UNIVERSITY OF BIRMINGHAM

School of Computer Science
College of Engineering and Physical Sciences

MSc. Project

Analysing news articles about Russia's war on Ukraine using Latent Dirichlet Allocation based topic modelling

---

Submitted in conformity with the requirements
for the degree of MSc. Artificial Intelligence and Machine Learning
School of Computer Science
University of Birmingham

Nelson Quintanilla Castro
Student ID: 2291960
Supervisor: Dr Mohammed Bahja

September 2022

MSc. Project
Analysing news articles about Russia's war on Ukraine using Latent Dirichlet Allocation based topic modelling

Nelson Quintanilla Castro

Contents

Table of Abbreviations

- IEEE – Institute of Electrical and Electronics Engineers
- IEEE – Institute of Electrical and Electronics Engineers
- IEEE – Institute of Electrical and Electronics Engineers

List of Figures

1 Overview

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.1 Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.2 Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

2 Introduction

The ongoing conflict between Russia and Ukraine is a topic of pivotal importance not only for the protagonist's regions but also for the whole world. This war is already causing many negative economic, political, and social consequences on a different scale for many countries Caldara et al. (2022). As a result, this has become of paramount concern for governments and, subsequently for all types of media.

This is why there is currently a tremendous amount of information that has been generated about the Russo-Ukrainian conflict and there is no systematic way to analyse this corpora to examine what is currently happening regarding this conflict, to discover the main themes contained within it, and possibly draw conclusions about it.

At the same time, it is not humanly possible to read and study large quantities of articles with this thematic in a short period of time. To do this, experts in machine learning have created probabilistic topic modelling, a collection of several algorithms which are used to identify and interpret sizeable sets of documents that contain topical information. We can use topic modelling to identify the primary ideas that run through a sizable and otherwise unstructured collection of papers. We can arrange the collection in accordance with the themes found by applying topic modelling Blei (2012).

For the purpose of this work, latent dirichlet allocation (LDA) is going to be the topic modelling algorithm of choice. A collection of texts can be broken down into their key topics using LDA, where a topic is defined as a probability distribution over a vocabulary Blei (n.d.).

A set number of topics are proposed by LDA, and it is assumed that each document in a collection of documents reflects a combination of those themes. Under these presumptions, probabilistic inference techniques identify an embedded theme structure in a document collection. LDA offers a technique to easily summarise, peruse, and search huge document collections with this structure Blei et al. (2010).

The main motivation of this work is to capture the main themes or topics that can be found on news articles available on the internet that cover events related to Rusia's war on Ukraine. This represents an important topic to study because there is a need to derive insightful information from this vast amount of news about the war that are being generated in mass all over the world.

Furthermore, it is of utter importance to observe and detect significant patterns that are being communicated to the general public through news articles so we can predict consequences of the current state of the situation on the short term such as the economical, political, and social impact on human lives.

In this paper we provide an analysis of news articles retrieved from The Guardian newspaper that cover events related to the Russian invasion of Ukraine by utilising LDA based topic modelling with the purpose of obtaining abstract topics within a corpora of news articles. We also offer a way to assess the effectiveness and outcomes of our model. Additionally, we present the validation of our model by comparing it against human judgement. In order to allow others to duplicate our findings, we have also made our data sets publicly available. Further to this, we use pyLDAvis to interpret the topics obtained by our model in a more comprehensible and visual way. This software offers a global view of the topics and how they differ from one another while also enabling a close examination of the terms most closely associated with each specific topic Sievert & Shirley (2014).

The structure of this paper is as follows. In section 3 we examine the most recent studies in this field. In Section 4 we present the specifics of our approaches, including a description of data extraction, preprocessing, parameter tuning of our model, and evaluation metrics implemented. We provide our findings and engage in a discussion in Section 5. The discussion is presented in part 6, and our conclusion is presented in section 7.

3 Literature Review

3.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3.2 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4 Methodology

In this section we explain the steps we took in order to build and prepare our dataset, the model generation and parameter tuning process, methods of evaluation for our model, and lastly visualisation and interpretation of our resulting topics. The Figure 1 summarises the method we just mentioned.

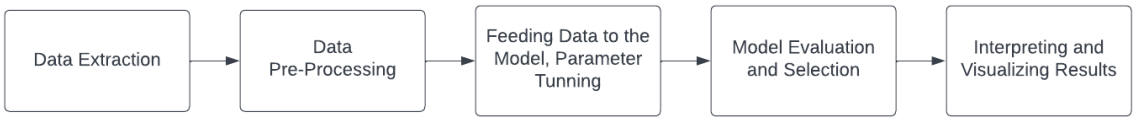

Figure 1: *Diagram that summarises steps taken during the development of our work.*

4.1 Data Extraction

For the Data

4.2 Data Pre-Processing

4.2.1 Dataset Relevant Features

4.2.2 Normalisation and Tokenisation

4.2.3 Stop Word Removal

4.2.4 Lemmatisation

4.2.5 Transforming the documents in a vectorised form (Bag of Words Representation)

4.2.6 Pre-Processed Data information

4.3 Model Training and Parameter Tunning

4.3.1 Training Multiple Models for different values of 'Number of Topics' Parameter

4.3.2 Choice of Stop Word List

4.3.3 Addition of Custom Words to the Stop Word List

4.3.4 Change of function to train the model

4.4 Model Evaluation and Selection

4.4.1 Visualising topic models

4.4.2 Using trained models on sample of documents

4.4.3 Model Perplexity and Coherence

4.4.4 Coherence

4.4.5 Perplexity

4.4.6 Human Judgement

4.4.7 Word intrusion

4.4.8 Topic intrusion

4.5 Topics Interpretation and Visualisation

5 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.1 Dataset

5.2 LDA model parameters

5.3 Experimental comparison between coherence and perplexity

5.3.1 Coherence

5.3.2 Dataset

5.4 Experimental comparison using human judgement

6 Discussion

6.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6.1.1 Subsubsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7 Conclusion

7.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7.2 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7.2.1 Subsubsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

Blei, D., Carin, L. & Dunson, D. (2010), 'Probabilistic Topic Models', *IEEE Signal Processing Magazine* **27**(6), 55–65. Conference Name: IEEE Signal Processing Magazine.

Blei, D. M. (2012), 'Probabilistic topic models', *Communications of the ACM* **55**(4), 77–84.
URL: *https://dl.acm.org/doi/10.1145/2133806.2133826*

Blei, D. M. (n.d.), 'Latent Dirichlet Allocation', p. 30.

Caldara, D., Conlisk, S., Iacoviello, M. & Penn, M. (2022), 'The Effect of the War in Ukraine on Global Activity and Inflation'.
URL: *https://www.federalreserve.gov/econres/notes/feds-notes/the-effect-of-the-war-in-ukraine-on-global-activity-and-inflation-20220527.html*

Sievert, C. & Shirley, K. (2014), LDAvis: A method for visualizing and interpreting topics, *in* 'Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces', Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 63–70.
URL: *http://aclweb.org/anthology/W14-3110*

8 Appendix One: Accompanying Archive and Instructions

8.1 Directory Structure

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

8.2 Running the Provided Code

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.