UNIVERSITY OF
GOTHENBURG

Investigating the Effects of MWE Identification in Structural Topic Modelling

Dimitrios Kokkinakis
Ricardo Muñoz Sánchez
Sebastianus C. J. Bruinsma
Mia-Marie Hammarlin

Background & Motivation

- Structural topic modelling
  - Topic modelling using LDA
  - Model how the the topics change according to a covariate variable (e.g. time, location, etc.)

- We want to check whether adding MWEs improves the explainability and interpretability of the topic modelling

- Our approach: a quantitative phase followed by a qualitative one

Case Study

- We are studying vaccine skepticism
- What do people talk about and how do they talk about it?
- As a case study we use social media reactions to a study made by the University of Lund

Dataset: Starting Point

- Use the University of Lund paper as a base
  - "*Intracellular Reverse Transcription of Pfizer BioNTech COVID-19 mRNA Vaccine BNT162b2 In Vitro in Human Liver Cell Line*" by Aldén et al.

- Often cited to justify vaccine skepticism and hesitancy
  - There is a potential misconception that the mRNA vaccine alters the human DNA

Dataset: Social Media

- Swedish Tweets from February 2022 to November 2022
  - 1,870 Tweets from 858 different users

- Posts from the Swedish forum Flashback from the same time period
  - 8,900 unique posts

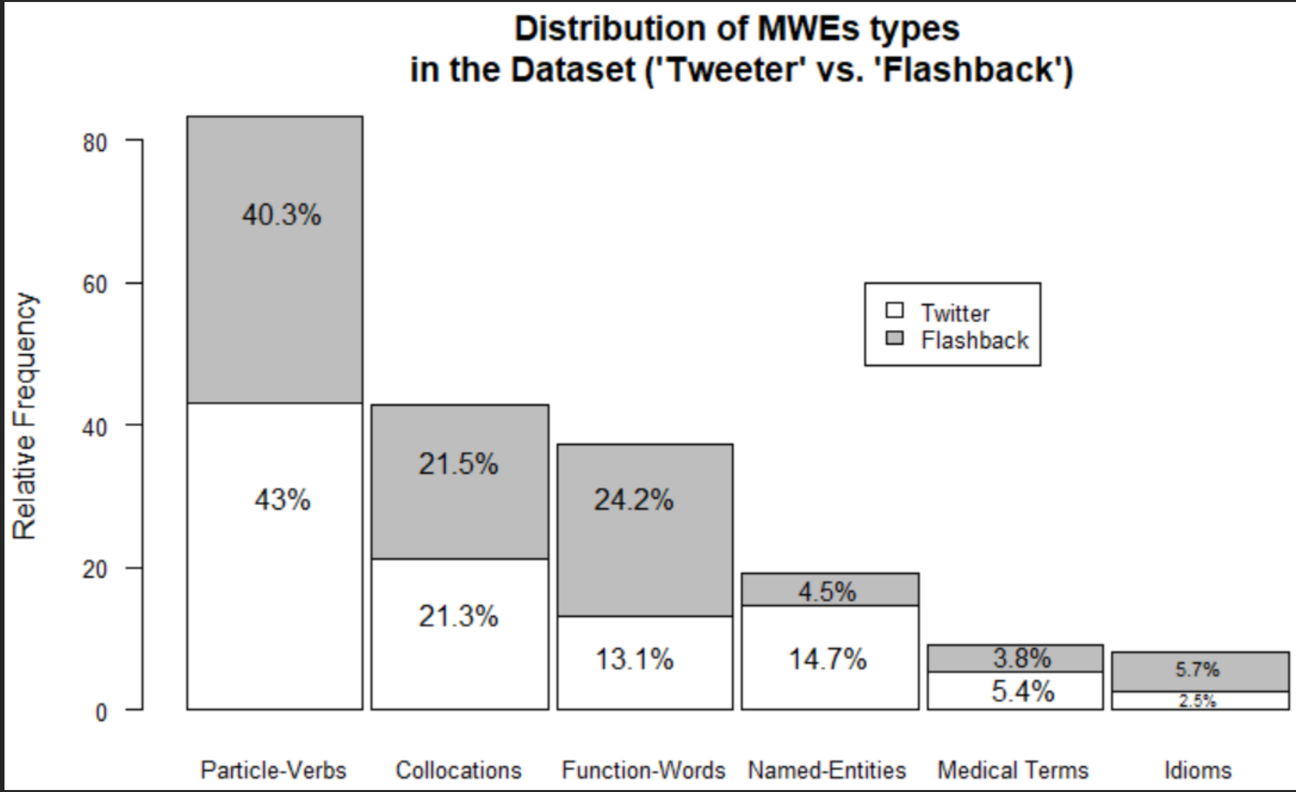
- We created two versions of each one with and one without MWEs

Swedish MWEs

- Lists of lexicalized idioms
  - bli blåst (to be cheated)
- Phrasal verbs
  - ställa upp (to participate)
- Function words
  - på grund av (because of)

Swedish MWEs

- Named entities
  – Lunds universitet (Lund University)
- Medical terminology
  – akut myokardit (acute myocarditis)
- N-gram collocations
  – smittsamt virus (infectious virus)


Distribution of MWEs types
in the Dataset ('Tweeter' vs. 'Flashback')
Relative Frequency
0
20
40
60
80
40.3%
43%
21.5%
21.3%
24.2%
13.1%
4.5%
14.7%
3.8%
5.4%
5.7%
2.5%
Twitter
Flashback
Particle-Verbs
Collocations
Function-Words
Named-Entities
Medical Terms
Idioms

Data Preprocessing

- MWE tokens were concatenated by underscore to a single token for uniformity
  - Robert Malone -> ‘Robert_Malone’

- Normalization
  - Lowercase
  - Punctuation & stopword removal
  - Removal of the top-10 most frequent words
  - Lexical normalization

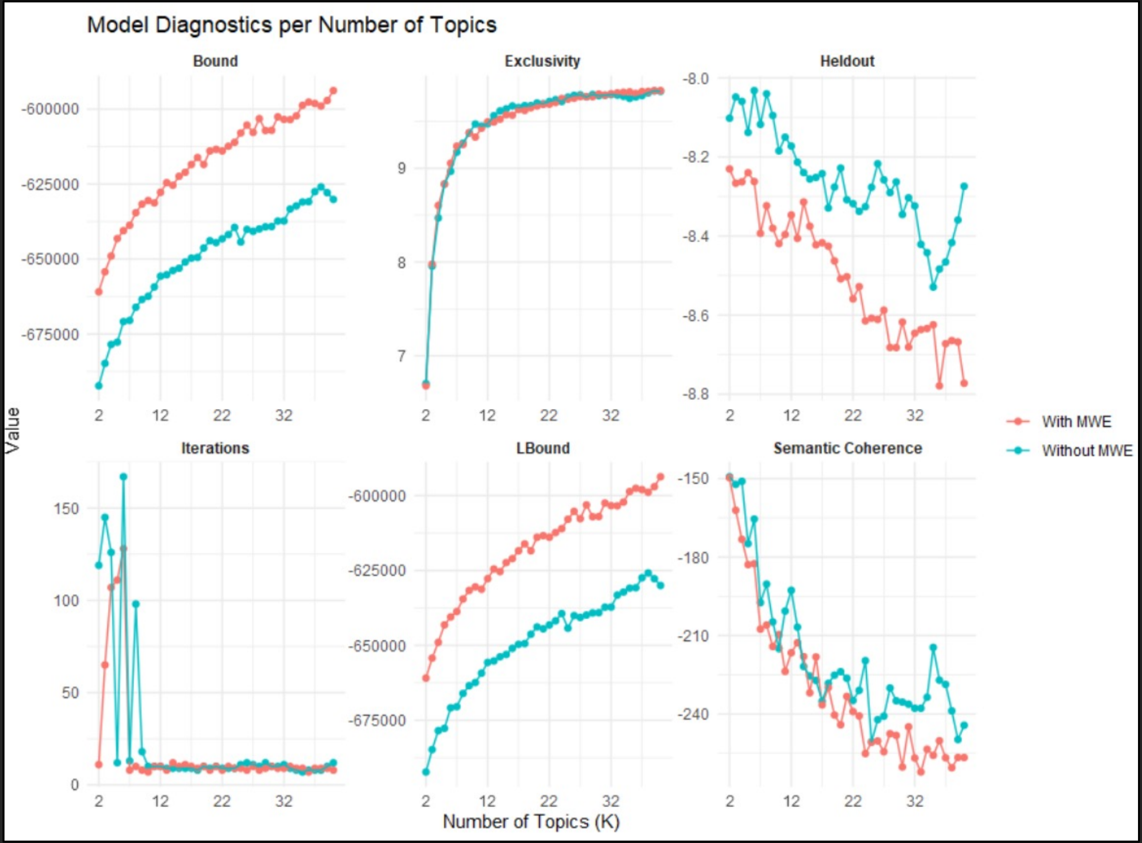
Structural Topic Modelling (STM)

- An extension to standard Latent Dirichlet Analysis (LDA)
- Allows for the integration of covariates (e.g.)
- We use it to model how our document collection changes over time to see how topics evolve

Selecting the Number of Topics

We want to look at two main things for this

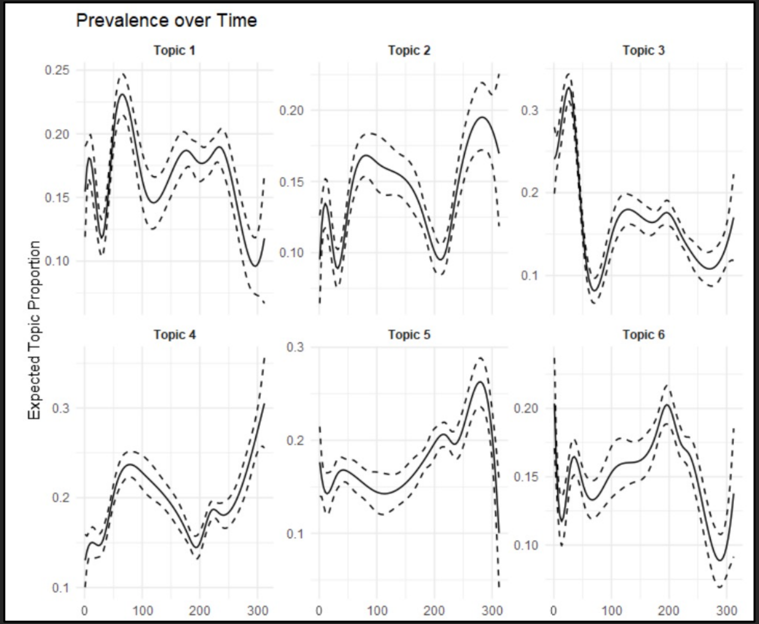
- Semantic coherence
  - The topics should be semantically interpretable
  - Low scored topics are usually artifacts of statistical inference
- Exclusivity
  - The top words in each topic should not appear as top words on other topics


Model Diagnostics per Number of Topics
Bound
Exclusivity
Heldout
Iterations
LBound
Semantic Coherence
Value
Number of Topics (K)
With MWE
Without MWE

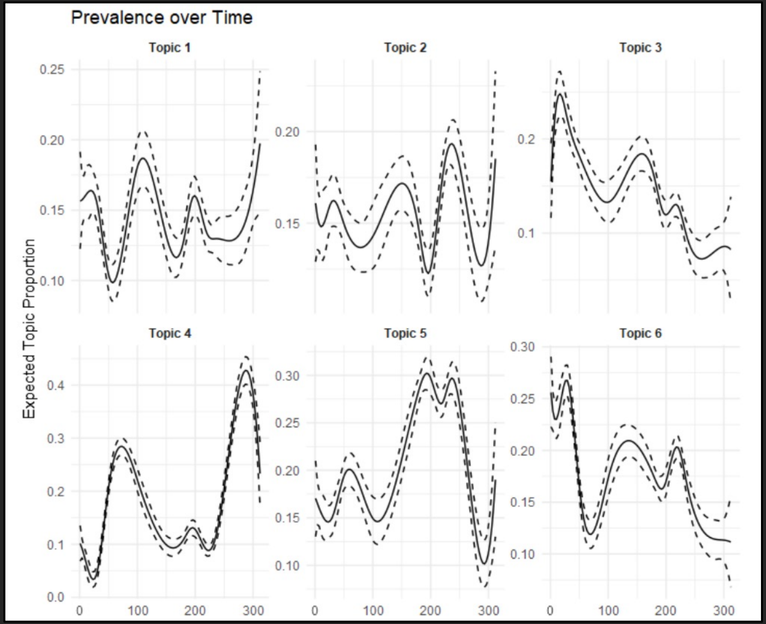
Manually assigned topic labels:

1. tone/sentiment
2. women’s health issues
3. freedom issues
4. truth seeking
5. power issues
6. body issues/side effects

Prevalence Over Time – No MWE



Prevalence Over Time – With MWEs



Results and Discussion

- *Hypothesis:* MWE identification can provide better, more targeted insights and enhance the interpretability and explainability of the generated topics

- *Evaluation*
  - keywords with the highest association for each topic
  - qualitative reading of the 50 most-representative tweets and/or posts for each topic

Results and Discussion

- Quantitative
  - There is only a slight improvement on the semantic coherence and the exclusivity when using MWEs
  - This might also be due to the increase of vocabulary size when using MWEs

- Qualitative
  - The different topics became clearer and easier to understand when looking at their top words

Limitations

- The dataset size is small
- The Twitter and Flashback searches were limited to a small list of keywords
- Lack of lemmatization could have affected the results
- All MWEs explored were contiguous

Future Directions

- Comparing different sources/lists of MWEs
  - Also explore other kinds of MWEs (e.g. non-contiguous MWEs

- Datasets
  - Use larger datasets
  - Check whether other reactions or discussions about other papers or arguments behave differently
  - Expand the queries used to find the datapoints

UNIVERSITY OF
GOTHENBURG

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io

This work is part of the project Rumour Mining (MXM19-1161:1) financed by the Bank of Sweden Tercentenary Foundation – Riksbankens Jubileumsfond

All images in this presentation are royal free, downloaded from https://pixabay.com/