Using Latent Dirichlet Allocation on Twitter data to gain insight in the behaviour and mindsets of young people in restrictive settings.

Sije van der Veen

rnw media

UNIVERSITY OF AMSTERDAM

Problem

- RNW writes about important and sensitive topics which young people in restrictive settings care about
- Costly supervised topic modeling for RNW media makers.
- Unseen topics by media makers should also be provided by content on the different platforms.

Approach

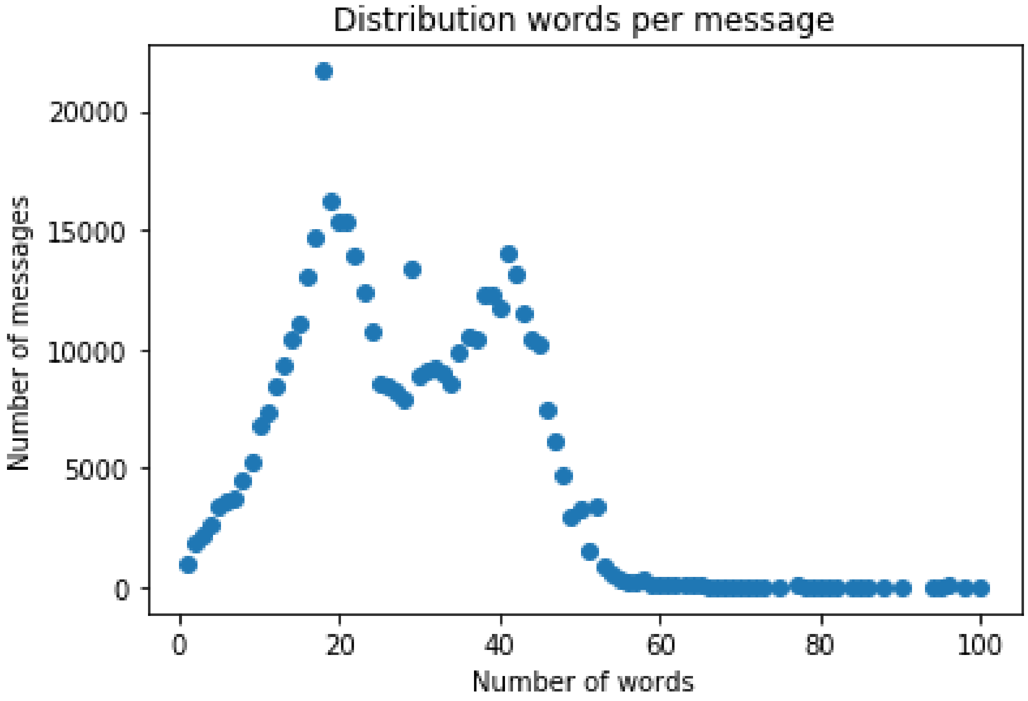
- Applying topic modeling on Twitter messages of the RNW Burundi Twitter platform.
- The assumption is that the topics will largely categorize and summarize what people of this platform talk about.

Subquestions

- How accurately can topic modelling be applied on Twitter data?
- What is the optimal number of topics that can be generated?
- Can LDA distinguish between Twitter message using hash-tags as true labels?

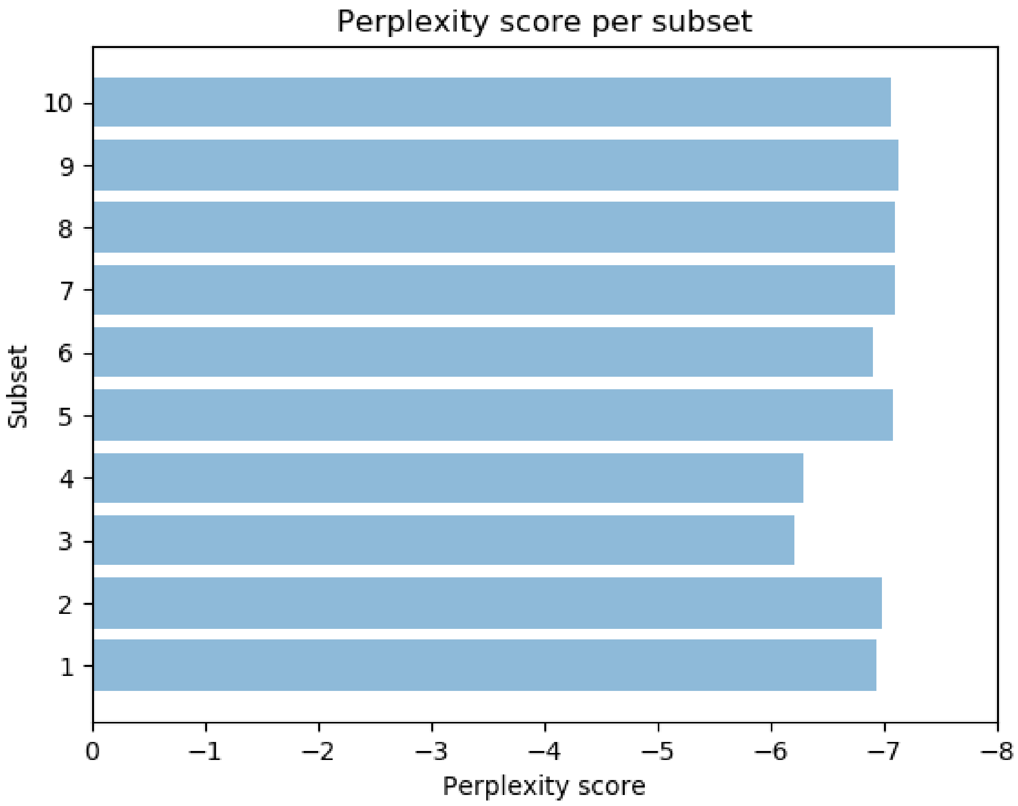
Dataset

- ~ 460.000 Twitter message and hash-tags
- RNW Burundi Twitter platform
- French, English and Rundi



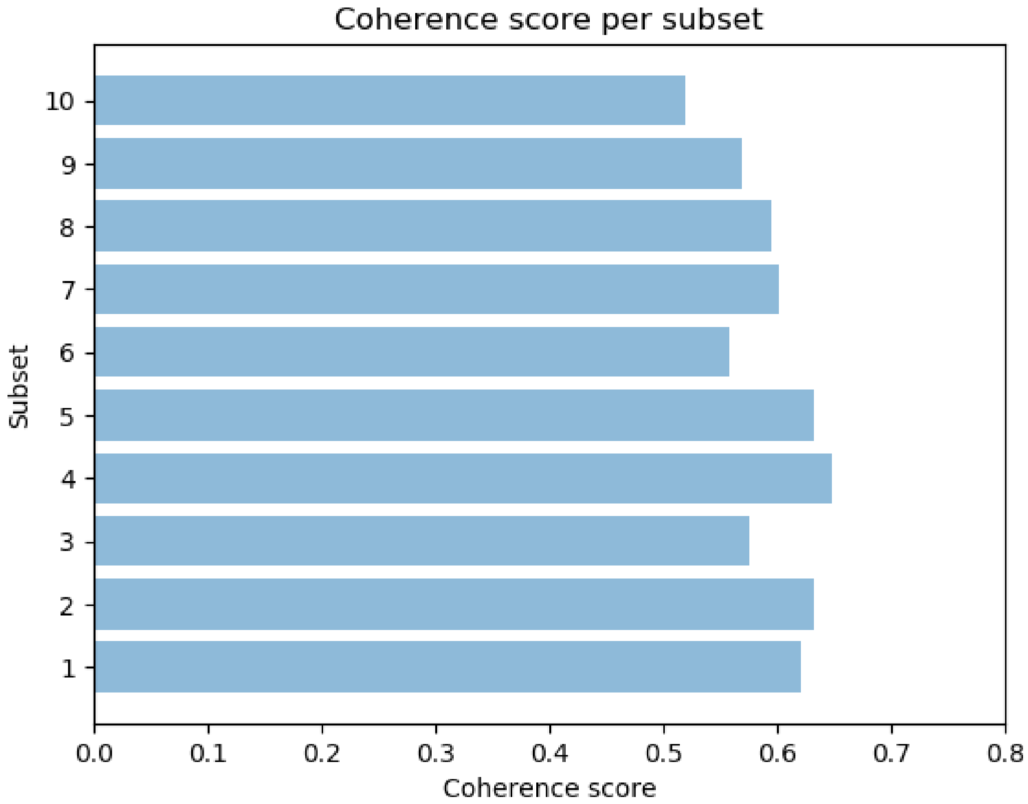
Subquestion 1

- Perplexity tested on subset of dataset
- The perplexity is performing stable between the subsets.
- Standard deviation of 2.87

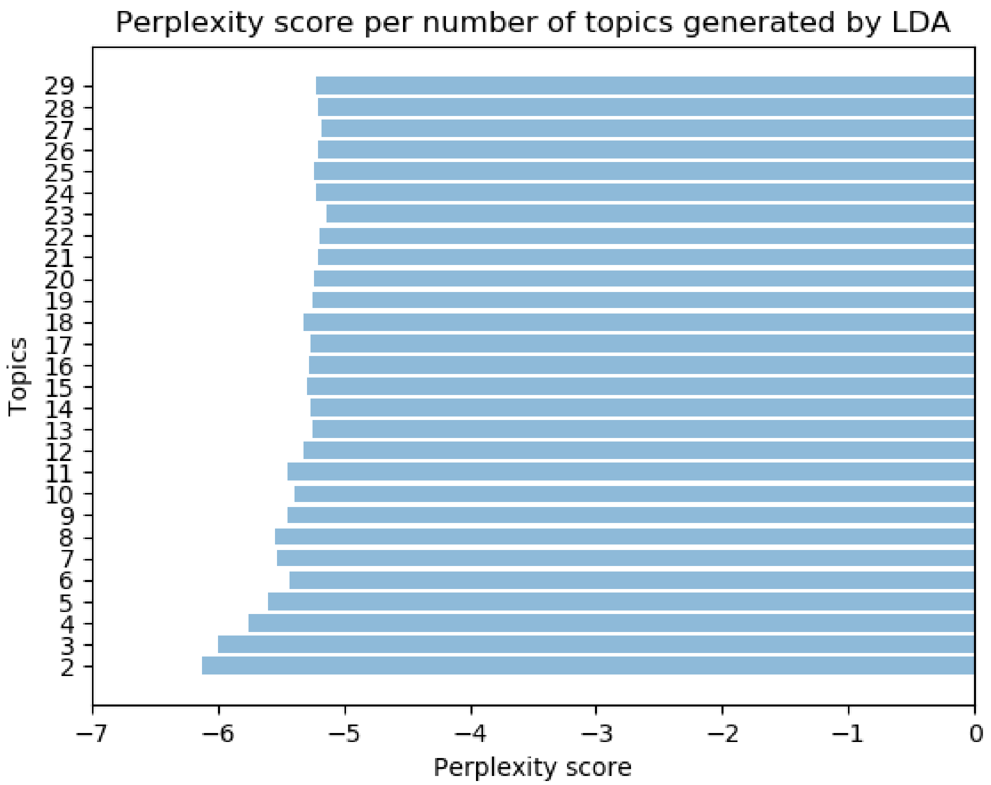


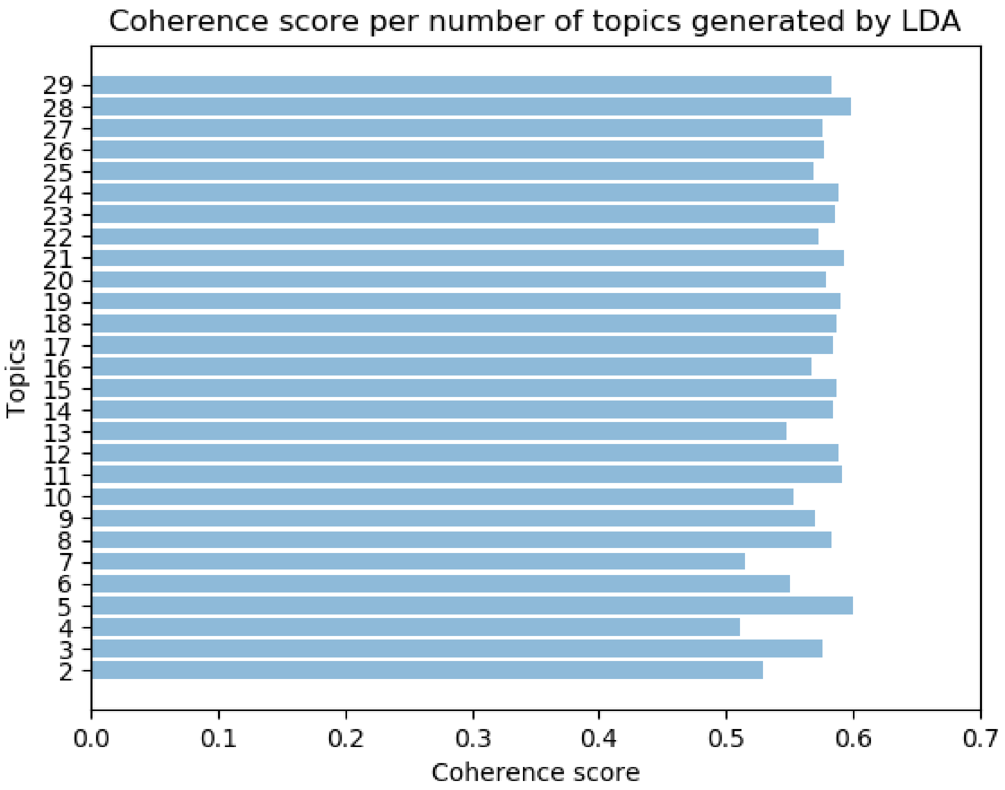
Subquestion 1

- Coherence tested on subset of dataset
- The perplexity is performing stable between the subsets.
- Standard deviation of 0.038



Subquestion 2





Perplexity optimum = 2

Coherence optimum = 5

Subquestion 3

- Can LDA distinguish between Twitter message using hash-tags as true labels?
- Are the message with in common hash-tag assigned to the same topic group?
- Important hash-tags will be selected.
- Accuracy will be used for evaluation.
- A LDA topic will be assigned to a hash-tag based on the biggest number labeled hash-tags within LDA topic.