Dynamic Fed Attention

Stanford CS224N Custom Project

Amar Venugopal
Department of Economics
Stanford University
amarvenu@stanford.edu

Abstract

The Federal Reserve ("the Fed") is the principal monetary policy body of the United States, and its policies have ramifications throughout the economy. As a result, significant attention is paid to the statements and speeches of its principal officers and committees. Past literature has sought to apply static topic models to these texts in an attempt to gauge the attention of the Federal Reserve (Jegadeesh and Wu, 2017). This project seeks to extend this literature to the use of dynamic embedded topic models (D-ETM), formulated by Dieng et al. (2019a), comparing the results to dynamic latent Dirichlet allocation (D-LDA) established in Blei and Lafferty (2006). I show that D-ETM can significantly outperform D-LDA in terms of quantitative measures such as topic coherence and diversity and investigate the variable effect of using different classes of embedding models for D-ETM.

1 Introduction

Topic modeling have gained significant ubiquity across a variety of fields in recent years, due in large part to their easy interpretability. Traditional methods of topic modeling have typically focused on probabilistic bag-of-words approaches, most notably Latent Dirichlet Allocation (LDA), which models topics as probabilistic distributions over words and documents as mixtures over topics (Blei et al., 2003). This approach has also been extended to dynamic topic models, most notably dynamic LDA (D-LDA) (Blei and Lafferty, 2006). These extensions allow for the compositions of topics to change across time, and are particularly compelling for corpuses that span a significant period of time. However, these traditional approaches fail to take advantage of the power of word embeddings, which have offered demonstrable performance improvements in other fields of NLP. More recent developments include the creation of the embedded topic model (ETM), which models topics as points in embedding space and utilizes the structure of that space to similarly model documents as mixtures over topics (Dieng et al., 2019b). Most recently, the same authors developed the dynamic embedded topic model (D-ETM), which seeks to address the aforementioned issues inherent to LDA and, by extension to D-LDA, by extending the authors' previous work on ETM to the dynamic setting (Dieng et al., 2019a). This algorithm provides a prescription to model a corpus of documents as mixtures over topics, where the topics themselves are now time-varying vectors in embedding space. The authors demonstrate that this deep-learning based approach outperforms D-LDA, both in terms of quantitative and qualitative benchmarks.

This project seeks to implement the D-ETM algorithm and apply it to a corpus of Federal Reserve Open Market Committee (FOMC) meeting statements and speeches spanning several decades and demonstrate performance improvement over benchmark methods demonstrated in existing papers. The FOMC is the principal policymaking body of the Federal Reserve, and as a result its meetings, statements, and decisions are highly scrutinized by the broader economics and finance community. Existing literature has focused on the application of more traditional topic modeling, such as LDA, to Fed-generated text data (Jegadeesh and Wu, 2017). Demonstrating that the use of embedding-based modeling frameworks can improve results over existing papers and benchmarks will demonstrate both the power of such models to improve performance without sacrificing interpretability. Fur-

thermore, since the corpus of FOMC statements and speeches is relatively small compared to most machine learning objectives, the success of this approach in this setting demonstrates the value of pretraining/transfer learning, primarily through the use of pretrained embedding models. These results help serve as a proof of concept for further adoption of these methods in the economics and finance literature, which often deal with datasets significantly smaller than is standard in computer science.

In particular, I examine D-ETM performance based on underlying embeddings produced by BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and GloVE (Pennington et al., 2014). I also consider different methods of producing word embeddings from the former (transformer-based) models. The results are then compared to the baseline D-LDA algorithm. Quantitative results show that D-ETM powered by BERT embeddings offers substantial performance improvements over D-LDA, while some form of embedding produced by each of the 3 models demonstrates qualitatively better topic quality as compared to D-LDA. There is, however, significant variability in both quantitative and qualitative performance across different embedding types, with not all outperforming D-LDA in all settings.

2 Related Work

One of the earliest topic models, probabilistic latent semantic analysis (PLSA), was put forth by Hofmann (1999) and models word and document co-occurrence as a mixture over multinomial distributions. Latent Dirichlet Allocation (LDA), formulated by Blei et al. (2003), built on this work by introducing a Dirichlet prior for the distributions of documents over topics and topics over words. This modification has the benefit of enforcing our prior belief that documents are mixtures over small numbers of topics, and that topics are similarly mixtures over small numbers of words. LDA quickly gained significant popularity and spawned an increasing literature of extensions and refinements. One key application of this method was to a corpus of FOMC meeting minutes by Jegadeesh and Wu (2017), in which the authors demonstrate that LDA can capture the varying attention that the Fed's leadership pays to various topics of economic interest.

The most notable extension of LDA, in the context of this project, is the introduction of dynamic LDA (D-LDA), postulated by Blei and Lafferty (2006). Rather than simply viewing topics as static objects, D-LDA assumes that topics evolve over time, with the parameters governing the distributions of topics over words at time $t + 1$ drawn from a normal distribution parameterized by the parameter values at time t, thereby ensuring smoothness in topic evolution. D-LDA provided a significant extension of LDA that was particularly useful when applied to corpuses of documents that span significant lengths in time, in which case assuming static topic distributions is infeasible.

In parallel, embedding models, which seek to represent words as elements of a vector space, gained significant popularity. Mikolov et al. (2013a) introduced the skipgram embedding model, soon extending it and creating the word2vec embedding model in Mikolov et al. (2013b). These neural network-based models demonstrated significant performance improvements over past iterations, including those based on latent semantic analysis. In years since, more sophisticated models, such as GloVE (Pennington et al., 2014), extended these approaches and leveraged improved computational performance, more complex models, and greater corpuses of data to create extensive dictionaries of static word embeddings. Most recently, however, these models have been largely overshadowed by the emergence of transformer models (Vaswani et al., 2017), which incorporate mechanisms such as attention to better model the relationships and interdependencies between text and leverage embeddings as part of this process. BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) are highly popular examples of such models.

Recognizing the power of these new approaches, a key extension to topic modeling was proposed by Dieng et al. (2019b) in the form of the embedded topic model (ETM), which modifies the distributional assumptions underlying LDA to incorporate word embeddings; topics are instead modeled as points in embedding space, with the inner products between these topic vectors and word vectors a natural proxy for topic-word assignment. This approach documented significant performance improvements over LDA, and quickly spawned a dynamic extension, D-ETM (Dieng et al., 2019a), which was similarly shown to outperform D-LDA across a variety of settings and metrics when utilizing pretrained, skipgram word embeddings.

3 Approach

The primary approach of this project is to apply D-ETM to a corpus of Federal Reserve statements and speeches and demonstrate performance improvements relative to more traditional baseline models, such as D-LDA. In doing so, I first collect the datasets and clean the text to result in a corpus of cleaned and filtered documents. I then split the corpus into 4 distinct time segments, corresponding to the tenures of the 4 different Federal Reserve Chairs who served during the time span the data contains.[1] While not all speeches are made by these Chairs and the statements in the corpus are issued by committee, Fed Chairs play an extensive role in setting monetary policy and so I use their tenures as a proxy for regime changes in what the Fed may be primarily concerned with. Their terms often tend to coincide with presidential terms, which allows them to proxy for the broader political climate to a certain extent.

I then run my primary baseline model, D-LDA, via a wrapper (provided by python package `gensim`) around the original C++ code provided by the original authors (Blei and Lafferty, 2006).[2] I then run the D-ETM model, using as a base the code provided by the authors Dieng et al. (2019a) in their implementation.[3] Note that I modify their code in order to accommodate the new dataset and my particular pre-trained embeddings.

For the embeddings themselves, I consider static (or static-analog) representations produced by BERT[4], GloVE[5], and T5[6]. While GloVE embeddings are inherently static, BERT and T5 are models designed to produce contextual embeddings, in which the embeddings of tokens vary depending on their context. The D-ETM model, however, takes as an input a time- and instance-invariant mapping from words to embeddings, requiring a degree of "staticness" to the embeddings produced. I therefore consider "static analog" embeddings produced by BERT and T5. In the case of BERT, these embeddings are obtained via the underlying embedding weight matrix, which is included as part of the `huggingface` BERT model implementation (henceforth referred to as "SBERT"). For T5, which lacks this feature, I consider embeddings produced via the last hidden state of the model when it is faced with a given word ("T5L") and a pooled average over all hidden layers of the model ("T5P").

I then further consider a static representation of contextual embeddings produced by BERT ("CBERT"). In particular, I sample a subset of 200 documents from my corpus, achieving coverage of approximately 90% of the underlying total vocabulary, and compute contextual embeddings for each word in each document. I then aggregate over the entire sample and average all embeddings for each word across all appearances in these documents. The result produces a one-to-one mapping from words to embeddings that is produced based on contextual information contained in the corpus. I can then compare these results with the aforementioned context-independent static representations.

For all transformer models, I generate a single vector for each word by averaging over the static (or static-analog) vector representations of all tokens generated by the model's corresponding tokenizer, thereby taking into account any relevant subword embeddings. In the case of GloVE and CBERT, if a given word in the vocabulary is not covered by the available embeddings, it is simply assigned an embedding vector of 0.[7]. These missing generally words correspond to the rarer words in the corpus, particularly in the case of CBERT since they are not covered by the subsample of the data from which contextual embeddings are generated. As a result, the impact of their exclusion on final model outputs is assumed to be negligible.

[1] The chairs in question are: Alan Greenspan (1987-2006), Ben Bernanke (2006-2014), Janet Yellen (2014-2018), and Jay Powell (2018-present)

[2] In doing so, I use a pre-built binary found at `https://github.com/magsilva/dtm/tree/master/bin`

[3] `https://github.com/adjidieng/DETM`

[4] `https://huggingface.co/bert-base-uncased`

[5] `https://nlp.stanford.edu/data/glove.6B.zip`

[6] `https://huggingface.co/t5-base`

[7] This corresponds to approximately 1% of the overall vocabulary for GloVE and 11% for CBERT

4 Experiments

4.1 Data

The primary dataset for this project is the corpus of FOMC meeting statements and speeches made by Federal Reserve Governors. Speeches are obtained from an existing repository[8] and span from 1996-2020, while the meeting statements are pulled from the FedTools[9] API and span from 1994-2023. The preprocessing steps taken on this dataset include: removing stopwords; employing a part-of-speech tagger to remove words corresponding to "uninformative" text that do not correspond to underlying economic discussion (e.g. proper nouns, etc.); lemmatizing words to remove the effect of specific conjugations; and removing words that occur in fewer than 1% or greater than 90% of the documents. The resulting cleaned dataset consists of 4938 unique words across 1650 documents.

4.2 Evaluation methods

The primary evaluation metrics used are topic coherence and topic diversity. Topic coherence attempts to provide a measure of the cohesion of a topic by measuring semantic similarity between its top words. In this project I adopt the C_V coherence metric developed by Röder et al. (2015), which the authors demonstrate outperforms other competing topic coherence metrics across a variety of settings.[10] Syed and Spruit (2017) provide a detailed breakdown of the C_V coherence measure, which operates by first segmenting each topic's top N words (a set denoted W) into pairs, matching each individual word with the set of all other words:

$$S = \{(w_i, W) | w_i \in W\}$$

Probabilities of word co-occurrence in the dataset are then calculated via a Boolean sliding window, where frequencies are counted based on expanding the cardinality of the document set by imposing a sliding window with a step size of one word. The co-occurrences of words w_i, w_j in each sliding window are then used to calculate the probabilities $p(w_i, w_j)$. The cohesion between two words is obtained using normalized pointwise mutual information (NPMI):

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)}$$

Where ϵ is included to prevent evaluating logarithms at 0. In order to calculate the semantic support of the top N words in W, for every element in S context vectors are created to represent the first and second elements of the tuple by u, w, respectively. These are calculated by summing all the pairwise NPMI values in the given tuple element and raising each term in the sum to some power γ, in order to weight higher NPMI values more. Each $S_i \in S$ can then be assigned a "confirmation measure":

$$\phi_{S_i}(u, w) = \frac{\sum_{i=1}^{|W|} u_i^T w_i}{\|u\|_2 \|w\|_2}$$

The final topic coherence score for the given topic is then the mean over all these measures:

$$C_V = \sum_{S_i \in S} \phi_{S_i}(u, w)$$

Since this is a per-topic coherence measure, in our multi-topic setting the aggregate coherence measure is simply the mean over the individual topic coherence measures.

For topic diversity, I implement and utilize a measure known as "proportion unique words", which measures the percentage of unique words in the top words of all topics (Dieng et al., 2019b). For topics enumerated $1, ..., n$, this is calculated as:

$$PUW = \frac{|\bigcup_{i=1}^{n} W_i|}{n|W_i|}$$

[8] `https://www.kaggle.com/datasets/natanm/federal-reserve-governors-speeches-1996-2020`

[9] `https://github.com/David-Woroniuk/FedTools`

[10] The python package `gensim` provides a built-in measure for this

Where we implicitly assume that we consider the same number of top words for each topic (that is, $\forall i, j, |W_i| = |W_j|$).

In keeping with the original D-ETM paper, I also consider a measure of topic "quality", defined to be the product of coherence and diversity (Dieng et al., 2019a):

$$Q = C_V \cdot PUW$$

For all quantitative metrics presented, a higher measure is more desirable.

I also consider qualitative evaluations of the models by examining the topics produced. In particular, I examine the top 10 words associated with each topic at various time periods for each model and assess whether the result appears to provide adequate coverage of the underlying economic context that the text data is meant to capture. For selected models, I dig deeper and examine the evolution of the weight applied to certain words by certain topics across time, which provides a measure of how the dynamic topic produced by these models capture changes in the economic environment.

4.3 Experimental details

I run all models using 10 topics, in keeping with previous applications to Federal Reserve text data (Jegadeesh and Wu, 2017). All models are fitted with the same (cleaned) text data, described in Section 4.1, in order to allow for a standardized comparison across modeling methodologies. For all D-ETM implementation, I train for 150 epochs with a batch size of 100 documents and a learning rate of 0.005; these are generally consistent with those used by the authors in their original implementation, albeit with a smaller batch size in keeping with my relatively small dataset (Dieng et al., 2019a). The number of epochs was selected based on observation of the training loss, ensuring that it achieves a sufficient plateau.

For the baseline D-LDA model, the primary hyperparameter is the maximum number of EM steps used to fit the constituent LDA models; this value is set at 20, consistent with the original implementation (Blei and Lafferty, 2006). When evaluating the models, I consider the top 20 words of each topic to calculate coherence and diversity measures. Both models are trained on CPUs. Training times were comparable, with average times of approximately 17 minutes for D-LDA and 25 minutes for D-ETM.

4.4 Results

My primary findings indicate that while D-ETM can achieve significant performance improvements over D-LDA, this effect is highly dependent on the type of embedding used. Only the SBERT systematically outperform D-LDA, in terms of topic quality. A full comparison of model topic quality over time can be seen in Figure 1.

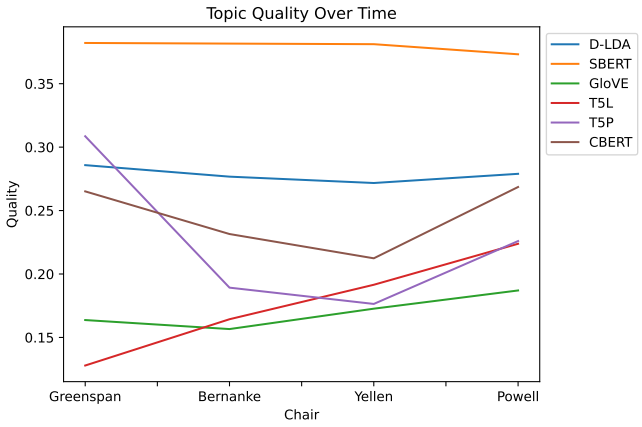


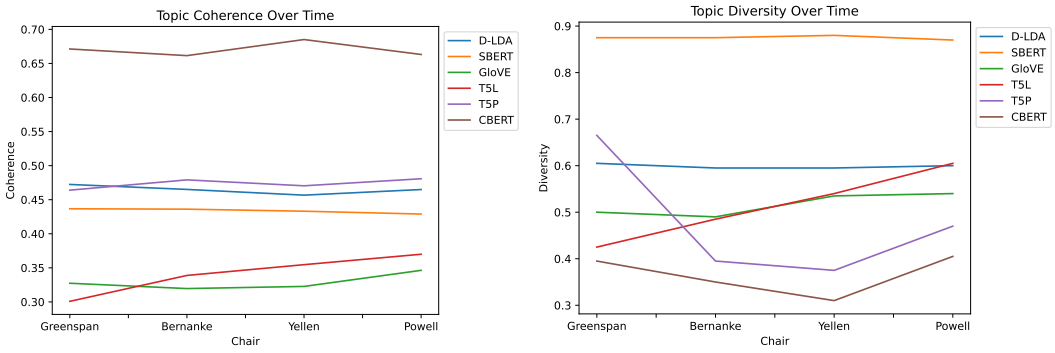
Figure 1: Topic qualities

Interestingly, GloVE embeddings, which are actually designed to be static and therefore seemed *a priori* to be strong candidates for D-ETM compatibility, exhibits poor performance across the board and is one of the lowest-scoring models overall. GloVE does, however demonstrate strong

qualitative performance (see Section 5). The clear performance difference between SBERT and GloVE two models, however, is quite interesting. One possible explanations include GloVE's lower dimensional representation (300 vs. 768 for BERT), which means that the embeddings must encode less information by construction. Another possible explanation could be that the the transformer architecture simply produces higher quality embeddings, but that the process by which these embeddings are made "static" is key. This could explain both the high performance of SBERT, relative to GloVE, as well as the low performance of CBERT and the T5 approaches, which may have failed to significantly outperform D-LDA due to sub-optimal embedding distillation strategies.

5 Analysis

Digging further into these quantitative results, an examination of the components of topic quality, namely coherence and diversity, are illuminating. They can be seen in Figure 2. From these plots, we can see that while BERT is not particularly strong in terms of coherence, it achieves excellent topic diversity; its topics are far more distinct from one another than those produces by other models. CBERT, however, displays the opposite effect; its topics are extremely coherent, but not at all diverse. The same occurs for T5P, which exhibits good coherence but poor diversity.



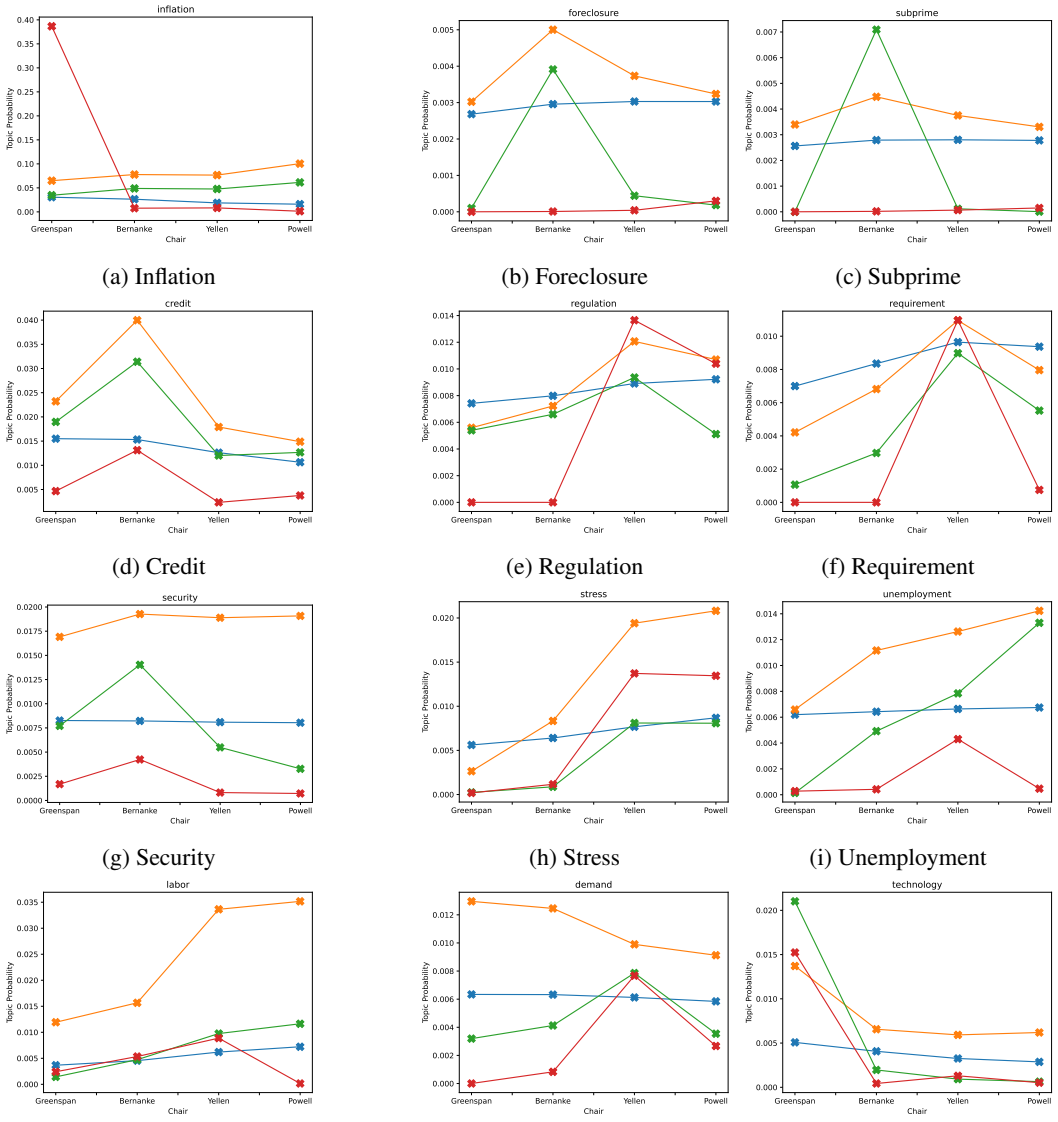
(a) Coherence

(b) Diversity

Figure 2: Topic coherence and diversity

These results are mirrored in qualitative assessments of topic quality as well. Specifically, the models that perform best according to the above quantitative metrics, particularly SBERT, appear to show the most reasonable evolution of word probabilities for topic assignment over time. In Figure 3 we can see the evolution of the probabilities with which several key words are assigned to their primary topic in the relevant model. Going through the words, we see some clear trends: words such as "foreclosure", "subprime", "credit", and "security" all have clear relations to the subprime mortgage crisis that fueled the financial crisis of 2008-2009, during Bernanke's chairmanship. SBERT and T5L seem to capture the bulk of these effects, but the other models largely fail to do so consistently. "Regulation" and "requirement" took on more importance during the post-crisis period as discussion turned towards how to prevent another collapse, during Yellen's tenure as chair; this is clearly captured by SBERT, T5L, and GloVE, but not by D-LDA. "Unemployment" and "labor" have increased in concern recently, as the period of strong economic growth post-crisis has led to record-low unemployment; "technology", in the meantime, has taken on diminished importance since the 1990s in Greenspan's tenure, when its role as a disruptor of the labor market was more pronounced. These trends are fully captured by SBERT and T5L, but not by D-LDA or GloVE. "Stress" has taken on greater importance in recent years as stress-testing has become an integral part of financial health assessments and investment risk; SBERT, T5L and GloVE seem to capture this, but D-LDA fails to do so. "Inflation" is becoming much more significant today as we are in a period of historic price increases, which is reflected in SBERT and T5L but not in GloVE or D-LDA. Finally, "demand" took on greater significance in the aftermath of the financial crisis when the Fed was concerned with stimulating demand in order to grow the economy, which is reflected by T5L and GloVE but not in SBERT or D-LDA.

No single model seems to capture every desired trend, but BERT is clearly the most successful overall. Interestingly, the results of this qualitative assessment seem to suggest that T5L and GloVE clearly outperform D-LDA in aggregate, which is not reflected in the quantitative measures discussed



(a) Inflation (b) Foreclosure (c) Subprime

(d) Credit (e) Regulation (f) Requirement

(g) Security (h) Stress (i) Unemployment

(j) Labor (k) Demand (l) Technology

Figure 3: Selected word evolutions for SBERT, D-LDA, T5L, GloVE

previously. This suggests that topic diversity, the metric by these models perform better, may actually be more reflective of our human assessment of topic quality than topic coherence. This is reinforced by the particularly poor qualitative performance of T5P and CBERT; these models are omitted from the plots above due to their poor topic diversity, which corresponds to very poor topic resolution as evidenced by even a cursory inspection of their top words per topic. Their inclusion would therefore fail to add useful information while adversely impacting readability of the Figure.

6 Conclusion

The experimental results demonstrate that D-ETM is capable of significantly performing D-LDA, demonstrating the capacity of word embeddings to provide performance improvements over more standard bag-of-words approaches, even when applied to a small dataset such as Fed speeches and statements, with D-ETM run on select embedding models demonstrating superior quantitative and qualitative performance relative to D-LDA. However, D-ETM demonstrates significant sensitivity to the choice of embedding used, appearing to favor static implementations over those that seek to

leverage in-sample context. Even so, static embeddings produced by BERT, a transformer model, significantly outperform those produced by GloVE, despite the latter being specifically designed to produce static word vectors.

This suggests that the increased sophistication of transformer models translates into higher quality static word representations, and that the demonstrated limitations of those models in this project is reflective more of the process through which static representations were created, rather than the power of transformer architecture for this use case. A key direction for future research would therefore be to utilize the static embeddings produced by Gupta and Jaggi (2021), who put forth a methodology to generate high-quality static word embeddings from contextual transformer models, including BERT. The results of this project suggest that this approach has the potential to provide significant performance improvements and yield more intuitive and useful topics.

References

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019a. The Dynamic Embedded Topic Model. ArXiv:1907.05545 [cs, stat].

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019b. Topic Modeling in Embedding Spaces. ArXiv:1907.04907 [cs, stat].

Prakhar Gupta and Martin Jaggi. 2021. Obtaining Better Static Word Embeddings Using Contextual Embedding Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5241–5253, Online. Association for Computational Linguistics.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. Association for Computing Machinery.

Narasimhan Jegadeesh and Di Wu. 2017. Deciphering Fedspeak: The information content of FOMC meetings. *Monetary Economics: Central Banks–Policies & Impacts eJournal*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. Association for Computing Machinery.

Shaheen Syed and Marco Spruit. 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.