Technical Annex on the Natural Language Processing section of the article on "The Word and the Bullet: Out-Grouping and Threat Framing as Predictors of Terrorist Targeting by ISIS, 2015-2019"

Gillian Kant

University of Göttingen, Göttingen, Germany.

Contributing authors: gillian.kant@stud.uni-goettingen.de;

1 Annex

This annex aims to provide additional technical background information regarding the specific preprocessing techniques and analysis carried out in the paper "The Word and the Bullet: Out-Grouping and Threat Framing as Predictors of Terrorist Targeting by ISIS, 2015-2019". Despite the advancements in Natural Language Processing (NLP) and Topic Modeling techniques, analyzing Arabic text data still remains a niche area within the social sciences. Although large language models trained on billions of documents have made it easier to work with Arabic text data, Topic Models such as Latent Dirichlet Allocation (LDA) require specific preprocessing of the data to produce satisfactory results. However, one of the major challenges faced when working with Arabic texts is the lack of proper data preprocessing frameworks. For instance, the popular NLP-Software library SpaCy, which provides a range of powerful and helpful functions for preprocessing, does not support Arabic. As a result, several self-created solutions for the preprocessing were devised to enable the LDA analysis, which will be described in detail in this section. Furthermore, we describe the technical details of the Granger Causality Test procedure.

The "Al Naba" dataset comprises of numerous PDF files that were parsed by Optical Character Recognition (OCR) solutions, which were not optimized for the Arabic language. After trying multiple parsers, we discovered, through qualitative

review, that the "PDFplumber" framework provided acceptable results. However, during processing, the words were "mirrored", which resulted in certain words such as "PKK" being loaded as "KKP" in the dataset. Consequently, the data had to be reversed after being read-in. We then proceeded to remove the non-informative parts from the resulting strings by eliminating control characters, punctuation, numbers, email addresses, and other irrelevant elements. Despite the OCR results being qualitatively acceptable enough to move forward, we encountered numerous instances where words were inaccurately read-in, resulting in dataset contamination. As the Arabic language consists of many distinct words (approximately 12 million) compared to English (around 200,000 words), there is a lack of NLP-preprocessing frameworks to filter our vocabulary. Hence, to address this issue, we filtered the corpus against a reliable and substantial source of Arabic words. The Alittihad newspaper corpus was selected as a comparison template for our study due to its diverse spectrum of political and religious topics. However, due to the distinct communication style developed by ISIS, which involves the use of potential "slang" words and expressions unique to their organization, we decided to supplement the comparison template by manually reviewing and adding the top 1000 most prevalent words from the "Al Naba" corpus to the filtered dictionary, provided they were valid words. To remove the stop words, we used the Arabic Stopword dictionary from the Natural Language Toolkit (NLTK). However, this dictionary was not comprehensive enough to catch the majority of stop words. Thus, we decided to enhance the stop word dictionary ourselves. This further emphasizes the need for the scientific community to improve the coverage of Arabic language in NLP frameworks to facilitate working with this type of text data.

Given the sub-optimal data preprocessing and the analysis focus on outgroups, we devised a novel approach for training data selection in LDA. Rather than training the model on the entire corpus, we restricted it to a selected subset that predominantly pertains to outgroup topics. In addition to direct outgroup mentions, we performed a quantitative analysis of the 40 surrounding words to identify further terms closely associated with outgroups. Through this process, we were able to determine pages of the corpus for the training process in which the outgroups were mentioned only as synonyms or related remarks. To be selected for training, a page had to contain more than three of any of the signal words. When applied to the entire corpus, the LDA model used only tokens from the training corpus, thereby creating more focused topics in a sub-optimal data scenario. However, this approach may result in the loss of information about outgroups that could have been discussed using different vocabulary on pages that were not chosen or identified for training. We tried to balance this trade-off by carefully selecting signal words. The resulting analysis, particularly the increased prevalence of the "residual topic" number 3 over time, which corresponds to the declining power projection capabilities of ISIS, suggests a good balance between our trade-off considerations. We deliberately decided against the usage of a "seed word"-guided LDA analysis, which is using a prior distribution of certain words of the vocabulary and therefore a-priori clusters those terms into a predefined topic. The reasoning behind this is that we wanted to see if any of the outgroups that we are interested in are actually rhetorically treated similar, in other words, are

Table 1 Variable list for the LDA model

K	Number of topics
V	Number of words in the vocabulary
D	Number of documents in a corpus
d	Document
$\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\lambda_\beta})$	Word distribution for topic k
$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\lambda_\alpha})$	Topic distribution for document d
$\boldsymbol{\theta}$	Document topic matrix with rows $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$
$\boldsymbol{\lambda_\beta}$	Parameter of the Dirichlet prior on the per-topic word distribution
$\boldsymbol{\lambda_\alpha}$	Parameter of the Dirichlet prior on the per-document topic distributions
N_d	Number of words for document d
$z_{nd} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$	Topic for the nth word in document d
$\boldsymbol{z}$	Global topic assignments
$w_{nd} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{nd}})$	nth word in document d
$\boldsymbol{\beta}_{z_{nd}}$	Prevalence of the nth word for topic z

discussed by ISIS in the manner that it is factually appropriate to join those different outgroups into the same topic.

1.1 Latent Dirichlet Allocation

Finally, we want to mathematically describe the main Topic Model used in this paper. As Weisser et al. [1] describe the LDA algorithm in an excellent manner in their paper, we closely follow their notation.

Latent Dirichlet Allocation (LDA) is a statistical model that has gained widespread popularity for uncovering hidden topics in textual data. The model was first introduced by Blei et al. [2] in a seminal paper that proposed a generative process for documents, assuming that each document is a mixture of underlying topics, where the mixture proportions are distributed according to a latent Dirichlet distribution. A topic is defined as a probability distribution over all words in the corpus. To avoid duplicate indexing of documents, each document is associated with a unique integer value ranging from 1 to D. The LDA model assumes that each document d in a corpus consisting of $d = 1, \ldots, D$ documents is generated as a weighted combination of topics, which can be defined as follows:

1. K topic distributions are determined as $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\lambda}_\beta)$, where $\boldsymbol{\lambda}_\beta = (\lambda_{\beta 1}, \ldots, \lambda_{\beta V})$ represents the word relevances in a topic k.
2. The distribution over topics for document d is determined as $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\lambda}_\alpha)$, where $\boldsymbol{\lambda}_\alpha = (\lambda_{\alpha_1}, \ldots, \lambda_{\alpha_K})$ represents the vector of topic relevances for the corpus.
3. To generate the N_d words w_{nd}, $n = 1, \ldots, N_d$ for document d,
   (a) a topic $z_{nd} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$ is chosen, and
   (b) the corresponding words $w_{nd} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{nd}})$ are determined, where $\boldsymbol{\beta}_z$ is the vector of word occurrence probabilities $p(w|z)$ given topic z.

The LDA model involves hyperparameters represented by Dirichlet parameters $\boldsymbol{\lambda}_\beta$ and $\boldsymbol{\lambda}_\alpha$. The former consists of all topic-specific word occurrence probabilities β_{kn}, while the latter includes all document-specific topic occurrence probabilities θ_{dk},

which can be construed as the likelihood that the topic k generated the document d. By marginalizing over the latent topics, the generating process for the words of a document d can be expressed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$:

$$p(w_{nd}|\boldsymbol{\theta}_d, \boldsymbol{\beta}) = \sum_{k=1}^{K} p(w_{nd}|z = k, \boldsymbol{\beta}) p(z = k|\boldsymbol{\theta}_d), \tag{1}$$

revealing that the LDA model is considered a type of mixture model, with the word-specific multinomial models $p(w_{nd}|z, \boldsymbol{\beta})$ serving as mixture components and the topic probabilities $p(z|\boldsymbol{\theta}_d)$ acting as mixture weights. In other words, the LDA model combines different probability distributions to generate documents, with each distribution representing a different topic and the weights indicating how much each topic contributes to the document.

One can express the document generation process for a document d as the multiplication of word probabilities $p(w_{nd}|\boldsymbol{\theta}_d, \boldsymbol{\beta})$ and an integration over $\boldsymbol{\theta}_d$:

$$p(d|\boldsymbol{\lambda}_\alpha, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}_d|\boldsymbol{\lambda}_\alpha)(\prod_{n=1}^{N_d} \sum_{k=1}^{K} p(w_{nd}|z = k, \boldsymbol{\beta}) p(z = k|\boldsymbol{\theta}_d)) d\boldsymbol{\theta}_d. \tag{2}$$

Finally, one can obtain an estimate of the posterior distribution of the latent variables by employing either Gibbs sampling or Variational Inference [2].

1.2 Granger Causality Test

Originally introduced by Clive Granger [3], the Granger Causality Test serves as a statistical measure to assess whether one time series can predict another in a linear fashion. Crucially, this does not denote a strict causation. Within this study's scope, we employ this test to evaluate if the topics we've identified offer valuable forecasts concerning our outgroup-related variables. We delve into a bivariate case with two time series variables, represented as X and Y. In this context, α and γ are the Granger-Causality regression coefficients for Y and X respectively, τ is the Granger-Causality intercept, and p signifies the number of lags. The testing procedure unfolds as follows:

1. Examine each time-series for stationarity, using the Augmented Dickey-Fuller (ADF) test as a tool.
2. Ascertain the lag length p, which represents the count of past observations of X and Y integrated into the model, by using criteria like AIC or BIC.
3. Formulate the vector autoregression (VAR) model as:

$$Y(t) = \tau + \alpha_1 Y(t-1) + \ldots + \alpha_p Y(t-p) + \epsilon(t) \tag{3}$$

$$Y(t) = \tau + \alpha_1 Y(t-1) + \ldots + \alpha_p Y(t-p) + \gamma_1 X(t-1) + \ldots + \gamma_p X(t-p) + \epsilon(t) \tag{4}$$

Test for autocorrelation employing the Durbin-Watson test. If detected, increment p by 1 until we eliminate it[1].

[1] https://davegiles.blogspot.com/2011/04/testing-for-granger-causality.html

4. Assess the *absence* of Granger causality using the specified VAR model. Here, the null hypothesis H_0 posits: Past values of X lack any extra predictive power for forecasting Y beyond the information contained in the past values of Y alone. In mathematical terms, H_0: $\gamma_1 = \ldots = \gamma_p = 0$.
5. Engage in hypothesis testing, resorting to an F-test and Chi-Squared-based test, to compare the models.

   - A rejection of the null hypothesis H_0 hints at X Granger-causing Y, signifying the historical values of X enhance the prediction of Y.
   - Contrary, if the null hypothesis H_0 stands, it suggests X doesn't Granger-cause Y.

A salient observation from our efforts is that, despite rigorous data transformations, we couldn't secure normality in our data residuals. This puts the reliability of the F-Test-Statistics into question. Thus, our analysis will focus solely on the variables where statistical significance has been established by any tests that use the Chi-squared distribution as a basis for their test statistics.

In our specific context, our primary interest lies in discerning the causal influence of the topics on selected outgroup-related variables. However, the potential influence of these outgroup-related variables on the topics has also been tested. It's important to note that Granger causality tests were conducted for various lag lengths, of which each test is discrete, meaning it is not affecting the others. Consequently, it is entirely feasible to encounter a significant result indicating a Granger causal relationship at one lag length, while finding an insignificant result at another lag length.

References

[1] Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T., Säfken, B.: Pseudo-document simulation for comparing lda, gsdmm and gpm topic models on short and sparse text using twitter data. Computational Statistics (2022)

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. NIPS'01 (2001)

[3] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424–438 (1969)