Adapting Topic Modeling for Computational Analysis of Framing Processes

Adapting Topic Modeling for Computational Analysis of Framing Processes

by

Chase Mattingly

A Thesis

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Data Science

Lehigh University

May 2023

Thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science in Data Science.

Adapting Topic Modeling Techniques for Social Media Framing
Chase Mattingly

Date Approved

Eric PS. Baumer

Parv Venkitasubramaniam

TABLE OF CONTENTS

LIST OF FIGURES

Figure 1. Different frames deployed in the *New York Times* and *The Guardian* [1]

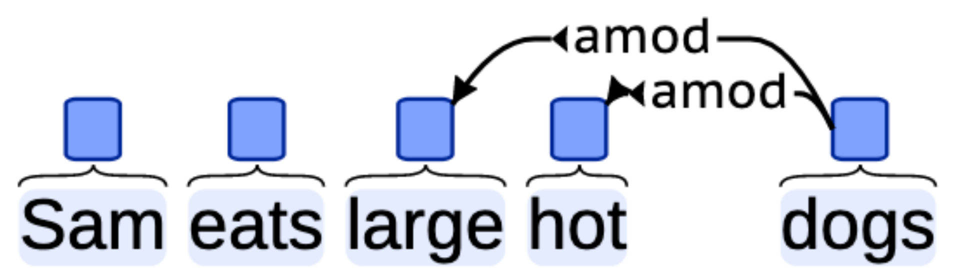

Figure 2. *amod* grammatical relation visualization

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	MWE	Loose	Special	Other
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figure 3. Universal Syntactic Relations used in UD v2. [3]

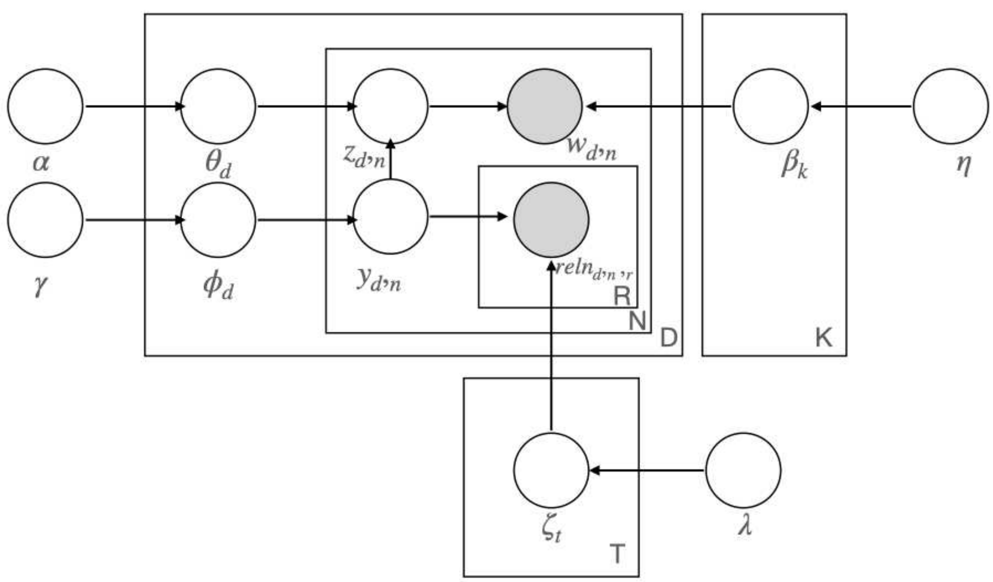

Figure 5. Latent Theta Role Plate Diagram

```
Require: words w ∈ corpus D = (d_1, d_2, . . . , d_m)
procedure LDA-GIBBS(w, α, β, T)
    randomly initialize z and increment counters
    loop for each iteration
        loop for each word w in corpus D
            Begin
                word ← w[i]
                tp ← z[i]
                n_{d,tp} −= 1; n_{word,tp} −= 1; n_{tp} −= 1
                loop for each topic j ∈ {0, . . . , K − 1}
                    compute P(z_i = j|z_{−i}, w)
                tp ← sample  from  p(z|.)
                z[i] ← tp
                n_{d,tp} += 1; n_{word,tp} += 1; n_{tp} += 1
            End
    Compute φ^(z)
    Compute θ_d
    return z, φ^(z), θ_D                                      ▷ Output
end procedure
```

Figure 6. Psuedocode for LDA Gibbs sampling

ABSTRACT

This thesis investigates a new approach for leveraging hierarchical topic modeling techniques to analyze and compare dominant frames found during major current events. We focus on the COVID-19 pandemic as it was an international crisis at an unprecedented magnitude, and one of the first of its kind to have full media coverage and social media discussion. We present the latent theta role model, a computational approach to framing analysis that develops latent variables in the form of distribution over words and distributions over grammatical relations to help understand the link between words and grammatical relations. With this newfound understanding of topics and theta roles, this technique can provide clearer insights about framing over Latent Dirichlet Allocation (LDA) topic modeling results. As a result, frames can be developed or solidified from previous qualitative framing analysis.

INTRODUCTION

Communication in the digital age has made framing a critical component in understanding how people perceive information. We define framing as the concept in which information shapes the presentation to influence how people perceive and react to it. The vast expanse of social media platforms and the multitude of information sources have made framing more intricate and difficult to handle. Therefore, there is an increased demand for computational tools to assist in framing analysis. Although framing can be used in various contexts, including politics, marketing, and other sectors, this paper will focus on framing related to the COVID-19 pandemic. It is essential to note that the findings and techniques discussed in this thesis are not limited to the COVID-19 pandemic and can be applied to framing analysis in general.

Since framing analysis is predominantly qualitative, in which we examine a small dataset with manual coding, the vast amount of data generated on social media platforms makes it challenging to analyze and identify the dominant frames that shape public discourse on the pandemic. Easy access to this data prompts researchers in both computation and social sciences to utilize various computational methods to explore frames in large-scale datasets [1].

Various prior studies have utilized topic modeling as a powerful tool for computational framing analysis [4]. Since topic modeling can efficiently extract latent topics from large text, it is a great base to find large-scale frames within a large corpus -- which would typically be time consuming and labor intensive with qualitative text coding. However, existing topic modeling techniques have their limitations, primarily relying on bag-of-words (BoW) techniques. Since framing revolves around how information is being discussed, and not what, it is important to note that BoW techniques can not consider the semantic relationships between words in a text. Therefore, techniques that consider how information is being communicated are necessary to improve the analysis of large-scale framing data.

This thesis aims to address the limitations of existing topic modeling techniques and provide a more in-depth approach for analyzing framing data. The proposed technique utilizes both hierarchical topic modeling techniques and the grammatical relationship between words to help provide a better understanding of frames in comparison to vanilla LDA techniques. By doing so, we hope to contribute to the development of more robust computational tools for framing analysis and help in determining the concepts and logical structures that individuals and communities develop to understand major current events.

FRAMING

Framing theory is based on the principle that how information is presented can influence readers perception and interpretation of said information. This eventually can shape the way users think about issues and events, influence their attitudes and beliefs, and even impact their behavior in a manner unprecedented.

The frames themselves refer to the specific ways in which the same message is presented -- emphasizing or changing certain parts of the message while downplaying or removing others. Figure 1 shows the same event presented in two different frames as a result of the Buffalo supermarket shooting in 2022. We can infer from the titles that a reader from The New York Times would view the gunman in a harsher manner compared to a reader from The Guardian with the changing of just a few words.

A prominent definition, widely used in both traditional and computational framing studies, was provided by Entman [7],

> "*perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.*"

Entman's definition describes four different components that make up a frame. The first component, problem definition, refers to how an issue or problem is defined or described. Framing can emphasize certain aspects of the issue and downplay others, shaping how people perceive and understand the problem. The second component, causal interpretation, pertains to how the problem's cause is explained or attributed. Framing can suggest different causes for the same problem, which can lead to different understandings of the issue. The third component, moral evaluation, refers to the ethical or moral values that are associated with the information stated. Framing can influence how people evaluate the information and the actions that should be taken to address the problem. Lastly, treatment recommendation relates to the actions and solutions that are proposed to address the problem. A problem's framing influences what actions are considered viable and or desirable, depending on how the problem is defined, what causes are assigned, and what moral values are invoked.

MODELING

We introduce a latent theta role model that looks to leverage aspects of the Latent Dirichlet Allocation (LDA) model with the addition of a latent theta role variable. A theta role, in generative grammar, is defined as the formal device for representing syntactic argument structure - the number and type of noun phrases - required syntactically by a particular verb [5]. LDA is a probabilistic model that allows for the identification of topics within a corpus of documents. The model looks to represent the corpus as k topics, where each topic is a distribution over words in the corpus's vocabulary. We extended this standard approach of hierarchical modeling by including a latent theta role during computation time. The addition of a latent theta role allows us to capture the variation in grammatical relations across the topics.

Each latent theta role is represented as a multinomial distribution of R grammatical relations. R is defined as the number of unique grammatical relations found within the corpus. Similarly, LDA topics are characterized by a multinomial distribution of words. Each grammatical relation represents a syntactic relationship between constituents in a clause. In Figure 2, we can a visualization of amod: an adjectival modifier of a noun (or pronoun) that is any adjectival phrase that serves to modify the noun (or pronoun).

In order to obtain said grammatical relations, we leverage Malter Parser and Stanford CoreNLP to obtain grammatical dependency trees over a given sentence. The table in Figure 3 lists the 37 universal syntactic relations used in UD v2. It is a revised version of the relations originally described in Universal Stanford Dependencies: A cross-linguistic typology (de Marneffe et al. 2014) [3]. Since a grammatical relation can be either a governor or a dependent, the number of possible universal syntactic relations doubles, leaving us with 72. The latent theta role is modeled as a Dirichlet distribution over these grammatical relations, with a hyperparameter γ controlling the strength of the prior distribution.

We fit our model using Gibbs sampling, which iteratively updates the topic and theta role assignments for each word in the corpus. Gibbs sampling is an effective method because it allows us to sample the posterior distributions over the latent variables without having to compute the normalization constant [6].
In the context of Latent Dirichlet Allocation (LDA), the normalization constant is found within the joint probability distribution of the words and latent variables. To compute the normalization constant for this distribution, we need to sum over all possible values of z, θ, and ϕ (in base LDA). However, the number of possible combinations of topic assignments, topic proportions, and word probabilities grows exponentially with the number of words and topics. It is even greater when we also include all possible variables within our latent theta model. Since it grows exponentially, it is computationally inefficient to compute the constant directly. Gibbs samploing was chosen over variational inference because we are looking to leverage accuracy over speed. Variational inference is a deterministic method that approximates the posterior distribution with a simpler distribution: mean-field Gaussian. Mean-field Gaussian assumes that each variable in the model is independent and normally distributed, with its own mean and variance parameters. This assumption simplifies the computation of the posterior distribution and can be faster and more scalable than Gibbs sampling. But, in turn, variational inference can be less accurate due to the use of simple posterior samples.

The inference process of our model is as follows and can be seen in Figure 4:

1) Let there be K latent topics, where each topic $\beta_k \sim Dir(\eta)$ is a multinomial distribution over the V words in the vocabulary, generated from a Dirichlet distribution parameterized by η.
2) Let there be T latent theta roles, where each latent theta role is a multinomial distribution over the R grammatical relationships, generated from a Dirichlet distribution parameterized by λ: $\zeta \sim Dir(\lambda)$.
3) For each document d
   a) Find the N number of words for the document d
   b) For each word w in document d :
      i) Assign a topic z using the probabilistic distribution of latent topics, each probability of a topic represented as:

$$k \in K \qquad P(z=k|w,d) = \frac{n_{d,k}+\alpha}{N+K\alpha} * \frac{n_{k,w}+\eta}{n_k+V\eta}$$

ii) Assign a theta role y using the probability distribution of latent theta roles:

$$t \in T \quad P(y=t|d,t,reln,z) = \sum_{reln=1}^{R} \frac{n_{d,t}+\gamma}{N+T\gamma} * \frac{n_{t,reln}+\lambda}{n_t+R\lambda} * \frac{n_{d,k}+\alpha}{N+K\alpha} * \frac{n_{t,k}+\gamma}{n_t+T\gamma}$$

iii) For each grammatical relationship found for word w, assign theta role y

4) Compute the θ matrix, which is a multinomial distribution over the K latent topics, drawn from a Dirichlet distribution parameterized by α: $\theta \sim Dir(\alpha)$. Each row of θ represents the topic distribution of a document d over the K latent topics.

5) Compute the ϕ matrix, which is a multinomial distribution over the T latent theta roles, drawn from a Dirichlet parameterized by γ: $\phi \sim Dir(\gamma)$. Each row of ϕ represents the distribution of each document d over the T latent theta roles.

6) Compute the ζ matrix, which is a multinomial distribution over the 64 grammatical *relns*, drawn from a Dirichlet parameterized by λ: $\zeta \sim Dir(\lambda)$. Each row of ζ represents the distribution of each theta role y over the R grammatical *reln*s.

7) Compute the β matrix, which is a multinomial distribution of the V words in the vocabulary, drawn from a Dirichlet parameterized by η: $\beta \sim Dir(\eta)$. Each row of β represents the distribution of each topic z over the V words.

As more research is needed to determine a computationally feasible method to evaluate latent theta role, this work will examine how the patterns identified in this model are in line with what researchers consider when examining framing processes.

IMPLEMENTATION

For our data within the project, we looked to leverage the Twitter API to obtain tweets during the COVID-19 pandemic. We collected general public tweets with keywords and hashtags relating to COVID-19. We also collected tweets from specific sources in news such as CNN and Fox, as well as government related organizations like the CDC. The tweets were collected from the start of the pandemic, December 2019, until the present.

The next set of data came from the Department of Health, where we scraped COVID-19 updates and related press releases in each state within the United States. In order to do so, it was necessary to use Scrapy. Scrapy is an open-source web crawling framework used to extract data from websites. It provides an efficient way to programmatically scrape and parse structured data from websites and is particularly useful for web scraping projects that require the extraction of a large amount of data. Its built-in schedule management of timing and requests was crucial due to the possibility of too many subsequent requests, locking us out of the website.

We then stored our tweets and documents in a database to provide a convenient and efficient way to store and retrieve large amounts of data. Since we had multiple organizations, keywords, and hashtags, we created relational tables for efficient access to subsamples of data. With a plan of adding more data from other news sources and platforms such as reddit, it was also useful to keep our data modular and organized for scalability and accessibility, especially since the project will be gaining new members for years to come.

The implementation of this model was carried out using the Python programming language along with several different libraries that provide natural language processing (NLP) capabilities. Specifically, for the use of NLP helper functions, we utilized the Natural Language Toolkit (NLTK) and Gensim. When looking for libraries that provide grammatical relationship parsing, we chose to leverage both the Malt Parser and

Stanford CoreNLP (Stanza) library. In regards to choosing a library for our probabilistic generative process, we utilize numpy for vectorization and fast computation times. In future work, computation times can be improved with GPU-accelerated libraries such as CuPy or Numba.

NLTK is a popular open-source library for NLP in Python, offering a wide range of tools for tasks such as tokenization, stemming, part-of-speech tagging, and sentiment analysis. We used NLTK extensively in this research to preprocess our data and leverage their Malter Parser wrapper. For example, we used the NLTK sentence and word tokenizer to break down documents into sentences and individual words, and then used their java Malter Parser wrapper to extract the grammatical dependency tree of each sentence.

The Malt Parser is another open-source NLP library that we used in this research. It is a transition-based parser that is designed to analyze the syntactic structure of sentences. As mentioned previously, we used Malt Parser to extract dependencies between words in sentences. These dependencies provide insight into the relationships between words and their syntactic functions, which can be useful when analyzing various linguistic states and *how* a sentence is being communicated. The Stanford CoreNLP library, at first, was used to parse grammatical dependency graphs, but the computation times were too long with its large neural model. For this reason alone, we switched to Malt Parser due to its fast and lightweight linear model and only use Stanza when said parser fails. Stanza also offers a wide range of NLP tools, like part-of-speech tagging, but none of which are used.

In Figure 5, we can see the psuecode for the implementation of Gibbs sampling for LDA from scratch. Since we wanted to make our algorithm efficient in sampling different latent variables, topics and latent theta roles, we needed to perform inference for said variables at the same time. In order to do so, we needed to implement Gibbs sampling from scratch using Python and numpy to incorporate the inference of the latent theta role. Our implementation steps can be seen in the inference steps section of the Modeling section. The main steps that were added are the probability calculations for the assigned latent theta roles to the different relationships of each word and the computation of the new ζ and ϕ matrices. We can find the related latent variables by sorting the rows of the ζ and β and getting the top number probabilities within each row.

RESULTS

Running our latent theta role model revealed possible underlying patterns for framing analysis through the connections of the variables and topics. A test analysis was conducted on CNN news articles.

The model identified K=10 topics, each topic containing the top 9 key words. The model also identified T=2 latent theta roles, each variable containing the top 9 grammatical relations. These results can be found below:

Topics:

Topic 0: 0.006806622496566505 * want, 0.006806622496566505 * people, 0.006655699603072791 * new, 0.006504776709579076 * years, 0.006353853816085363 * health, 0.006353853816085363 * get, 0.005599239348616792 * president, 0.005297393561629364 * back, 0.00514647066813565 * never, 0.004542779094160794 * best

Topic 1: 0.014128161852901527 * fire, 0.010837605748919954 * cnn, 0.009351548153573438 * people, 0.006804020847265123 * authorities, 0.006591726905072764 * thursday, 0.005317963251918607 * officials, 0.005105669309726247 * county, 0.004999522338630068 * department, 0.004893375367533888 * killed, 0.004574934454245348 * san

Topic 2: 0.009434982010222519 * world, 0.008354051174354145 * museum, 0.008199632483515806 * year, 0.006809864265970754 * time, 0.00572893343010238 * congress, 0.005111258666749024 * bill, 0.004802421285072345 * way, 0.004802421285072345 * billion, 0.00418474652171899 * bush, 0.00418474652171899 * according

Topic 3: 0.007753048911563651 * told, 0.007753048911563651 * mother, 0.0061766434027429465 * stiles, 0.0061766434027429465 * family, 0.0057467146276100264 * girl, 0.005603405035899053 * life, 0.004600237893922241 * man, 0.004313618710500294 * found, 0.004170309118789321 * cnn, 0.004026999527078348 * son

Topic 4: 0.006941938835890257 * simpson, 0.005931677466841779 * cnn, 0.005643031361399357 * people, 0.005354385255956934 * country, 0.005210062203235723 * government, 0.005065739150514512 * snow, 0.004921416097793301 * told, 0.004632769992350879 * musharraf, 0.0043441238869084565 * president, 0.004055477781466034 * navy

Topic 5: 0.01303902125906552 * iraq, 0.00863200452198239 * cnn, 0.00844039509863095 * military, 0.006524300865116546 * president, 0.006332691441765106 * iraqi, 0.005853667883386505 * bush, 0.005757863171710785 * security, 0.005374644325007904 * war, 0.005183034901656 4635 * air, 0.005087230189980744 * tuesday

Topic 6: 0.0075749047105936095 * children, 0.006288296691809132 * year, 0.005805818684764953 * first, 0.0056449926824168935 * international, 0.005484166680068834 * blackwater, 0.005001688673024655 * well, 0.0048408626706765955 * death, 0.004680036668328536 * state, 0.004680036668328536 * cnn, 0.0045192106659804766 * service

Topic 7: 0.005923599558450847 * new, 0.005612649450395684 * england, 0.005612649450395684 * london, 0.005457174396368104 * day, 0.005457174396368104 * around, 0.005301699342340522 * time, 0.004835274180257778 * team, 0.004213373964147453 * friend, 0.004057898910119872 * real, 0.0039024238560922903 * work

Topic 8: 0.008116675432486602 * first, 0.007121636412742186 * vick, 0.006410894255781888 * last, 0.006410894255781888 * home, 0.005558003667429531 * win, 0.004989409941861292 * season, 0.004847261510469232 * league, 0.004562964647685113 * match, 0.004278667784900994 * world, 0.004278667784900994 * friend

Topic 9: 0.012315687510570581 * south, 0.008779347775949815 * obama, 0.008471839972939314 * clinton, 0.00754931656390781 * carolina, 0.006011777548855303 * percent, 0.00555051584433955 * house, 0.0053967619428343 * democratic, 0.0053967619428343 * wine, 0.0053967619428343 * university, 0.005089254139823799 * cnn

Theta Roles:

Theta Role 0: 0.19043038241572882 * pobj.dep, 0.18858550984371292 * det.gov, 0.18687241388398385 * prep.gov, 0.15511425032285267 * amod.gov, 0.12928603431462984 * amod.dep, 0.11966634161768969 * nsubj.gov, 0.10372137153098068 * nn.gov, 0.09897741348865403 * nn.dep, 0.08290066678965817 * dobj.gov, 0.07973802809477373 * dobj.dep

Theta Role 1: 0.19330746978831542 * prep.gov, 0.19060396334045254 * pobj.dep, 0.18864392116575196 * det.gov, 0.1461312822731082 * amod.gov, 0.13207304874422124 * amod.dep, 0.12774743842764064 * nsubj.gov, 0.09287220525020952 * nn.gov, 0.09084457541431236 * dobj.gov, 0.0888169455784152 * nn.dep, 0.08847900727243234 * aux.gov

After analyzing the results, more understanding and research is needed to be done with our model in order to optimize the number of latent theta roles. We also found that the latent theta roles are hard to understand in relation to the topics, therefore, we will need a better method of displaying how they relate to topics for more efficient analysis. We will continue to work with our content analysis team in order to qualitatively analyze our results on our larger scraped dataset and improve the algortithm.

REFERENCES

[1] Mohammad Ali, Naeemul Hassan: “A Survey of Computational Framing Analysis Approaches”, 2022; [https://aclanthology.org/2022.emnlp-main.633.pdf]

[2] David Bamman, Brendan O’Connor, Noah A. Smith: “Learning Latent Personas of Film Characters”, 2023; [https://aclanthology.org/P13-1035.pdf]

[3] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning: “Universal Stanford Dependencies: A cross-linguistic typology”, 2014; [https://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf]

[4] John W. Mohr, Petko Bogdanov: “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding”, 2013; [https://www.sciencedirect.com/science/article/abs/pii/S0304422X13000661]

[5] Richard C. DeArmond: “Theta Roles”, 2002; [https://www.sfu.ca/person/dearmond/322/322.theta.roles.htm]

[6] George Casella, Edward I. George: “Explaining the Gibbs Sampler”, 1992; [http://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/Basic/CasellaGeorge1992.pdf]

[7] Entman, R. M. (1993) ‘Framing: toward clarification of a fractured paradigm’, *Journal of Communication* [https://onlinelibrary.wiley.com/doi/10.1111/j.1460-2466.1993.tb01304.x]

VITA

Chase Mattingly was born in the Bay Area of Californa on October 9^{th}, 2001. They attended grades Pre-K to 12 in the same area and graduated high school in 2019. The following August they entered Lehigh University and three years later, May of 2022, graduated with a Bachelor of Science in Computer Science and Engineering. The following summer, they started their Master of Science in Data Science at Lehigh University and received their degree in May 2023.