doogan_2021_topic_model_or_topic_twaddle_re_evaluating_semantic_interpretability_measures

Year

2021

Author(s)

Doogan, Caitlin and Buntine, Wray

Title

Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures

Venue

NAACL

Topic labeling

Manual

Focus

Primary

Type of contribution

Established approach

Underlying technique

Manual labeling from four Subject Matter Experts on the basis of the top 10 topic terms

Topic labeling parameters

\

Label generation

The four SMEs were shown the same topics consisting of the top-10 words ranked by term frequency that were generated by LDA and MetaLDA.
Their task was to provide a (single) descriptive label for each and to use 'NA' if they were unable to provide a label.

Motivation

Evaluating two of the four propositions:

- An interpretable topic is one that can be easily labeled.
- An interpretable topic has high agreement on labels.

To verify the robustness of established topic coherence metrics.

Topic modeling

- LDA
- MetaLDA (Zhao et al., 2017)

Topic modeling parameters

We constructed topic sets with the number of topics K = {10, 40, 20, 60, 80, 100, 150, 200}.

Default parameters as appearing in GitHub - ethanhezhao/MetaLDA: The code for MetaLDA in ICDM 2017

Nr. of topics

780 topics (390 topics per model,130 per model-dataset combination)

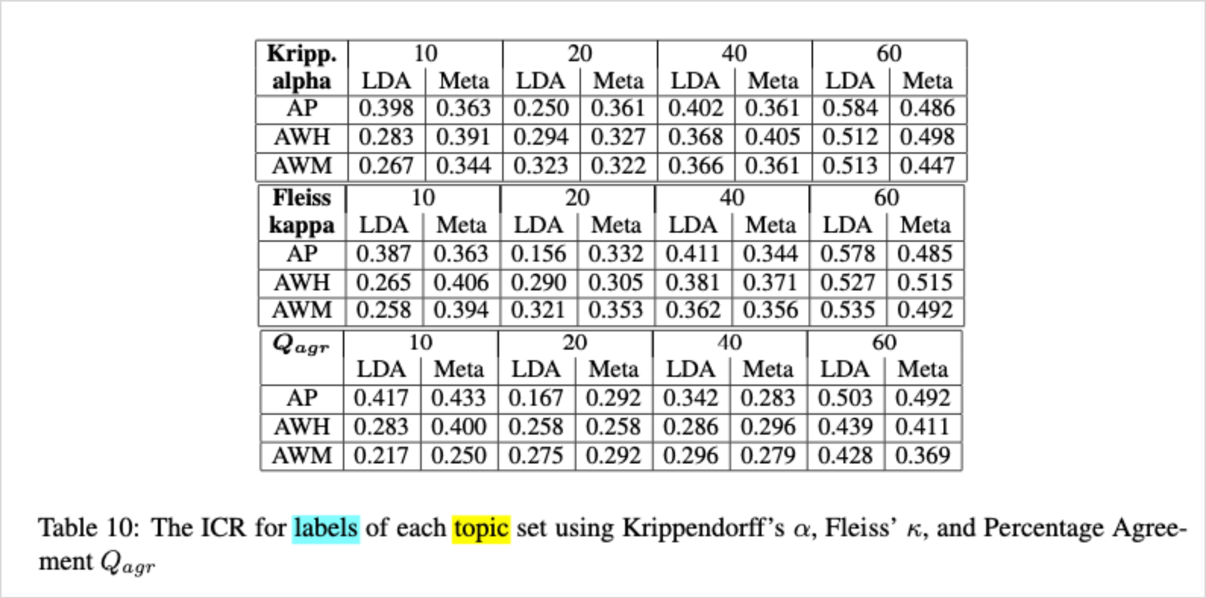
Label

Four (one per assessor) single or multi-world (humanly generated) labels per topic

Label selection

No final selection process. But level of agreement between assessors is measured using:

- Fleiss' kappa (κ) (Fleiss, 1971)
- Krippendorff's alpha (α) (Krippendorff, 2004)



Kripp.	10		20		40		60	
alpha	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.398	0.363	0.250	0.361	0.402	0.361	0.584	0.486
AWH	0.283	0.391	0.294	0.327	0.368	0.405	0.512	0.498
AWM	0.267	0.344	0.323	0.322	0.366	0.361	0.513	0.447

Fleiss	10		20		40		60	
kappa	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.387	0.363	0.156	0.332	0.411	0.344	0.578	0.485
AWH	0.265	0.406	0.290	0.305	0.381	0.371	0.527	0.515
AWM	0.258	0.394	0.321	0.353	0.362	0.356	0.535	0.492

Q_{agr}	10		20		40		60	
	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.417	0.433	0.167	0.292	0.342	0.283	0.503	0.492
AWH	0.283	0.400	0.258	0.258	0.286	0.296	0.439	0.411
AWM	0.217	0.250	0.275	0.292	0.296	0.279	0.428	0.369

Table 10: The ICR for labels of each topic set using Krippendorff's α, Fleiss' κ, and Percentage Agreement Q_{agr}

Label quality evaluation

/

Assessors

Four Subject Matter Experts from a multidisciplinary pool of researchers who were representative of the political-ideological spectrum and who were Australian English speakers.

Domain

Paper: Topic modeling evaluation
Dataset: Politics

Problem statement

Evaluating widely used topic coherence measures in an applied settings by considering four propositions:

- If coherence scores are robust, they should correlate.
- An interpretable topic is one that can be easily labeled.
- An interpretable topic has high agreement on labels.
- An interpretable topic is one where the document-collection is easily labeled.

Corpus

Origin: Twitter
Nr. of documents: 123,629
Details: Three subsets of Auspol-18 dataset on Australian politics

Document

Textual content of a single tweet.
Single tweet made up of up to 140 characters.

Pre-processing

- stopword removal
- POS-tagging
- Lemmatisation
- Exclusion of non-English tweets
- Duplicate removal
- Removal of tokens with a frequency n < 10
- Removal of tweets with n < 5 tokens
- Standardization of slang, abbreviations

```
@inproceedings{doogan_2021_topic_model_or_topic_twaddle_re_evaluating_semantic_
interpretability_measures,
    title = "Topic Model or Topic Twaddle? Re-evaluating Semantic
Interpretability Measures",
    author = "Doogan, Caitlin  and
      Buntine, Wray",
    booktitle = "Proceedings of the 2021 Conference of the North American
Chapter of the Association for Computational Linguistics: Human Language
Technologies",
    month = jun,
    year = "2021",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2021.naacl-main.300",
    doi = "10.18653/v1/2021.naacl-main.300",
    pages = "3824--3848",
    abstract = "When developing topic models, a critical question that should
be asked is: How well will this model work in an applied setting? Because
standard performance evaluation of topic interpretability uses automated
measures modeled on human evaluation tests that are dissimilar to applied
usage, these models{'} generalizability remains in question. In this paper, we
probe the issue of validity in topic model evaluation and assess how
informative coherence measures are for specialized collections used in an
applied setting. Informed by the literature, we propose four understandings of
interpretability. We evaluate these using a novel experimental framework
reflective of varied applied settings, including human evaluations using open
labeling, typical of applied research. These evaluations show that for some
specialized collections, standard coherence measures may not inform the most
appropriate topic model or the optimal number of topics, and current
interpretability performance validation methods are challenged as a means to
confirm model quality in the absence of ground truth data.",
}

@INPROCEEDINGS{zhao_2017_metalda_a_topic_model_that_efficiently_incorporates_me
ta_information,
  author={Zhao, He and Du, Lan and Buntine, Wray and Liu, Gang},
  booktitle={2017 IEEE International Conference on Data Mining (ICDM)},
  title={MetaLDA: A Topic Model that Efficiently Incorporates Meta
Information},
  year={2017},
  volume={},
  number={},
  pages={635-644},
  doi={10.1109/ICDM.2017.73}}

@article{
fleiss_1971_measuring_nominal_scale_agreement_among_many_raters,
title={Measuring nominal scale agreement among many raters.},
author={Joseph L. Fleiss},
journal={Psychological Bulletin},
year={1971},
volume={76},
pages={378-382}
}
```

```
@article{krippendorff_2004_measuring_the_reliability_of_qualitative_text_analysis_data,
  abstract = {This paper reports a new tool for assessing thereliability of text interpretationsheretofore unavailable to qualitative research.It responds to a combination of twochallenges, the problem of assessing the reliabilityof multiple interpretations --asolution to this problem was anticipated earlier(Krippendorff, 1992) but not fullydeveloped --and the problem of identifying unitsof analysis within a continuum of textand similar representations (Krippendorff, 1995).The paper sketches the family ofα-coefficients, which this paper extends, andthen describes its new arrival.A computational example is included in the Appendix.},
  author = {krippendorff, Klaus},
  date = {2004/12/01},
  date-added = {2023-02-28 14:45:12 +0100},
  date-modified = {2023-02-28 14:45:12 +0100},
  doi = {10.1007/s11135-004-8107-7},
  id = {krippendorff2004},
  isbn = {1573-7845},
  journal = {Quality and Quantity},
  number = {6},
  pages = {787--800},
  title = {Measuring the Reliability of Qualitative Text Analysis Data},
  url = {https://doi.org/10.1007/s11135-004-8107-7},
  volume = {38},
  year = {2004},
  bdsk-url-1 = {https://doi.org/10.1007/s11135-004-8107-7}}
```

#Thesis/Papers/Initial