m the
novation
299.



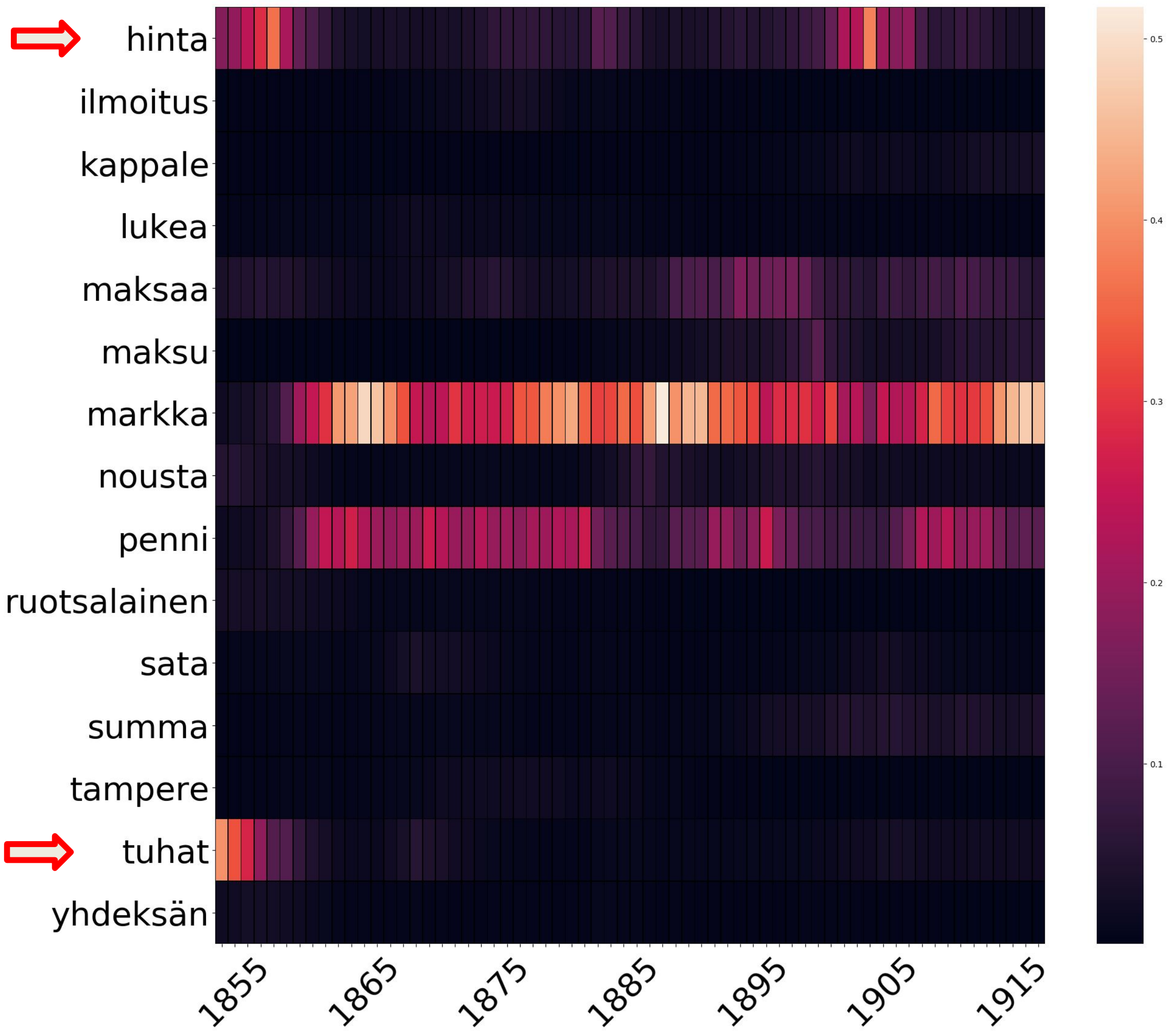
m the
novation
299.

But DTM does stretching (21: introduction of the mark in 1860)




hinta
ilmoitus
kappale
lukea
maksaa
maksu
markka
nousta
penni
ruotsalainen
sata
summa
tampere
tuhat
yhdeksän
1855
1865
1875
1885
1895
1905
1915
0.5
0.4
0.3
0.2
0.1

e
ion

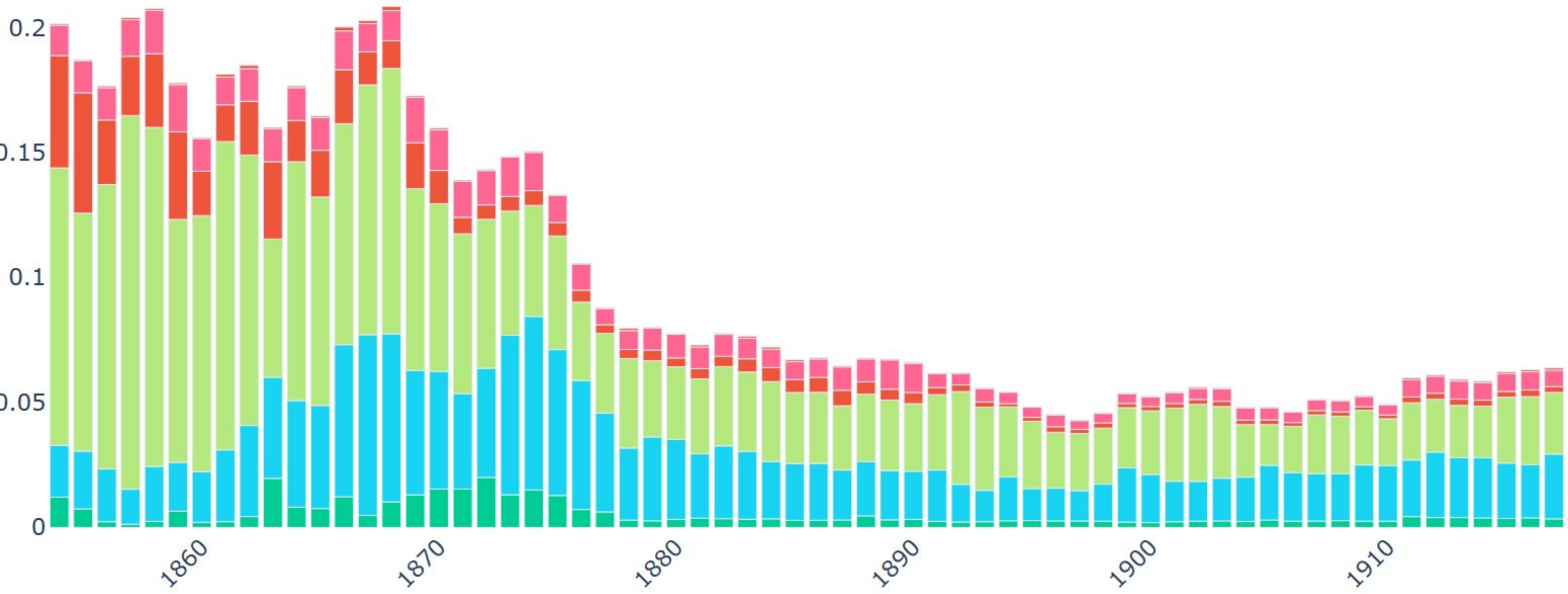

hinta
ilmoitus
kappale
lukea
maksaa
maksu
markka
nousta
penni
ruotsalainen
sata
summa
tampere
tuhat
yhdeksän
1855
1865
1875
1885
1895
1905
1915
0.5
0.4
0.3
0.2
0.1

e
ion

LDA Results

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

Topics in decline: church (3), imperial announcements (6), forests (18), vicar, chaplain (22), municipality, chapel (37), nature and famine (42)



Discussion about methods

- LDA does not allow for temporal change within a topic, where as DTM does
- DTM does "stretching" of topics (more prone to introduce anachronisms), whereas LDA is more sensitive to ruptures in the material
- LDA sometimes requires merging for humanistic interpretation

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Discussion about hypotheses

- We expected a decline in religious topics, and this seems to hold.
- A decline in agrarian topics holds to a certain extent, but we are uncertain about this
- The findings regarding the language struggle were genuine surprises
- We were hoping to catch more fine grained changes, but with our amount of topics, the processes found were very general
- Much easier to identify the things that are new or on the rise!

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Thank you!

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

Part II

Investigating the robustness of embedding-based topic models to OCR Noise

Elaine Zosa, Stephen Mutuvi,
Mark Granroth-Wilding, Antoine Doucet

University of Helsinki
University of La Rochelle

- Topic modelling performance degrades with OCR Noise (Walker et al., 2010; Mutuvi et al., 2018)
- This is due to distorted word co-occurrence statistics which stem from words being misspelled (Walker et al., 2010)
- We think *embedding-based topic models* might be more robust to noise than topic models that do not use embeddings like LDA

What are embedding-based topic models?

Topic models that *directly* incorporate information from word embeddings during training.

- Gaussian LDA (GLDA; Das et al., 2015)
- Latent Concept Topic Model (Hu and Tsuji, 2016)
- Spherical Hierarchical Dirichlet Process (Batmanghelich et al., 2016)
- Embedded Topic Model (ETM; Dieng et al., 2020)
- and many others...

What are embedding-based topic models?

Topic models that *directly* incorporate information from word embeddings during training.

- **Gaussian LDA (GLDA; Das et al., 2015)**
- Latent Concept Topic Model (Hu and Tsuji, 2016)
- Spherical Hierarchical Dirichlet Process (Batmanghelich et al., 2016)
- **Embedded Topic Model (ETM; Dieng et al., 2020)**

- Embedding-based models, *in theory*, can mitigate the negative impact of word misspellings from OCR errors because types with similar identities tend to cluster together in the word embedding space (e.g. *mvrket, markbt, mbrket, market, farket*)
- **Gaussian LDA (GLDA)** and the **Embedded Topic Model (ETM)** have been shown to improve over LDA on clean datasets (20Newsgroups, New York Times corpus)

➔ Question: How robust are these models, compared to LDA, on textual data with OCR noise?

Methodology

- **Data**
  - **Overproof dataset:** OCRed articles from the *Sydney Morning Herald* (1842-1954) digitized by the National Library of Australia
  - Gold standard (GS) articles through crowd-sourced corrections (Evershed & Fitch, 2014)
  - OCR and GS articles are aligned on character level
  - WER: 30%
- **Training**
  - Trained separate topic models on the OCR portion and the GS portion
  - K = 50 topics
  - For ETM and GLDA: experimented with word embeddings trained on Wikipedia and embeddings trained on the Overproof dataset

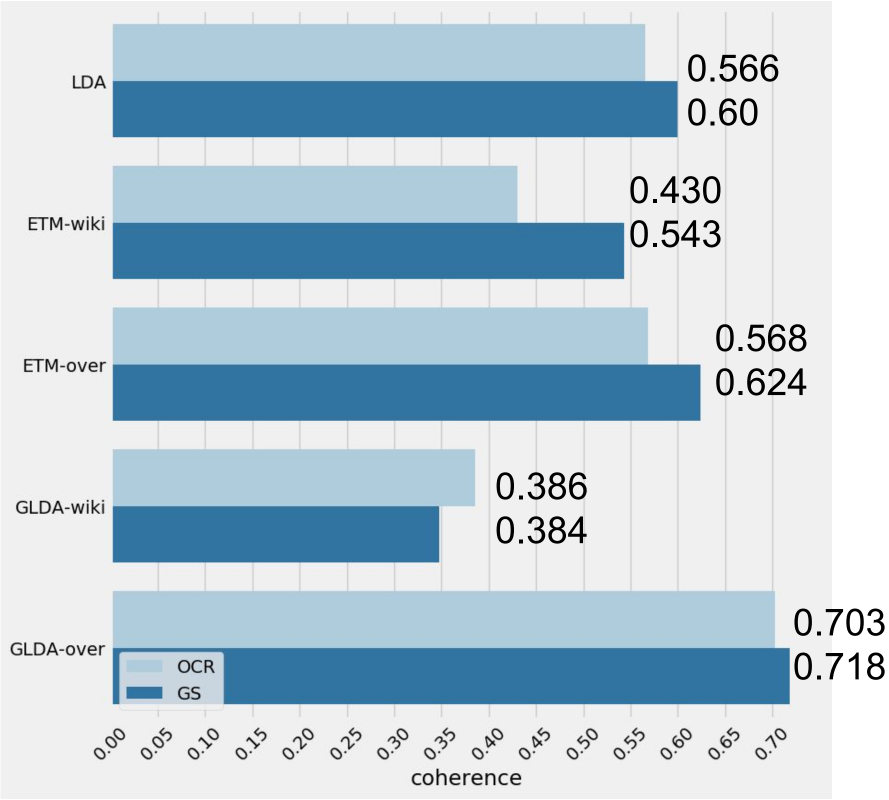
Evaluation measures

- **Mean topic coherence:** *interpretability* of a topic as represented by its most probable words; usually based on normalised pointwise mutual information (nPMI) [Röder, et al., 2015]
- **OCR-GS agreement:** measures the similarity between the topics found in the OCR portion and the topics from the GS portion by the same method. Based on Jaccard Index between topic pairs from two separate trained models (Belford, et al., 2018)
- **Diversity of topics**
- **Model stability**
- **Classification accuracy**

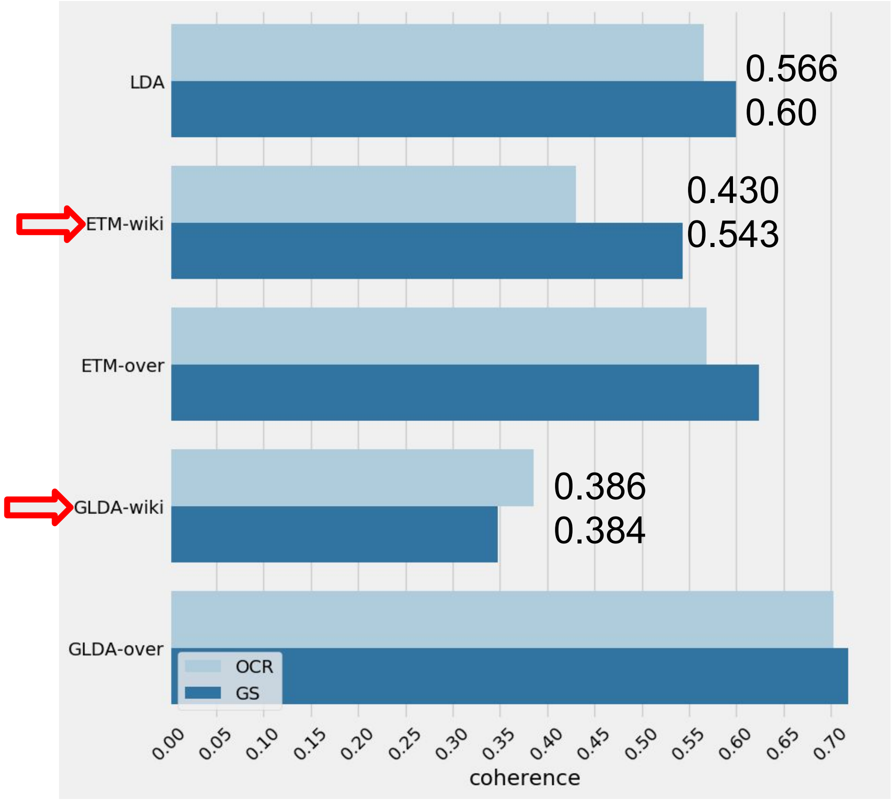
Top words -> Top 20 most probable words

Results

Topic coherence

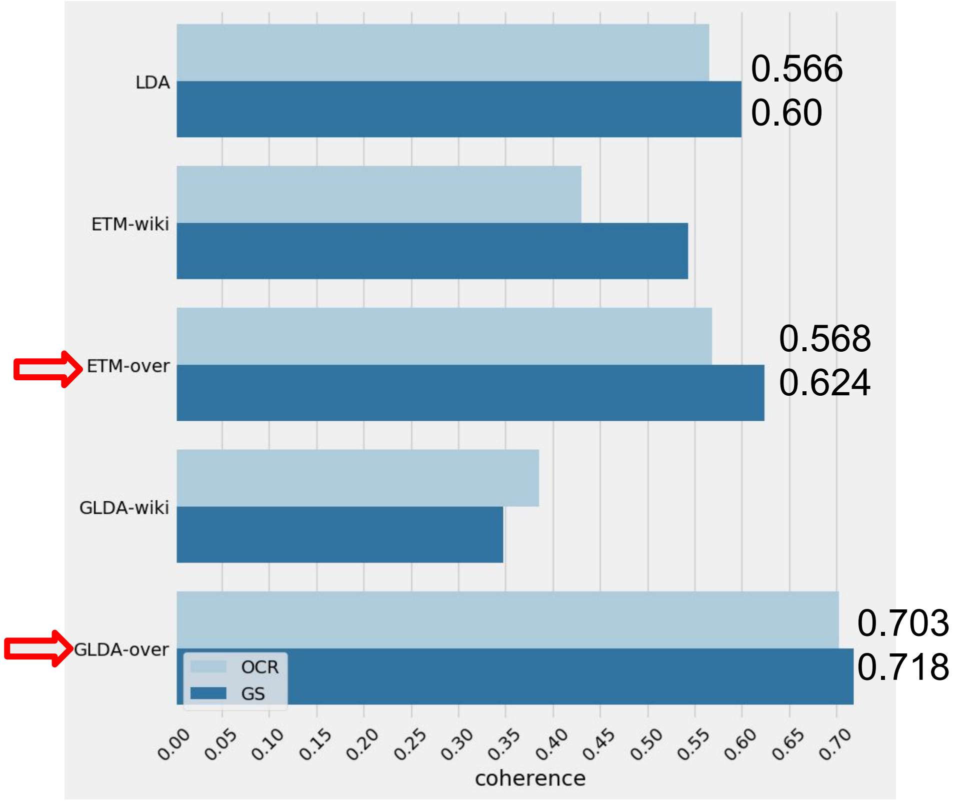


Topic coherence



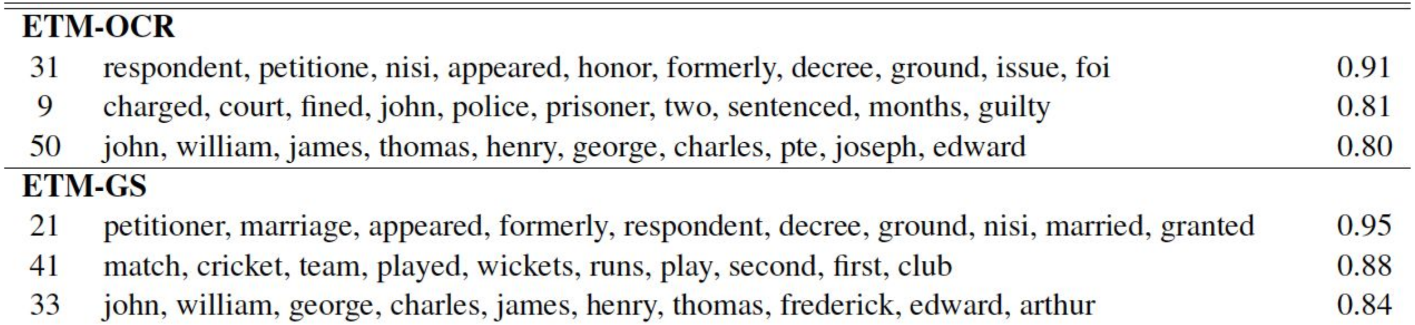
- Using **Wikipedia embeddings** for ETM and GLDA produces **worse** topics than LDA
- The embeddings actually *harm* the models

Topic coherence



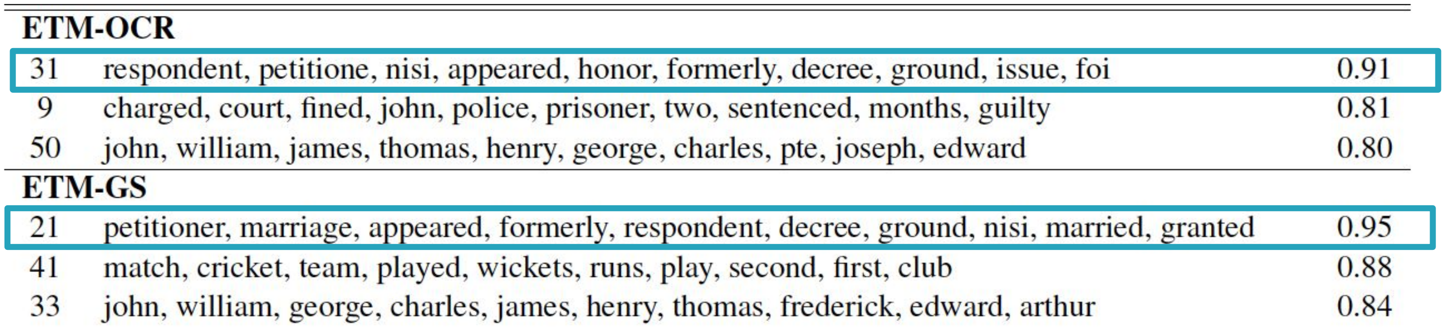
- Using **Wikipedia embeddings** for ETM and GLDA produces **worse** topics than LDA
- The embeddings actually *harm* the models
- Using the **Overproof embeddings**, ETM and GLDA has **more coherent topics** than LDA
- GLDA in particular shows high resilience to noise

ETM Topics



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
50	john, william, james, thomas, henry, george, charles, pte, joseph, edward	0.80
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
33	john, william, george, charles, james, henry, thomas, frederick, edward, arthur	0.84

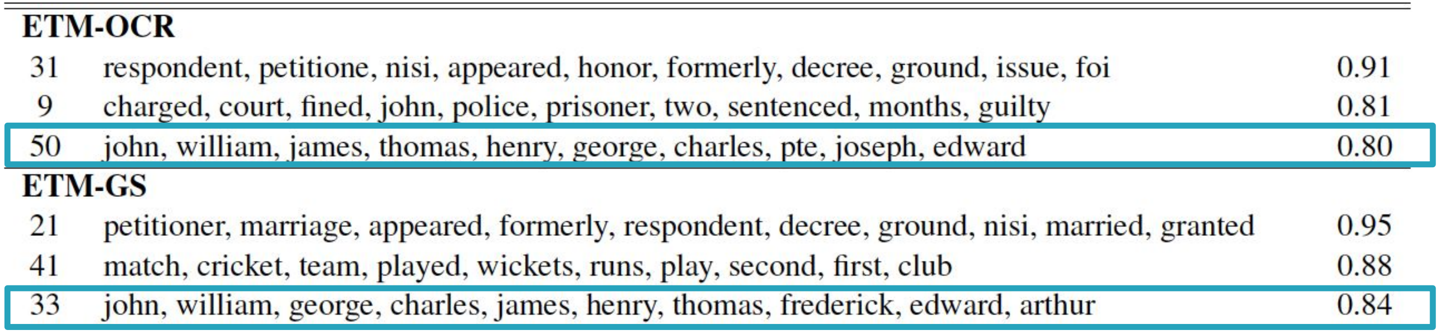
ETM Topics



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
50	john, william, james, thomas, henry, george, charles, pte, joseph, edward	0.80
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
33	john, william, george, charles, james, henry, thomas, frederick, edward, arthur	0.84

ETM-OCR Topic 31 and ETM-GS Topic 21 overlap of **17 out of 20 words**

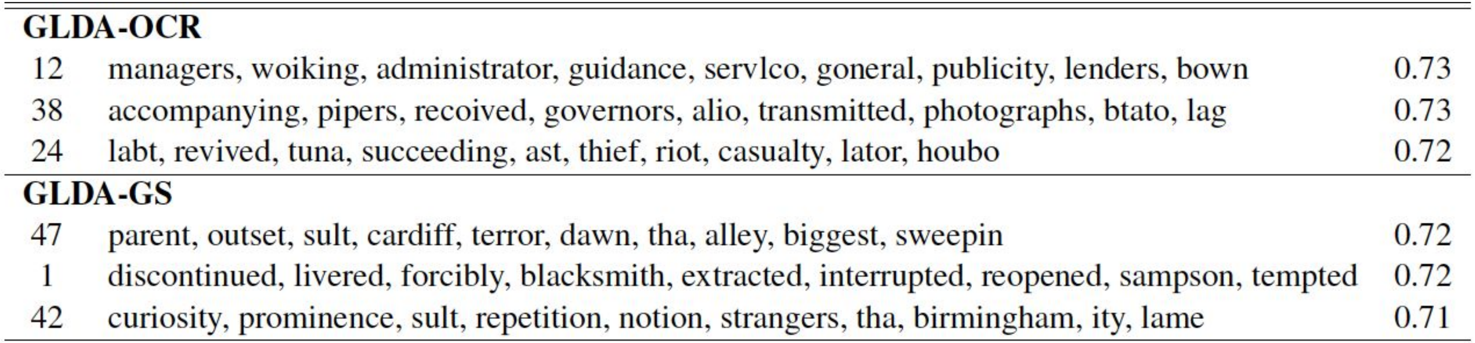
ETM Topics



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
50	john, william, james, thomas, henry, george, charles, pte, joseph, edward	0.80
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
33	john, william, george, charles, james, henry, thomas, frederick, edward, arthur	0.84

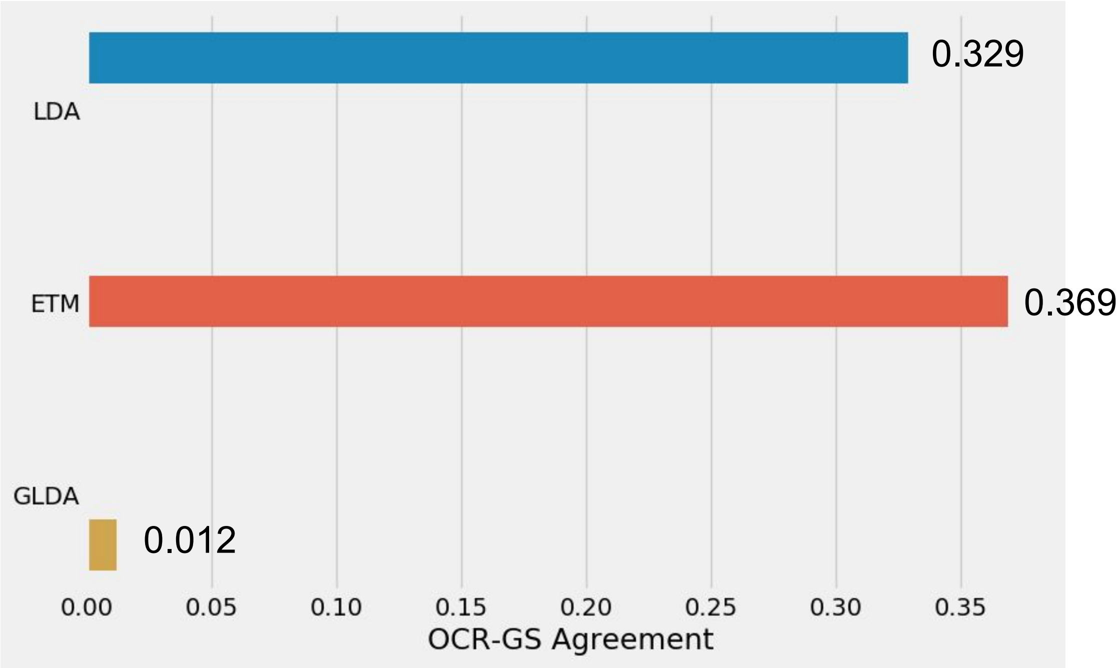
ETM-OCR Topic 50 and ETM-GS Topic 33 overlap of **15 out of 20 words**

GLDA Topics



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
24	labt, revived, tuna, succeeding, ast, thief, riot, casualty, lator, houbo	0.72
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
42	curiosity, prominence, sult, repetition, notion, strangers, tha, birmingham, ity, lame	0.71

No such overlaps found in the top GLDA topics


0.329
LDA
ETM
0.369
GLDA
0.012
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
OCR-GS Agreement

Conclusions

- Embedding-based TMs trained with word embeddings that are trained on the same dataset as the TM produces *more coherent* OCR and GS topics than LDA
- ETM-OCR and ETM-GS topics show a high degree of correspondence (more than LDA and GLDA)
- GLDA-OCR and GLDA-GS topics shows almost no correspondence
- Experiments on synthetic data with increasing levels of noise shows improved resilience from ETM and GLDA over LDA
- Embedding-based TMs offer benefits when working with noisy data but we have to take into account how the word embeddings were trained

Thank you!

References

- Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings.
- Mark Belford, Brian Mac Namee, and Derek Greene. 2018. Stability of topic modeling via matrix factorization.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation.
- Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings.
- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces.
- John Evershed and Kent Fitch. 2014. Correcting noisy ocr: Context beats confusion.
- Stephen Mutuvi, Antoine Doucet, Moses Odeo, and Adam Jatowt. 2018. Evaluating the impact of ocr errors on topic modeling.

References

- Michael R¨oder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures.
- Daniel Walker, William B Lund, and Eric Ringger. 2010. Evaluating models of latent document semantics in the presence of ocr errors.
- X. Wang, A. McCallum, Topics over time: a non-markov continuous-time model of topical trends
- D. M. Blei, J. D. Lafferty, Dynamic topic models