Words Paint a Thousand Pictures: Evaluating Topic Models Using FOMC Transcripts

A Dissertation Presented in Fulfillment of the Requirements for the Degree of Masters of Research

Luke Cayanan

B. App Fin (Hons), Macquarie University 2010

Department of Computing
Faculty of Science and Engineering
Macquarie University, NSW 2109, Australia

Submitted November 2022

Declaration

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Signed: .....

Date: .....

Dedication

To Mum.

To my beautiful wife, Ngoccy. And my three children Derek, Beatrix and Raiden.

I am nothing without all of you.

Abstract

At its core, a topic model's primary task is to uncover patterns from a huge collection of unstructured data in an automated way. *Fast*. This architecture is naturally well suited for analysing large corpora (collections of text data). When it comes to analysing text, a topic model typically returns a group of words which are semantically linked. These word groups are called *topics*. As such, topic models are appealing to researchers in fields other than the machine learning and natural language processing (NLP) domains. Domains such as higher education, sociology and finance and economics, which frequently deal with these types of data. The focus of this study is on finance and economic data.

Recent advancements in the topic modelling space have sought to improve the quality or interpretability of topics. In particular, I focus on newer topic models which incorporate word embeddings and user guidance that result in models learning better topics. However, the application of the newer approaches, has rarely transitioned outside the machine learning and NLP research fields. Those non-NLP domains have largely been confined to the application of classic Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods.

I extend the work of Hansen et al. (2018), who use traditional LDA to investigate the topical structure underlying the US Federal Open Market Committee (FOMC) transcripts. They were able to show econometrically, that these transcripts contained signals about the behaviour of inexperienced FOMC members (*rookies*). I take this result and frame it as a machine learning problem, where I task a classifier to predict whether an FOMC member is a rookie, given the text in the transcripts. I then assess the efficacy of the newer topic models against Hansen et al. (2018)'s benchmark LDA model in making these predictions. I also compare the topic quality of the newer topic models against the traditional LDA as measured by metrics in NLP. I find that while the newer topic models improve topic quality versus LDA, they were unable to outperform LDA in the classification task.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 What is a topic model?

At its core, a topic model's primary task is to uncover patterns from a collection of *words*, the mechanisms of which are grounded in statistics and probability theory. Think of it as an automated version of what a human brain does naturally, when uncovering a pattern or structure from the same collection. This collection can range from a set of images to a set of genes. As long as a collection is encoded into an appropriate digital format, then a computer can be tasked to uncover its topical structure. *Fast*. This type of architecture makes it well suited for processing large amounts of unstructured data. In particular, topic models are appealing for researchers who are interested in natural language processing (NLP) tasks such as organising, summarising, visualising, searching, predicting and exploring large collections of digital text documents — or *corpora*.

Three foundational methods, developed in the late 1990s and early 2000s, form the spine of modern topic modelling research. They are largely known for their adaptability and improved generalisation through dimensionality reduction (among others).

Latent Semantic Analysis (LSA) (Deerwester et al., 1990). It is an automated way for retrieving documents and underpins modern day internet search engines. LSA's

premise is that a document's meaning can be represented by latent semantic indexes, which can be extracted from a term-document matrix using linear algebra. The final output is a set of k largest singular values, where k is some tuning parameter determined beforehand. The k^{th} largest component can be thought of as the k^{th} *topic*. One of LSA's drawbacks, however, is that it has difficulty in dealing with polysemous terms, that is, a term that has different meanings in different contexts.

Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) overcomes some of LSA's shortcomings (e.g. polysemy). The pLSI is known as a mixture model. It assumes that a document is a mix of k latent topics, where the mixing proportions is a list of fixed values and k, again, is some pre-determined number. Then a document's semantic representation is assumed to be a generative probabilistic process where the terms are drawn out of some distribution from k topics. pLSI has a notable constraint, however. It struggles to assign non-null probabilities to unseen documents, thereby limiting its effectiveness to small, static datasets.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Like pLSI, LDA is a mixture model. But, unlike the pLSI, LDA uses a Bayesian probabilistic approach, which helps overcome some of pLSI's limitations. The LDA procedure assumes that the document-topic proportions are drawn from a Dirichlet prior, or a distribution of discrete probabilities. In fact, Girolami and Kabán (2003) proves that pLSI is a special case of LDA under a uniform Dirichlet prior.

A topic model's output is typically a group of words. These groups of words, when viewed together, represent *topics*. Table 1.1 gives an example of some topics estimated by an LDA topic model. The sample corpus is the collection of US Federal Open Market Committee transcripts (explained in detail in Chapter 3 and in Hansen et al. (2018)).

productivity	***growth***	***inflation***	***economic weakness***
product	growth	inflat	economi
increas	slow	expect	weak
wage	economi	core	recoveri
price	continu	measur	recess
cost	expans	higher	confid
labor	strong	path	eas
rise	trend	slack	neg
acceler	inflat	gradual	econom
inflat	will	continu	will
pressur	recent	remain	turn

Table 1.1: Example output estimated by a Latent Dirichlet Allocation (LDA) topic model (Hansen et al., 2018, p. 821).

The terms are ranked from top to bottom where the top-most term is given the highest weight for a given topic (ignore the spelling for now, details are explained in Chapter 3). Each word group is given a label or mnemonic which best describes that group. In this instance, the mnemonics represent economic concepts: *productivity*, *growth*, *inflation* and *economic weakness*, which is intuitive given the context of the data.

Topic models are easily accessible, with most software packages made freely available to the public.[1] As such, non-NLP research domains such as higher education (Park, 2020; Bowles and Carlin, 2020) and sociology (Lindstedt, 2019) have been able to benefit from the power of topic models when answering their own questions. Finance is another important application area, and the one that is central to this thesis, as elaborated below.

Topic models have evolved since the (20-year-old) foundational work described above, branching off in many directions. However, to list out all the work published

[1] For example, gensim; https://radimrehurek.com/gensim/ and Octis; https://github.com/MIND-Lab/OCTIS. There are lots more.

to date is not practical and is beyond the scope of this study. Boyd-Graber et al. (2017) provides a neat survey of the applications of topic models in NLP, which covers important developments in the area. In addition, Chapter 2 will discuss the newer topic models that I use in this study, (Dieng et al., 2019b; Yu et al., 2020), which incorporate *word embeddings* (Bengio et al., 2003; Mikolov et al., 2013b). Topic models combined with word embeddings are typically the representations used in modern NLP research. These embedding-enhanced topic models have improved over LDA in terms of standard NLP metrics (Mikolov et al., 2013b; Dieng et al., 2019b,a; Yu et al., 2020).

1.2 What am I doing, and why?

1.2.1 NLP in finance and economics

Focusing on the bigger picture for a moment, this study draws its inspiration from an important field of study where researchers are trying to better understand the *language* of business and finance. It is argued that important and *material* signals about a given firm's future *value* is not only contained in the structured, numeric data but also in its unstructured data — or *text* — that they produce. These unstructured data come in many forms, from interviews with CEOs, news media, management discussion and analyses, analyst calls etc . . . For example, Ahern and Sosyura (2015); Antweiler and Frank (2004); Bodnaruk et al. (2015); Cohen et al. (2020); Hoberg and Phillips (2010); Loughran and McDonald (2011); Tetlock et al. (2008) are only a handful of researchers which tie firm-specific textual content to firm-level stock returns.

Text within the news media and other commentary on the wider economy also contain important signals about the broader financial markets. Fang and Peress (2009); Tetlock (2007, 2011, 2014); Calomiris and Mamaysky (2019a,b) (not an exhaustive list) study the impact of such text data on foreign exchange and aggregate

equity markets.

1.2.2 The why

Turning back to my main focus, in modern central banking communications about their actions and plans, or *forward guidance*, plays an important role in shaping expectations of the future path of interest rates. Especially in times of zero or negative interest rate environments, like those currently seen in Japan, central banks have resorted to alternative tools like forward guidance and large-scale asset purchases to affect monetary policy (Blinder et al., 2008; Bernanke, 2020). These expectations translate into material changes real in economic activity which affect hundreds of millions of people. As such, Bholat et al. (2015), Hansen and McMahon (2016) and Hansen et al. (2018) are doing important work to better understand the language of *central banks* by combining NLP-techniques — *topic models* — with traditional econometric methods. The importance of understanding a key central banker's thought process and how they form policy decisions cannot be stressed enough.

However, the literature on central bank communications and topic models has largely been confined to the application of classical LSA and LDA models. This is the area in which I aim to make a contribution by investigating the use of newer topic models.

1.3 Research design and contributions

Broadly speaking, newer topic models typically improve on the original LDA according to metrics used by NLP researchers. However, non-NLP researchers may have different evaluation criteria, or different tasks in mind when they initially applied LDA. As such, the main purpose of this study is to assess whether newer topic models perform better than LDA when comes to those domain-specific tasks. This thesis,

to an extent, addresses to the research gap noted by Doogan and Buntine (2021) who drill into whether the evaluation metrics proposed within NLP are adequate tools for assessing the true usefulness of topic models in an applied setting. The authors stress the importance in understanding whether newer, state-of-the art models generalise well to specialised collections of text. Here, I use the United Sates Federal Reserve's FOMC meeting transcripts (explained in more detail in Chapter 3) as the analogue to Doogan and Buntine (2021)'s Twitter Data.

I unpack this broad question by structuring this study in the following way. First, I use Hansen et al. (2018)'s work as the basis for my analysis. The reason is because Hansen et al. (2018) was able to econometrically identify a change in the behaviour of certain FOMC members after an unanticipated event related to publicising their raw dialogue during policy deliberations — the *Transparency* event. This was most noticeable in less experienced, or *rookie*, members of the FOMC who changed their topical discourse after the transparency event. I then look at Hansen et al. (2018)'s findings from a different angle where I ask:

Given the topics, can I determine if the speaker is a rookie or not?

I investigate this question by applying newer topic modelling methods developed by Dieng et al. (2019b) and Yu et al. (2020), to the original data, which perform *better* by standard NLP metrics.

I then evaluate those newer models' performance against Hansen et al. (2018)'s original topic model (Blei et al. (2003)'s LDA) based on three criteria.

The first is a quantitative evaluation, where I measure how well each model performs in a bespoke downstream labelling task, which I frame as a document classification task. Second is a qualitative evaluation, where I assess the newer models' topical output against Hansen et al. (2018)'s customised topic-ranking index which, by-and-large, serves as a sanity test of whether a given model's output appears correct

within the context of a business cycle. The third evaluation criteria is a suite of NLP metrics, which measure the quality or *interpretability* of a given topic that was produced by a given model.

I propose to extend the *central bank communications* and *topic modelling* fields of research in the following ways. **First**, research using topic models to directly model central bank communications has not considered models based on current word embedding-based approaches. I extend this branch of literature by applying newer topic models which allow for the embedding space to be estimated jointly with the topics, or allow for the use of pre-trained embeddings. **Second**, I update the topic modelling literature by applying a downstream method for evaluation to the new topic models. Specifically, I assess the performance of each model's ability to classify whether a member of the United States Federal Open Market Committee (FOMC) is a *rookie* or a *veteran*. **Third**, the topic modelling literature typically evaluates the performance and *quality* of topic models on a standard set of corpora, for example, *BBC-News*, *20 Newsgroups*, *New York Times*, *PubMed Central abstracts* etc . . . (Terragni et al., 2021; Wallach et al., 2009). I extend both the central bank communications and topic modelling literature by evaluating the quality of the newer topic models against an economics-focused corpus. I then tie this all together by comparing the topic quality of each model against the results of their classification task.

1.4 Research questions

I crystallise the intersection of the fields of research stated above by framing my study with the following questions:

RQ1 Do the new topic models produce topic rankings in line with Hansen et al. (2018)'s LDA model?

RQ2 Do the new topic models in Dieng et al. (2019b) and Yu et al. (2020)

perform better than the LDA on a bespoke downstream task of predicting if a member of the FOMC is a rookie or a veteran? This is in the same spirit as the key question raised by Doogan and Buntine (2021): *How well will this model work in an applied setting?*[p. 3824]

RQ3 Do the newer topic models produce better quality topics as measured by the metrics found in Lau et al. (2014)?

1.5 What this study does NOT do

- It is *not* an *econometric* study. As such, I will not be conducting econometric tests for significance of any kind. My primary concern is how each model performs *relative* to each other based on a given set of assessment metrics.
- This study does not seek to optimise topic models, such that they maximise a given assessment metric (accuracy or quality etc...). The primary focus is on the set of features which are produced by each model, given a set of default parameters (see Chapter 6 for further details).
- This study only considers *static* or time-invariant topic models. Dynamic or time-varying topic models are out of scope.

The points noted above are opportunities for future work.

1.6 The rest of this document

The rest of this document is set out as follows. Chapter 2 provides background and related work. Chapter 3 gives details on replicating Hansen et al. (2018)'s work as well as a discussion on the underlying data. Chapter 4 briefly outlines the bespoke classification task which forms one of the evaluation criteria for this study. Chapter 5 provides a qualitative evaluation of the competing models of Dieng et al.

(2019b) and Yu et al. (2020) with respect to Hansen et al. (2018) topic-ranking index. Chapter 6 presents the results of the classification task described in Chapter 4 and topic quality based on Lau et al. (2014)'s coherence measures. Chapter 7 provides some concluding commentary and signs off with ideas for future work.

Chapter 2

Background and Literature Review

This chapter gives a summary of related work and is roughly divided into two parts. The first part describes a way of representing *text* called *topic models*, which is the main analytical tool I use in this study. I give brief and broad account of the lineage of the main topic models of interest, leading up to two major topic model frameworks that use the kind of word embedding representations that are the core of modern NLP, and present the main formulas and procedures.[1] I then briefly cover the literature on topic model evaluation. The second part is centred on a body of economics literature that leverages the use of topic models to analyse financial and economic text. My main focus is on the research that applies topic models to central bank communications, particularly Hansen et al. (2018)'s work which forms the core of this study.

2.1 Topic models

At its core, a topic model's primary task is to uncover patterns from a collection of *words*, the mechanisms of which are grounded in statistics and probability theory. And while topic models are fundamentally applied to words, they can be extended to other areas, like images and genes (He et al., 2014; Pan et al., 2015; Chen et al., 2010; La Rosa et al., 2015; Liu et al., 2016). As long as a collection is encoded

[1] Given space limitations, I will deliberately skip mathematical proofs and other technical aspects of each model and will point the reader to relevant details when needed.

into an appropriate digital format, then a computer can be tasked to uncover its topical structure. *Fast*. As such, this architecture is well suited for processing large collections of digital text – or *corpora*. In particular, topic models are useful for those who are NLP tasks such as organising, summarising, visualising, searching, predicting and exploring large collections of digital text documents, but do not have the time to it manually themselves.

2.1.1 Latent Semantic Analysis

An early topic model is called Latent Semantic Analysis (LSA) pioneered by Deerwester et al. (1990) whose core motivation was document retrieval. LSA is a generalised principal components analysis (PCA), where the components – or *topics* – are extracted using a singular value decomposition (SVD). SVD is fundamental to LSA because it is what maps the relationships between words across documents, taking into consideration those words that *are* and *are not* used (Hendry, 2012).

2.1.1.1 Singular Value Decomposition SVD

Consider a rectangular *term-document* matrix $M_{t\times d}$, which can be described as having t rows representing unique *terms* and d columns of *documents*. SVD implies that M can be represented by three characteristic matrices U, S and V^T. LSA, however, does not use the full decomposition, but rather, it takes the first k columns (where k is typically set to an amount much less than d for a given sample corpus) of U and V and the $k \times k$ upper-left matrix of S. The approximated matrix , $\hat{M} = U_k S_k V_k^T$ is said to be the least-squares best fit of M. Figure 2.1 provides a stylised picture of SVD. This *reduction* removes noise or other unwanted information (Hendry, 2012) and leaves important underlying semantic structure of the words and documents. Each element in $\hat{M}$ is some numeric value, which can loosely be interpreted as the *importance* of a given term for a given topic. Those top-ranking terms for each column, is can loosely be interpreted as the underlying *theme* representing the k^{th}

topic. The value k (the number of columns) can be thought of as the *number* of topics for a given corpus and the order of the columns reflects the importance of the k^{th} topic in that corpus.

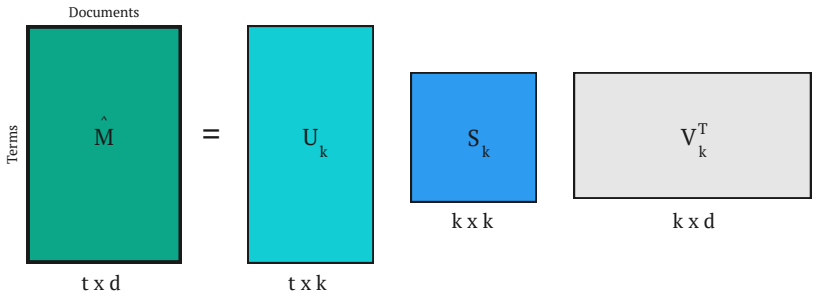


Figure 2.1: A stylised example of SVD.

One shortfall of LSA, however, is that it has difficulty in dealing with polysemous terms, or terms that have different meaning if they are used in different contexts. For example, the term *interest* could refer to a person's curiosity about something, or it could relate to a rate of return on some dollar amount. Instead of keeping keeping these different meanings separate, LSA combines them. In the language of linear algebra, these two senses are projected onto one line (Deerwester et al., 1990).

2.1.2 Probabilistic Latent Semantic Indexing

Hofmann (1999) introduces Probabilistic Latent Semantic Indexing (pLSI) which presents a generative model for data which is underpinned by more formal statistical theory. It is a model which associates an unobserved class variable $z \in Z = \{z_1, \ldots, z_k\}$ with each occurrence of a word $w \in W = \{w_1, \ldots, w_m\}$ in a document $d \in D = \{d_1, \ldots, d_n\}$. The pLSI assumes that documents are a mix of k latent aspects or *topics* and is called a mixture model, where k is strictly less than the number of documents n. The generative process for data is as follows:

- select a document d with probability $P(d)$,
- draw a latent class z with probability $P(z|d)$,
- generate a word w with probability $P(w|z)$.

It is assumed that the observation pair (d, w) is generated independently, that is, from a *bag-of-words*. Another assumption is, conditional on the latent class z, words are generated independently of a specific document d.

Therefore, the joint probability for the generative model can expressed by:

$$P(d, w) = \sum_{z \in Z} P(z)P(w|z)P(d|z) \tag{2.1}$$

Using Bayes' Theorem, Equation 2.1 is written as such because we observe the words and documents, but not the topics. The log-likelihood function is:

$$L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w) \tag{2.2}$$

The posterior distribution is intractable, and is approximated via Dempster et al. (1977)'s Expectation Maximisation (EM) algorithm. The EM procedure is set out in Hofmann (1999)[p.51] and maximises the expected log-likelihood objective function $E[L]$ in Equation 2.2.

2.1.3 Latent Dirichlet Allocation

Blei et al. (2003)'s Latent Dirichlet Allocation (LDA) addresses the shortcomings of Hofmann (1999)'s pLSI. It is a generative model for data at the *document* level which is grounded in Bayesian statistics. Like the pLSI, LDA is a mixture model, where documents are represented as a mix of topic proportions. However, unlike the pLSI which treats these topic proportions as some fixed list of quantities, the LDA assumes these topic proportions are generated from a Dirichlet prior which is conditioned on a hyperparameter α. In fact, Girolami and Kabán (2003) proves that

pLSI is a *maximum a posteriori* estimator of LDA under a uniform Dirichlet prior.

More formally, assume a corpus of D documents, that contains V distinct words. Let $w_{d\in\{1,\ldots,D\},n\in\{1,\ldots,V\}}$ be the n^{th} word of the d^{th} document. There are K topics in a document. Each topic $k \in \{1,\ldots,K\}$ is assigned weight β_k and together form some distribution over a vocabulary and is some distribution over a vocabulary. Next, there are topic proportions (θ_d) assigned to each document d, where $\theta_{d,k}$ is the extent to which the k^{th} topic is expressed in document d.

LDA's generative process assumes each word is assigned to topic k with a probability $\theta_{d,k}$, and is then drawn from the distribution β_k. The process for each document is shown below:

1. Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For each word n, in document d:
   (a) Draw topic assignment $z_{d,n} \sim \text{Cat}(\theta_d)$.
   (b) Draw word $w_{d,n} \sim \text{Cat}(\beta_{z_{d,n}})$.

Where, $\text{Cat}(\cdot)$ is the categorical distribution. LDA places a Dirichlet prior on the topics, $\beta_k \sim \text{Dirichlet}(\eta)$, where α and η are fixed model hyperparameters, and are concentration parameters of the Dirichlet distribution.[2] The generative process for LDA is characterised by the following joint distribution over the corpus, given the latent (hidden) topics (β's), topic assignments (z's) and topic proportions (θ's):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}|\alpha,\eta) = \prod_{i=1}^{K} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \quad (2.3)$$

Like the pLSI, the posterior distribution is intractable and must be approximated.

[2] A higher value for α and/or η leads to a more uniform spread of probabilities across documents and across the vocabulary.

Here is a non-exhaustive list of methods for approximating the posterior and in no particular order.

- Mean field variational methods (Blei et al., 2003)
- Expectation propagation (Minka, 2001)
- Markov Chain Monte Carlo (Griffiths and Steyvers (2004), and is the method used in Hansen et al. (2018)).[3]
- Amortized inference (Srivastava and Sutton, 2018)

2.1.4 Topic models and word embeddings

Word embeddings (Bengio et al., 2003) are said to produce accurate syntactic and semantic word relationships (Mikolov et al., 2013b). Word embeddings represent words as vectors in $\mathbb{R}^n$. These vectors can be thought of as a word's context or *meaning*. As such, words that have similar meaning are characterised by their corresponding vectors being relatively *close* together in $\mathbb{R}^n$.

With this in mind, Dieng et al. (2019b) combines LDA and word embeddings to make the *embedded topic model* (ETM).

The ETM model assumes a *continuous bag-of-words* (CBOW) (Mikolov et al., 2013a). The likelihood of each word w_{dn} is:

$$w_{dn} \sim \text{softmax}(\rho^\top \alpha_{dn}). \tag{2.4}$$

The embedding matrix ρ is an $L \times V$ matrix whose columns contain the embedding representations of the vocabulary, $\rho_v \in \mathbb{R}^L$. The vector α_{dn} is the *context embedding*, which is the sum of the context embedding vectors (α_v for each word v) of the words surrounding w_{dn}.

[3] For more detail, see Hansen's Online Technical Appendix in The Quarterly Journal of Economics.

In the ETM, the k^{th} topic is a vector $\alpha_k \in \mathbb{R}^L$ in the embedding space. α_k is called a *topic embedding* and is a distributed representation of the k^{th} topic in the semantic space of words. The ETM uses α_k to form a per-topic distribution over the vocabulary. It then assigns high probability to a word υ in topic k by measuring the agreement between the word's embedding and the topic's embedding (Dieng et al., 2019b, p.4). Denote the $L \times V$ word embedding matrix by ρ; the column ρ_υ is the embedding of υ. The generative process for the d^{th} document under ETM is then:

1. Draw topic proportions $\theta_d \sim \mathcal{LN}(0, I)$.
2. For each word n, in document d:
   (a) Draw topic assignment $z_{d,n} \sim \text{Cat}(\theta_d)$.
   (b) Draw word $w_{d,n} \sim \text{softmax}(\rho^\top \alpha_{d,n})$.

$\mathcal{LN}(\cdot)$ is a logistic-normal distribution; step 2(a) is the same as in LDA; step 2(b) refers to the likelihood under the CBOW variant mentioned earlier (see Dieng et al. (2019b) for specific details).

2.1.5 Category-Name Guided Text Embedding

Similar to the ETM, Yu et al. (2020)'s Category-Name Guided Text Embedding (CatE) incorporates word embeddings when estimating topics. The difference, however, is that the CatE gives a user the opportunity to guide the topic discovery process by providing an initial set of *categories* or topics based on their interest or prior knowledge about a given corpus. CatE is also fundamentally different to the LDA and ETM in that it discovers topics in a (weakly) supervised way, while LDA and ETM are unsupervised.

Essentially, CatE performs what the authors call a *discriminative topic mining* task, where a user provides a set of category names, $\mathscr{C} = \{c_1, ...c_n\}$, for a given corpus $\mathscr{D}$. The search then aims to retrieve a set of *representative* terms $\mathscr{S}_i = \{w_1, ..., w_m\}$ from

$\mathscr{D}$ for each category c_i such that each term in $\mathscr{S}_i$ semantically belongs to category c_i *only*. That is, terms in $\mathscr{S}_i$ cannot belong to any other category j where $j \neq i$.

The process under user providing n category names follows three steps:

1. a document d is generated conditioned on one of the n categories (this can be thought of the as a *topic assignment*);
2. each word w_i is generated conditioned on the semantics of document d (this defines the *global context*);
3. surrounding words w_{i+j} in a local context window ($-h \leq j \leq h, j \neq 0, h$) of w_i are generated conditioned on the semantics of word w_i (this defines the *local contexts*).

The probability of the underlying generative process for a corpus, conditional on a user's categories of interest, $\mathscr{C}$, is as follows:

$$P(\mathscr{D}|\mathscr{C}) = \prod_{d \in \mathscr{D}} p(d|c_d) \prod_{w_i \in d} p(w_i|d) \prod_{\substack{w_{i+j} \in d \\ -h \leq j \leq h, j \neq 0}} p(w_{i+j}|w_i), \tag{2.5}$$

where c_d is the latent category of document d.

Taking the negative log-likelihood as the objective $\mathscr{L}$ and is expressed as:

$$\mathscr{L} = -\sum_{c \in \mathscr{C}} \sum_{w \in c} p(c|w) + B - \sum_{d \in \mathscr{D}} \sum_{w_i \in d} \log p(w_i|d) - \sum_{d \in \mathscr{D}} \sum_{w_i \in d} \sum_{\substack{w_{i+j} \in d \\ -h \leq j \leq h, j \neq 0}} \log p(w_{i+j}|w_i), \tag{2.6}$$

where B in the first term is a constant (Yu et al., 2020, p.3).

Next let $\mathbf{u}_w$ be the word embedding for w and let the embedding cosine similarity between w and category c_i be $\cos(\mathbf{u}_w, c_i)$. Let $\kappa_w \geq 0$ be a value of Word Distributional Specificity (WDS). The larger the κ_w the less polysemous word w becomes,

that is, w converges to a more *specific* meaning.[4] Then the selection criteria for determining a representative term is defined by the following:

$$w = \arg_w \min rank_{sim}(w, c_i) \cdot rank_{spec}(w) \quad \text{s.t. } w \notin \mathscr{S} \text{ and } \kappa_w > \kappa_{c_i}, \tag{2.7}$$

where $rank_{sim}(w, c_i)$ is the ranking of w by $\cos(\mathbf{u}_w, c_i)$ (highest to lowest); $rank_{spec}(w)$ is the ranking by of w by WDS (from lowest to highest).

The procedure for retrieving representative terms given the initial category names is outlined by the following algorithm (Yu et al., 2020, p.5):

Algorithm 1: Discrminative Topic Mining

Input: A text corpus $\mathscr{D}$; a set of category names $\mathscr{C}\{c_i\}|_{i=1}^n$;
Output: Discriminative topic mining results $\mathscr{S}_i|_{i=1}^n$;
for $i \leftarrow 1$ *to* n **do**
  $\mathscr{S}_i \leftarrow \{c_i\}$; `/* initialise` $\mathscr{S}_i$ `with category names */`
for $t \leftarrow 1$ *to max_iter* **do**
  Train $\mathscr{W}, \mathscr{C}$ on $\mathscr{D}$; `/* according to Equation 2.6 */`
  for $i \leftarrow 1$ *to* n **do**
    $w \leftarrow$ Select representative word of c_i ; `/* by Equation 2.7 */`
    $\mathscr{S}_i \leftarrow \mathscr{S}_i \cup \{w\}$
  for $i \leftarrow 1$ *to* n **do**
    $\mathscr{S}_i \leftarrow \mathscr{S}_i \backslash \{c_i\}$; `/* exclude category names */`
Return $\mathscr{S}_i|_{i=1}^n$

The topic models outlined so far are *static* or non-time varying. That is, these models do not consider topic evolution across time. As such, time-varying or *dynamic* topic models are a natural extension to static topic models. Dynamic topic models, while interesting, are beyond the scope of this thesis. See Blei and Lafferty (2006) for some background.

[4] See Section 3.3 in Yu et al. (2020) for more detail on WDS and proofs.

2.1.6 Evaluating topic quality

One difficult hurdle facing topic models can be summarised by the following scenario. Imagine a dialogue between a group of NLP researchers which goes like this:

> **Researcher 1**: *So, do these topics make sense to you?*
> **Researcher 2**: *Yes*.
> **Researcher 3**: *No*.
> **Researcher 4**: *Maybe*.
> **Researcher 5**: *What do you mean by 'makes sense'?*

A given topic model's output, in the context of natural language processing, is typically some group of words that are semantically linked (according to that model). But, whether these groups of words *make sense* to a person, is based on the judgement by that person. It is not guaranteed that judgments will align across different people. So, how does a researcher make such an assessment of how *good* a topic is?

Human judgements, however, are *still* the gold standard, despite the issue highlighted in the pretend scenario above. Chang et al. (2009)'s word intrusion task presents a robust framework in which humans their objective judgements. The task is described as follows. The subject (a person) is presented with a set of randomly ordered words. The subject's task is to find the word (the intruder) that does not belong in that group. The *goodness* of a topic model is measured by how easily an intruder word is identified.

In terms of topic quality, the earliest measures were:

- **perplexity** (Shannon, 1948), a monotonic function of the estimated probability of a text, where the probability comes from the learned topic model,
- and **human judgements** (Chang et al., 2009) as noted earlier.

On one hand, it is not obvious if perplexity corresponded to anything humans care about and on the other, human judgements are expensive to get (Chang et al., 2009;

Röder et al., 2015). However, Newman et al. (2010) took a significant leap forward with their proposal to measure topic quality through coherence or interpretability of LDA-based topic models. A more computationally useful approach came as an extension of this, with a proposal to use metrics (for example, pairwise mutual information) that align with human judgements, and an experimental evaluation that measured Pearson correlations between metrics and human judgements (Lau et al., 2014).

In this thesis, I will use Lau et al. (2014)'s automated framework to evaluate the quality of topics produced by the different models used in this study. The three coherence measures that are used to evaluate the estimated topics (t) are:

- **Pairwise PMI** (pmi) of top-N topic words:

$$pmi(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)P(w_j)} \tag{2.8}$$

- **npmi** a variant of 2.8

$$npmi(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_j, w_i)}{P(w_i)P(w_j)}}{-\log P(w_j, w_i)} \tag{2.9}$$

- **Pairwise log conditional probability** (lcp) of top-N topic words (Mimno et al., 2011).

$$lcp(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)} \tag{2.10}$$

Pmi and npmi metrics are essentially measures of association based on the number of times a word-pair (w_i,w_j; $i \neq j$) appears in some sliding window. The lcp measure is only slightly different to pmi and npmi, in that its sample word counts come from the training corpus itself as opposed to an external source (Lau et al., 2013). A higher value means *better*.

2.2 Topic models and central bank text

Central bank communication is an important medium for implementing monetary policy, particularly for developed economies who target inflation through interest rates (Blinder et al., 2008; Blanchard and Sheen, 2013). These communications, whether delivered through speeches, press conferences, meeting minutes, transcripts etc... all shape capital market participants' (such as banks and other large financial institutions) expectations of *future* interest rates; the active use of communications by the central bank to shape expectations is referred to as *forward guidance*. As such, teams of highly-paid economists and analysts (armed with their tarot cards, crystal balls and chicken bones) are tasked with deciphering every clue and clues of clues contained within such communications to try and predict what a given central bank will do at its next policy meeting.

While all the manner of economic witch-doctory is popular for some, a handful of researchers take a more sensible approach. They use technology, like a *computer*, and more importantly, apply solid statistical theory to aid in their quest to analyse these cryptic central bank hieroglyphs. With respect to their methods, these researchers fall into two broad groups. One: those that use LSA-based methods, and two: those that use LDA-based methods.

2.2.1 LSA-based methods

Hendry and Madeley (2010) and Hendry (2012) apply LSA to extract topics contained in the Bank of Canada's (BoC) Monetary Policy reports, press conferences and media news, and tries to identify which topics affect other macroeconomic variables. Boukus and Rosenberg (2006) do something similar, but use the United States Federal Open Market Committee (FOMC) policy meeting *minutes*. FOMC minutes are not to be confused with the FOMC *transcripts* which are the core documents that I use in this study. The FOMC minutes provide only a summary of

what was discussed among policy members and are made public not long after the policy meeting. FOMC transcripts, on the other hand, are documents containing the *raw* dialogue spoken among policy members. These documents are released five years after the fact. More detail is provided in Chapter 3. In addition to FOMC minutes, Acosta (2015) applies LSA to FOMC transcripts and investigates how well the minutes reflect what was actually discussed during a given meeting.

2.2.2 LDA-based methods

Moniz and de Jong (2014) use a mix of supervised and unsupervised (LDA) methods to analyse the Bank of England's (BoE) Monetary Policy Committee Minutes to predict market participants' interest rate expectations. In a closely related study Hansen et al. (2019) model the BoE's Inflation Report to gauge long run expectations. Hansen and McMahon (2016) in a similar trajectory to Boukus and Rosenberg (2006), Hendry and Madeley (2010) and Hendry (2012) use LDA in conjunction with econometric methods. They estimate a 15-topic model and extract the *tone* of these topics. They find that the tone of topics related to future expectations about inflation, growth and unemployment have a significant impact on economic variables like short- and long-term bond yields, market volatility (VIX), equity market indexes and some commodity prices.

Fligstein et al. (2014) analyse FOMC transcripts using LDA (with other methods) and argue that policy members were unable to connect discussions related to the financial crisis together — they failed to *connect the proverbial dots*. To quote the authors:

> "The fact that the group of experts whose job it is to make sense of the direction of the economy were more or less blinded by their assumptions about how that reality works, is a sobering result. Fligstein et al. (2014)[p.46]"

Like Acosta (2015), Hansen et al. (2018) is interested in transparency — a decision to publicise the raw transcripts of what was discussed during FOMC policy meetings. Transparency is an important aspect of forward guidance; however, Hansen et al. (2018) observe that there are both positive and negative sides to it. They describe the *discipline effect*, where transparency increases accountability and induces policy makers to "work harder and behave better"; on the other hand, they describe the *conformity effect*, where, out of career concerns (Holmström, 1999), board members might limit their speech, use less dissenting language, etc (Meade and Stasavage, 2008).

To investigate this, Hansen et al. (2018) formulates a 'natural experiment' based on the fact that board members from before 1993 did not know that FOMC transcripts were going to be released, while those after did (see Section 3.1.1). Hansen et al. (2018) then uses econometric models,[5] including LDA topics as variables to represent the content of transcripts, and found both the discipline and conformity effects in play. Specifically, using members' experience in policy making as a proxy for career concerns, members were divided into the relatively less experienced — the *rookies* — and their other more experienced peers — the *veterans*. Hansen et al. (2018)'s econometric analysis found that their behaviours differed significantly: in the part of the transcripts that deal with economic situation discussion, the rookies speak more quantitatively and on a wider range of issues (the discipline effect), but in the policy strategic discussion, they limit their speech (the conformity effect). This is an important finding, not only about human nature in general, but also about the specific context of setting policy in central banks — this select group of FOMC members are responsible for making policy decisions on behalf of the largest economy in the world.

[5] For those interested these standard econometric tests include fixed-effects panel regression and difference-in-difference regressions.

This work by Hansen et al. (2018) is the starting point for the application of topic models in this thesis. Where that work characterised its use of LDA as a methodological contribution in the economics space, this thesis investigates more recent topic models from NLP.

While dynamic methods are out of scope for this study, it is still worth a brief mention.[6] More recent studies use dynamic approaches to model the variation of topics over time, but do not go beyond concluding that the *topics which central banks talk about or focus on, may (or may not) change over time* (Cross and Greene, 2020; Windsor, 2021).

2.2.3 Making sense of a central bank's two cents

The spectre of topic quality and interpretability still looms, however. And it is made worse by the very nature of central bank communications itself, which is rife with esoteric jargon. Even policy makers, who utter those very words, would find it hard to understand themselves (Blinder et al., 2002).

Of the authors in the central bank literature cited thus far, only Hansen et al. (2018) and Fligstein et al. (2014) make a serious attempt at using some method for setting parameters such that their models yielded good quality topics. The former used trial and error and their judgement, while the latter use mix of trial and error, their judgement and a coherence measure Mimno et al. (2011). The other authors typically only mention that topic quality is based on subjective judgement, but rarely apply their own.[7] Cross and Greene (2020) did not even bother to address the issue (as interpretability was not their focus).

[6] No more than *two* cents worth, though.

[7] I am also guilty of this, though my purpose here is not to fine tune the models to yield the best topics.

2.3 Summary

This chapter briefly covers the important statistical methods that I use in this study, namely LDA, ETM and CatE. I give a high-level view of important mathematical results (ex-proofs) and the inference problems. I then give a short outline of the related literature in terms of the contextual domain, i.e. central bank communications. I also briefly mention dynamic methods, though these are out of scope in this study. The next chapter details on replicating Hansen et al. (2018)'s study, which forms the core of this thesis.

Chapter 3

Replicating Hansen et al. 2018

In Chapter 2, I highlighted the significance of Hansen et al. (2018)'s comprehensive study on FOMC deliberations. The authors investigate how important economic policy decisions (which affect a lot of people) are made by using a solid framework that combines NLP techniques with robust econometric methods. This is why I choose Hansen et al. (2018)'s paper as the core foundation of this thesis. This chapter aims to set up the subsequent analyses in later chapters which primarily revolve around a bespoke classification task which is discussed in Chapter 4.

I treat Hansen et al. (2018)'s topic model and results as the baseline or — *gold standard* — for comparison against other topic models discussed in later chapters. However, because intermediate topic model data from Hansen et al. (2018) is not available — in particular, the original LDA *document-topic-proportions* — and this is necessary for the comparative evaluation in this thesis, I had to replicate Hansen et al. (2018)'s model in order to estimate and store these data myself. I will discuss the classification task in more detail in later chapters. In this chapter, I provide the following:

- a description and some background on the data;
- a detailed description of my replication method;[1] and

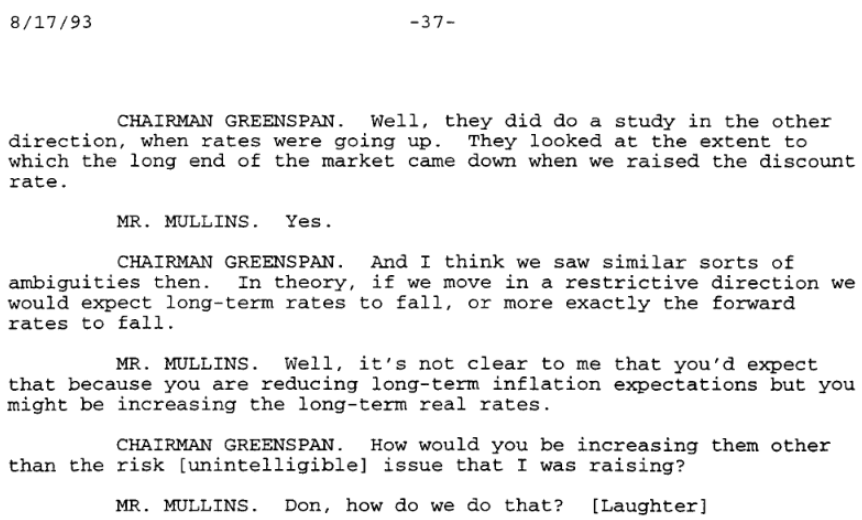
[1] Reading the original paper is also an alternative, but I thought it more convenient to articulate the process here.

- a detailed comparison between my replicated output and Hansen's original output to show that my replicated LDA model is a valid proxy to the *gold standard* benchmark that is used in the subsequent analysis in later chapters.

3.1 Data

3.1.1 Background: a natural experiment

Hansen et al. (2018) analyses the United States Federal Open Market Committee (FOMC) raw meeting transcripts.[2] These transcripts are important documents because they allow us to see the internal deliberations, discussions and debates among key FOMC policy makers, when deciding monetary policy settings for the largest economy in the world. Moreover, these transcripts give us insights on the thought process of those FOMC members during that time. Figure 3.1 shows an excerpt of the dialogue between Fed Chair Greenspan and Governor Mullins during the FOMC policy meeting held on August 17th, 1993.



8/17/93 -37-

CHAIRMAN GREENSPAN. Well, they did do a study in the other direction, when rates were going up. They looked at the extent to which the long end of the market came down when we raised the discount rate.

MR. MULLINS. Yes.

CHAIRMAN GREENSPAN. And I think we saw similar sorts of ambiguities then. In theory, if we move in a restrictive direction we would expect long-term rates to fall, or more exactly the forward rates to fall.

MR. MULLINS. Well, it's not clear to me that you'd expect that because you are reducing long-term inflation expectations but you might be increasing the long-term real rates.

CHAIRMAN GREENSPAN. How would you be increasing them other than the risk [unintelligible] issue that I was raising?

MR. MULLINS. Don, how do we do that? [Laughter]

Figure 3.1: A sample extract from the August 17th, 1993 policy meeting transcript [pp 37]. The historical documents can be found at: https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm

[2]Technically, these transcripts were *near raw* as they were partly edited from October 1993 to assist with readability.

Since the 1970s, deliberations and discussions during FOMC meetings were held behind closed doors, but were tape-recorded in order to assist with preparing meeting minutes. The committee members at the time *knew* that the tapes were erased thereafter. However, in October 1993 it was discovered that the tape recordings were *transcribed* and *archived* (prior to being erased), unbeknown to those same committee members. Following this discovery, it was decided that the Federal Reserve release these documents to the public. This meant that all historical transcripts prior to October 1993 would be released, as well as all of the ***future*** transcripts going forward — albeit with a five-year lag.[3]

It was this decision — to release all future transcripts — that a 'natural experiment' arises (Meade and Stasavage, 2008), because we can observe two distinct periods of time, where the serving committee members **did** and **did not** believe their deliberations would be made public (Hansen et al., 2018, p. 804). As a result committee members, particularly those who were *less* experienced relative to their more experienced counterparts, changed their behaviour in subsequent meetings (Meade and Stasavage, 2008; Hansen et al., 2018). Hansen et al. (2018) labels these less experienced members as ***rookies***. A core part of the econometric analysis of Hansen et al. (2018) concerns identifying the effects of this change in communicative behaviour of rookies versus others. I will discuss the significance of this for this thesis later in Chapter 4.

3.1.2 Preparing the FOMC data

I present Hansen et al. (2018)'s processing method in the steps below and show a side-by-side comparison of my replicated output with his original figures throughout.

Step (1) Identify 1) bi-grams and 2) tri-grams. This is to identify a sequence of

[3] Lindsey (2003) provides more detail regarding the change in the FOMC transparency arrangements.

words with a particular meaning. For example, the bi-gram *unemployment rate* represent a single economic concept. To identify collocations, words are tokenized and part-of-speech tagged (Toutanova et al., 2003), then specific combinations, such as adjective-noun or noun-noun-noun, are counted (Justeson and Katz, 1995).[4] Bi-grams (tri-grams) which occur more than 100 (50) times are retained. Then the identified bi- and tri-grams are joined to create a single token. To use the previous example, *unemployment rate* becomes *unemploymentrate*.

Step (2) Remove stop words. These are common tokens such as *the*, *a* and *of*, which add little meaning or context.[5]

Step (3) Stem tokens to produce a token's linguistic root. These stems need not be an English word. For example, *continued, continuing* and *continue* is stemmed to *continu* and is counted as a single stem.[6]

Step (4) Compute the *term frequency-inverse document frequency* (tf-idf) for each stem. The tf-idf measure is calculated in the following way:

i. Let n_v be the number of times stem v occurs in the corpus. The term-frequency $tf_v = 1 + \log(n_v)$.

ii. Let D be the number of documents in the corpus. D_v is the number of documents in which term v appears. The document frequency is $df_v = \log(\frac{D}{D_v})$.

iii. The combined measure is tf-idf $= tf_v \times df_v$

Step (5) Rank the stems in descending order by the tf-idf weight and drop those

[4]The specific combinations are: adjective-noun; noun-noun; adjective-adjective-noun; adjective-noun-noun; noun-adjective-noun; noun-noun-noun and noun-preposition-noun.

[5]The stop word list is found here: http://snowball.tartarus.org/algorithms/english/stop.txt

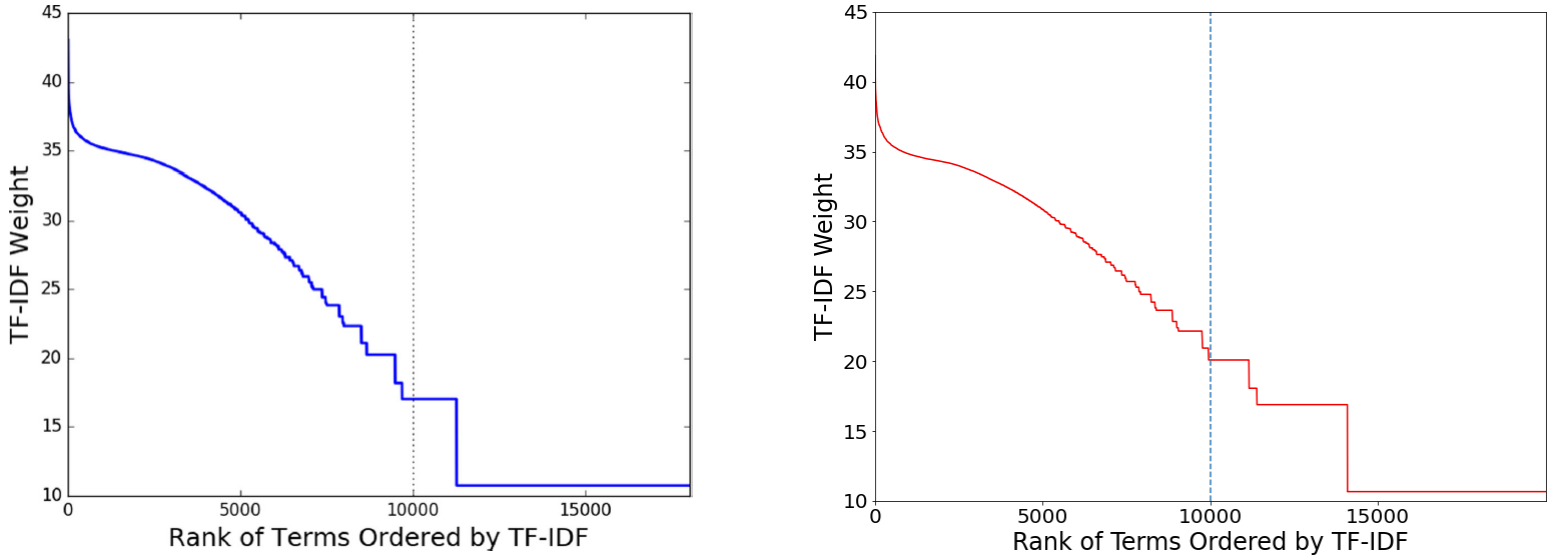
[6]The Porter stemmer method is applied using Python's NLTK package.

stems that rank below 10,000[7] and remove those stems that appear in two or fewer documents.

Step (6) Remove the stem ***think***. It appears in many documents and is effectively a stop word (Hansen et al., 2018, p.819).

After completing the cleaning process described above, the final number of unique stems becomes **9,651** across **42,183** documents. Figure 3.1.2 gives a side-by-side comparison between Hansen et al. (2018)'s original figure (Panel (a); left) and my replicated figure (Panel (b); right). While not exact, I was able to reproduce a chart that closely resembles the original.

[7] Hansen et al. (2018) p.818 indicates that the cut-off rank is 9,000. However, the dashed line on his original chart depicts a cut-off rank of 10,000. I chose to follow the *original* chart, using the 10,000 cut-off rank for my replication. Nevertheless, regardless of what it was meant to be, the cut-off rank is just an arbitrary choice based on visually inspecting the chart.



(a) **Original**

(b) **Replication**

Figure 3.2: Side-by-side comparison between Hansen et al. (2018)'s original TF-IDF figure and my replication. Panel (a) shows the original figure (blue); see (Hansen et al., 2018, p 819). Panel (b) shows the replicated figure (red).

3.2 Fitting the LDA Model

After completing the text pre-processing steps above, I fit Blei et al. (2003)'s popular Latent Dirichlet Allocation (LDA) model to the whole sample. I use Python's *lda* module, which is located in the Python Package Index.[8]

3.2.1 Model setup: hyperparameter settings

For the hyperparameter settings, I follow Hansen et al. (2018):

- The number of topics, k, is set to **40**. Hansen et al. (2018) highlights the difficulties of setting an appropriate value for k because of the trade-off between a better model fit (a higher k) and topic interpretability (a lower k) (Chang et al., 2009). In this instance Hansen et al. (2018) takes a pragmatic approach and favours interpretability.
- I set $\alpha = \frac{50}{k}$ as recommended by Griffiths and Steyvers (2004). This value controls the mixture of topics within a given document. A higher (lower) value results in a higher (lower) mixture of topics in a document.
- I set $\eta = 0.025$ s recommended by Griffiths and Steyvers (2004). A lower value results in topics that feature word distributions that have fewer highly-weighted words.

3.3 Replication results

My replicated topics broadly align with Hansen et al. (2018)'s original output. Importantly, I emphasise the following points. First, my replicated topics yielded word distributions that made sense overall. Second, the hyperparameter settings made sure that each topic had a fewer number of words with large probabilities

[8] Hansen used his own LDA package to estimate the model in the original paper (see here). However, after subsequent discussions with Hansen, I was advised to use the LDA package in the Python Package Index.

relative to the number of words with low probabilities. Third, as Hansen et al. (2018) points out, the topics themselves exhibit a certain degree of **logical** labelling. Table 3.1 below gives a few examples of replicated term-topic distributions for illustration. **Topic 23**: *productivity and growth*, **Topic 24**: *fiscal policy* and **Topic 26**: *recession*.[9] I will present a figure showing all of the replicated 40 term-topic distributions later in this chapter.

Examples of replicated topics

Topic 23	Topic 24	Topic 26
product	tax	financi
increas	govern	economi
wage	effect	recess
growth	fiscal	recoveri
cost	state	financialmarket

Table 3.1: Shows the top five terms for a sample of replicated topics. Topic 23 relates to *productivity*, Topic 24 relates to *fiscal policy* and Topic 26 is *recession*.

3.3.1 Calculating the Procyclicality Index

Next, I check to see if I can reproduce a topic ordering that is consistent with a ranking system based on what Hansen et al. (2018) calls a *Procyclicality Index* (PI). The PI is a way to frame the estimated topics in the context of an economic cycle. Economic cycles are defined by periods of *growth* or *boom* and *recession* or *bust*. Topics with a PI greater (less than) than zero are those topics that feature more prominently in FOMC policy discussions during times of boom (bust). These are called procyclic (countercyclic) topics.

The PI is computed in the following way:

Step (i) Label all FOMC documents in the full sample that correspond to **recessions** and group them together.[10] Those FOMC documents that do not

[9] I avoided using the same examples Hansen et al. (2018) gave originally.

[10] Recessions in the United States are as defined by the National Bureau of Economic Research

correspond to recessions are assumed to be in economic *expansion*, *growth*, *boom* (or, whichever term is defined as the opposite of recession). Group the non-recession documents together.

Step (ii) Compute the average document-topic-proportion for each topic over those documents corresponding to non-recessions.

Step (iii) Compute the average document-topic-proportion for each topic over those documents corresponding to recessions.

Step (iv) Then for each topic, subtract the number derived in **Step (iii)** from the number derived in **Step (ii)**. This difference is called the PI. The number of PIs should be equal to the number of topics (k).

Step (v) Rank the topics in descending order according to the PI computed in **Step (iv)**. A positive (negative) index suggests that a given topic is relatively more prominent during economic expansions (recessions).

As Hansen et al. (2018) discovers, sorting the estimated topics in the way described in **Step (v)** serves as a sanity check to help verify that these topics actually have *meaning* given their context. The contextual domain in this instance is made up of two elements. First, the subject matter is focused on *monetary policy in the United States* and all other related economic issues. Second, the setting in which the subject matter is discussed is strictly within the bounds of FOMC policy meetings.

At first glance, the first element described above may appear broad or wide-ranging. One could argue that FOMC members could speak about *any* topic if they wished. However, central bank policy makers in modern inflation-targeting countries, usually have a fixed set of issues which they focus on when forming their policy decisions. For example, the topics of *inflation*, *growth* and *unemployment* are standard for policy discussion. Any first- or second-year undergraduate textbook in macroeconomics

(see https://fred.stlouisfed.org/series/USREC).

will provide all that is needed on what drives monetary policy. Blanchard and Sheen (2013)'s *Macroeconomics* is a good starting point. This along with the structure of the policy meetings (the second element) act as a natural constraint on what is discussed.[11]

Indeed, Hansen et al. (2018)'s original PI produced a topic ranking that reflects the correct intuition.

For example, topics related to *productivity* and *inflation* (the most procyclical topics) were ranked at the top, while topics related to economic weakness and policy stimulus (the most countercyclical topics) were ranked at the bottom. I treat these original topic rankings as the ***gold standard*** from which I compare my replicated topic rankings. The next sections present a detailed comparison between my replicated output and the original output.

3.3.2 Comparing original output and replicated output

3.3.2.1 A visual inspection

Figure 3.3 shows a comparison between Hansen et al. (2018)'s original output (Panel (a); top) and my replicated output (Panel (b); bottom). Topics are ordered from top to bottom according to the PI measure described above. Each row represents an estimated topic as a heat map. Each term for a given row is ordered from left to right according to their estimated probabilities, with the left-most term being the darkest shade. I limited each topic to its top-five terms to help with space and visuals.

[11] Hansen et al. (2018) devotes a section of the paper (pp.811-813) on how the FOMC conducted their policy discussions under Alan Greenspan.

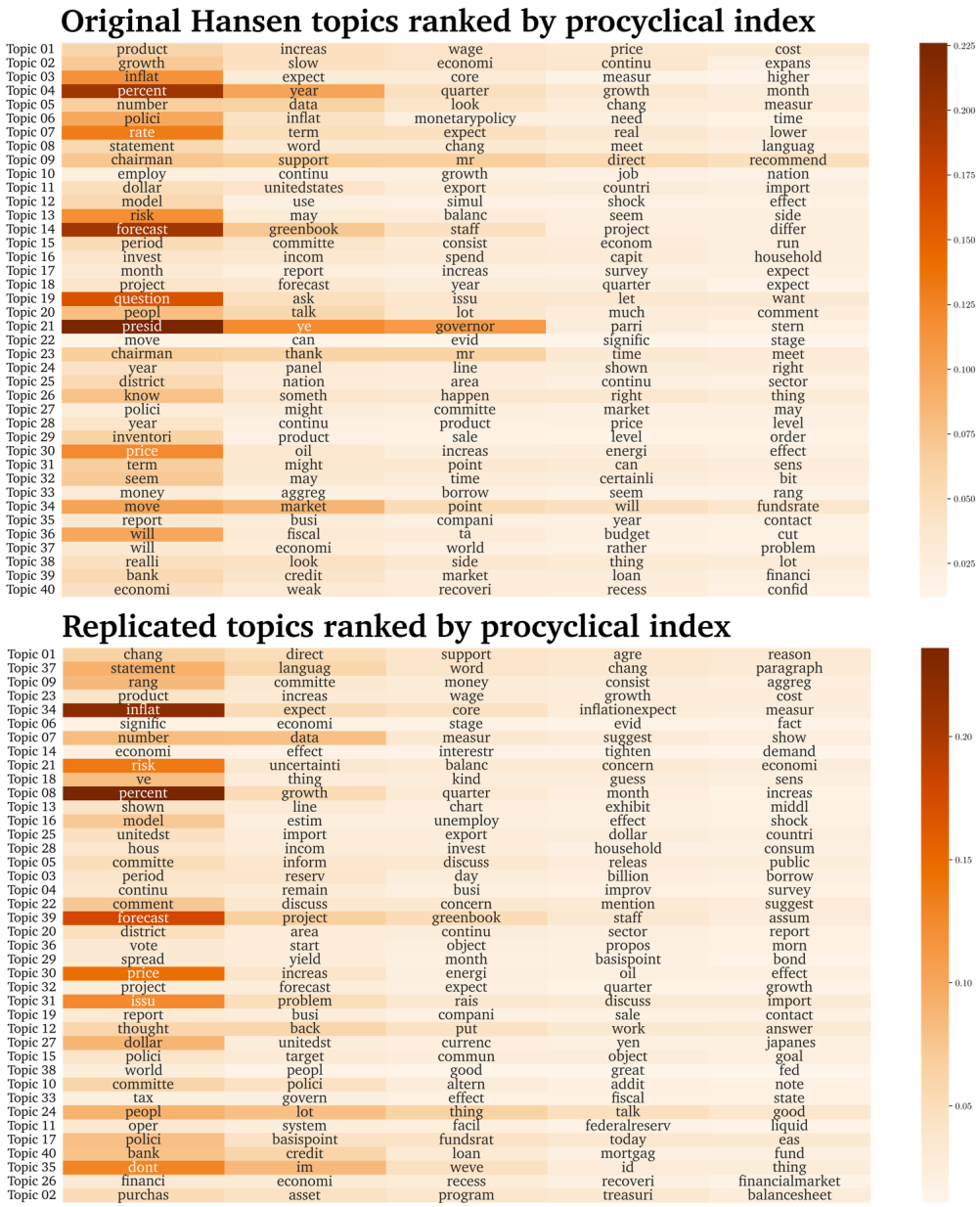


Figure 3.3: Comparison between Hansen et al. (2018)'s original term-topic distributions (top panel) and my replicated term-topic distributions (bottom panel).

Looking at the bottom panel it appears that I was able to reproduce a PI-based topic-ranking that was consistent with the economic intuition highlighted in the original paper. For example, my replicated Topic 23, ***productivity*** is ranked fourth from the top and Topic 34, ***inflation*** is ranked fifth from the top. These positions suggest that

Topics 23 and 34 are relatively procyclical. Meanwhile, the most countercyclical topics were Topic 26, ***recession*** and Topic 2, ***balance sheet expansion*** with rankings of ranked 39th and 40th respectively (at the bottom). I report the procycilatilty index values in Table A.3.

To validate the visual results in Figure 3.3, I compare the relative rankings between Hansen et al. (2018)'s most procyclical topics and my replicated topics in Table 3.2. I also report the topic label (as labelled in the original paper) and the intersection of terms between the original topic and the corresponding replicated topic in the same table.

Hansen's most procyclical topics

Topic label*	Original rank	Replicated rank	Intersection
productivity	1	4	product increas wage cost labor rise acceler trend worker tight job demand
growth	2	4	growth slow economi trend recent demand
inflation	3	5	inflat expect core measur higher slack gradual remain suggest low

*As labelled by Hansen et al. (2018)

Table 3.2: Common terms between Hansen et al. (2018)'s most procyclical topics and replicated topics, their topic labels and their relative rankings.

My replicated rankings for the most procyclical topics were consistent with Hansen et al. (2018)'s original rankings. For example, the original rank for the topic labelled *productivity* was 1 compared to my replicated rank of 4. The same could be said for the topics labelled (original rank 2; replicated rank 4) and *inflation* (original rank 3; replicated rank 5), whose replicated ranks were not more than two steps away

from the original ranks.[12]

Similarly, Table 3.3 indicates that the replicated rankings for the most countercyclic topics, *economic weakness*, *the financial sector* and *fiscal issues*, were all within three steps away from their original rankings.

Hansen's most countercyclical topics

Topic label*	Original rank	Replicated rank	Intersection
economic weakness	40	39	economi weak recoveri recess neg
the financial sector	39	37	bank credit loan financi debt lend fund financ spread asset commerci institut mortgag liquid larg
fiscal issues	36	33	fiscal budget cut govern effect state spend feder reduct packag

*As labelled by Hansen et al. (2018)

Table 3.3: Common terms between Hansen et al. (2018)' most countercyclical topics and replicated topics, their topic labels and their relative rankings.

[12] *Step* is defined as the absolute difference between the original rank and the replicated rank.

3.3.2.2 Topic correlations

The visual inspection depicted in Figure 3.3 and Tables 3.2 and 3.3 above appear to yield expected results. However, to ensure that this did not happen by chance, I check the correlation between a given replicated term-topic distribution and all of the original distributions. Intuitively, if the replicated distributions are a good representation of the gold standard.

I use the Jaccard Similarity Index (Equation 3.1) as the measure for the topic correlation between the original distributions and replicated distributions. In this context, the Jaccard Similarity Index measures the similarity between the two different *sets* of terms that make up individual topics (original versus replicated).

$$\text{Jaccard\_Similarity} = \frac{|A \cap B|}{|A| + |B| - |A \cup B|} \tag{3.1}$$

The index represents the fraction of common terms between sets A and B of the union of sets A and B. Two sets become more similar as the index approaches 1. The heatmap in Figure 3.4 shows the correlations (similarity indexes) between the original topics (left axis) and the replicated topics (top axis).

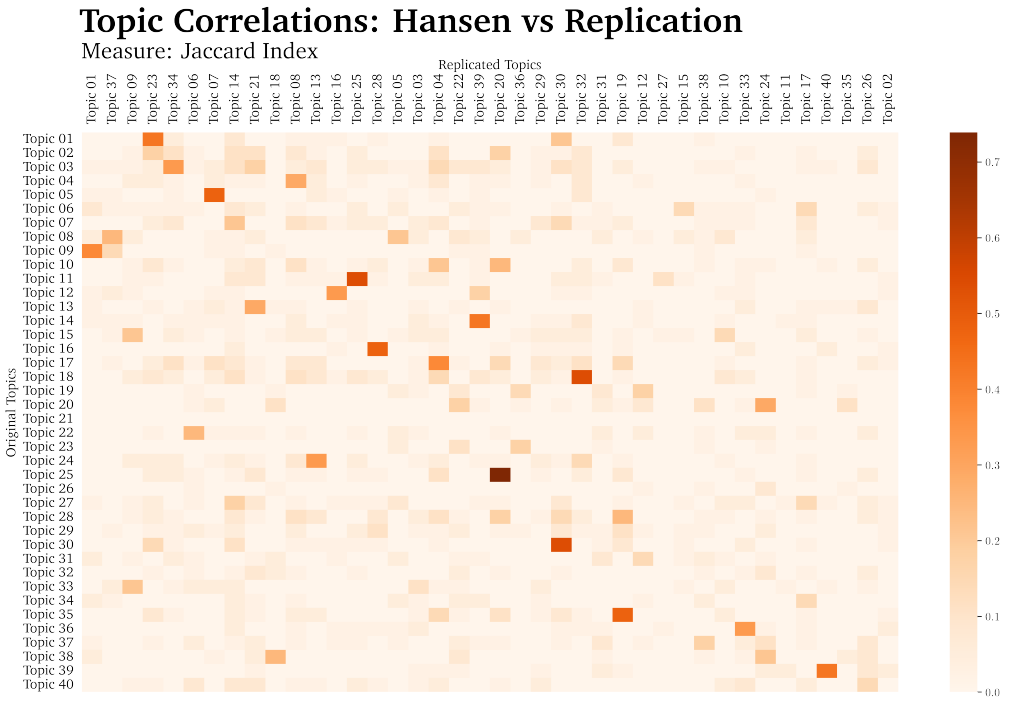


Figure 3.4: Heat map of *Jaccard Similarity* measures between Hansen et al. (2018)'s original term-topic distributions (left axis) and the replicated term-topic distributions (top axis).

As noted earlier, topics are ranked by the PI value calculated in Section 3.3.1 from most procyclical to most countercyclical. The original topics are ranked from top to bottom while the replicated topics are ranked from left to right. The correlations themselves are also in line with the intuition highlighted earlier. There are few highly correlated pairs (darker shades) with many low to zero correlation measures (lighter shades). Moreover, those highly correlated topic-parings are largely centred around the main diagonal (top-left quadrant to the right-bottom quadrant), meaning that replicated topics exhibit high correlation with original topics that fall within a similar ranking based on the PI measure, while being is relatively uncorrelated elsewhere. This further reinforces that the replicated topics are indeed a good proxy for the original topics.

Topic Correlations: Jaccard Similarity Index

Original (1)	Replica (2)	*Rank (3)*	diff (4)	JImax (5)	Original (6)	Replica (7)	*Rank (8)*	diff (9)	JImax (10)
Topic 01	**Topic 23**	**4**	**3**	**0.4286**	Topic 21	Topic 01	1	20	0.0000
Topic 02	**Topic 23**	**4**	**2**	**0.1765**	Topic 22	Topic 06	6	16	0.2500
Topic 03	**Topic 34**	**5**	**2**	**0.3333**	**Topic 23**	**Topic 36**	**22**	**1**	**0.1765**
Topic 04	**Topic 08**	**11**	**7**	**0.2903**	Topic 24	Topic 13	12	12	0.3333
Topic 05	**Topic 07**	**7**	**2**	**0.4815**	**Topic 25**	**Topic 20**	**21**	**4**	**0.7391**
Topic 06	Topic 15	30	24	0.1429	Topic 26	Topic 24	34	8	0.0811
Topic 07	**Topic 14**	**8**	**1**	**0.2121**	Topic 27	Topic 14	8	19	0.1765
Topic 08	**Topic 37**	**2**	**6**	**0.2500**	**Topic 28**	**Topic 19**	**27**	**1**	**0.2500**
Topic 09	Topic 01	1	8	0.3793	Topic 29	Topic 28	15	14	0.1111
Topic 10	Topic 20	21	11	0.2500	**Topic 30**	**Topic 30**	**24**	**6**	**0.5385**
Topic 11	**Topic 25**	**14**	**3**	**0.5385**	**Topic 31**	**Topic 12**	**28**	**3**	**0.1429**
Topic 12	**Topic 16**	**13**	**1**	**0.3333**	Topic 32	Topic 21	9	23	0.0811
Topic 13	**Topic 21**	**9**	**4**	**0.2903**	Topic 33	Topic 09	3	30	0.2121
Topic 14	**Topic 39**	**20**	**6**	**0.4286**	**Topic 34**	**Topic 17**	**36**	**2**	**0.1429**
Topic 15	Topic 09	3	12	0.2121	Topic 35	Topic 19	27	8	0.4815
Topic 16	**Topic 28**	**15**	**1**	**0.4815**	**Topic 36**	**Topic 33**	**33**	**3**	**0.3333**
Topic 17	**Topic 04**	**18**	**1**	**0.3793**	**Topic 37**	**Topic 38**	**31**	**6**	**0.1765**
Topic 18	**Topic 32**	**25**	**7**	**0.5385**	Topic 38	Topic 18	10	28	0.2500
Topic 19	Topic 12	28	9	0.1765	**Topic 39**	**Topic 40**	**37**	**2**	**0.4286**
Topic 20	Topic 24	34	14	0.2903	**Topic 40**	**Topic 26**	**39**	**1**	**0.1429**

Table 3.4: Most correlated topic pars between the original topics (Original) and replicated topics (Replica) based on the highest Jaccard Similarity Index (JIMax) in Eq. 3.1. The column(s) "Rank" represents the rank of *replicated topic*$_k$ based on the PI measure. Column(s) "diff" indicates the distance between the original topic's rank and the replicated topic's rank.

Lastly, Table 3.4 presents the original topic ranking alongside the replicated topic rankings and is another way to view the data underlying Figure 3.4. The way to interpret this table is as follows. Take the Original-Replica topic pair of *Topic 01-Topic 23* on the first row of the left-side of the table. This says that Original Topic 01 (*productivity*) is most correlated with Replica Topic 23 with a Jaccard Similarity Index score of 0.4286 (column 5). Replica Topic 23 was ranked as fourth (column 3) according to the PI measure. Its distance from Original Topic 01's rank is *three* steps away (column 4). A shaded row represents those topic pairs where the values in columns 4 and 9 are less than one standard deviation, or eight steps away from the Original Topic's rank. A green (red) shade represents a procyclical (countercyclical) topic pair.

The average similarity for procyclical topic pairs is around 0.4, while the average for countercyclical topic pairs is 0.3. With more than half of the most correlated

topic pairs lying broadly along the main diagonal in Figure 3.4 and their average similarities being sufficiently greater than zero, I can conclude that I can a) treat the replicated model as a proxy for the gold standard model, and b) use the document-topic-proportions that were estimated in Section 3.3.1 as the inputs to the logreg classifier which I discuss in Chapter 4.

3.4 Summary

Replicating Hansen et al. (2018)'s original was necessary to obtain inputs that are needed for a classification task. This chapter gives detailed steps of the process, from data cleaning to fitting Hansen et al. (2018)'s LDA model. Lastly and most important, my successful replication gives justification to use my output as a proxy for Hansen et al. (2018)'s model. Unless I make the distinction otherwise, any reference to the *benchmark*, *gold standard* or *Hansen et al. (2018)'s LDA model* from now on, refers to the replicated model. The next chapter outlines the bespoke classification task used to evaluate the competing models against the benchmark model.

Chapter 4

A classification task

As I highlighted in the previous chapter, much of the analysis in this thesis revolves around a bespoke classification task. This classification task is related to Hansen et al. (2018)'s findings mentioned in Chapter 2 where they were able to observe distinct behaviour between *rookies* and *veterans* by using a topic model (LDA) to analyse the topical structure underlying the FOMC text. In addition to being an issue of great practical importance, the framework that Hansen et al. (2018) established is a perfect opportunity to evaluate a set of topic models (including Hansen et al. (2018)'s LDA model) in an applied setting. Notably, while the topic modelling literature does sometimes use topic classification as an extrinsic evaluation method, the classification task in this thesis requires *more* of topic models, with the indicators of rookie status being more subtle, related not just to specific words indicating the issue being discussed but to many other indicators such as a member's manner of speaking.

4.1 The classification task

The setting in this case is, if the transcripts contain signals about a given FOMC member's type (rookie or veteran) — and this is what Hansen et al. (2018)'s analysis found — then we would expect to be able to train a model to predict that member's type based on the transcripts. I use a standard logistic regression classifier (logreg)

to undertake this task. Analogously to Hansen et al. (2018)'s use of document topic proportions as independent variables in their econometric models, I use the document topic proportions estimated (discussed later in this chapter) by each topic model as the input features to the logreg. Each topic model will then be evaluated by the logreg's accuracy in predicting a member's type (to be discussed in Chapter 6).

4.2 Defining the classification window

Firstly, it is important to define the *classification window* because it a critical element for the following analysis. While, the full dataset covers the period from 1987 to 2011, I follow Hansen et al. (2018) and reduce the sample to an *eight year* window covering the period between 1989 and 1997. This reduces the number sample documents to 9,783. This classification window singles out those 19 FOMC members who were present at the meetings just before and after the transparency event in October 1993. As noted earlier, this structural break allowed Hansen et al. (2018) to make statistical inferences about the behaviour of FOMC members during policy debates, given the classification window. To put it another way, Hansen et al. (2018) was able to statistically identify a *structure* within the FOMC transcripts themselves, that contained signals about the *type* of member through the topics they spoke about.

4.3 Logistic regression classifier input features

I treat the document topic proportions ($\hat{\theta}_d$)estimated from each topic model as a proxy for the topics a given FOMC member spoke about for a given policy meeting. I use $\hat{\theta}_{d,k}$'s as the inputs to a simple logistic regression classifier (Equation 4.1) to estimate the probability p of whether FOMC member i is a rookie or not (I will omit the subscripts going forward for neatness). Here, $z = \mathbf{X}'\beta$, where $\mathbf{X}$ is a vector of

input features ($\hat{\theta}$s from topics 1 to 40) and β is a vector of coefficients or weights (estimated by OLS) associated with each feature. The predicted outcomes ($\hat{y}$) are articulated in Equation 4.2. I repeat the process for the competing topic models and compare the classification results against the benchmark LDA model. I discuss the results later in Chapter 6.

$$p = \frac{1}{1 + e^{-z}} \tag{4.1}$$

$$\hat{y} = \begin{cases} 1 \text{ (Rookie)}, & \text{if } p \geq 0.5 \\ 0 \text{ (Veteran)}, & \text{otherwise} \end{cases} \tag{4.2}$$

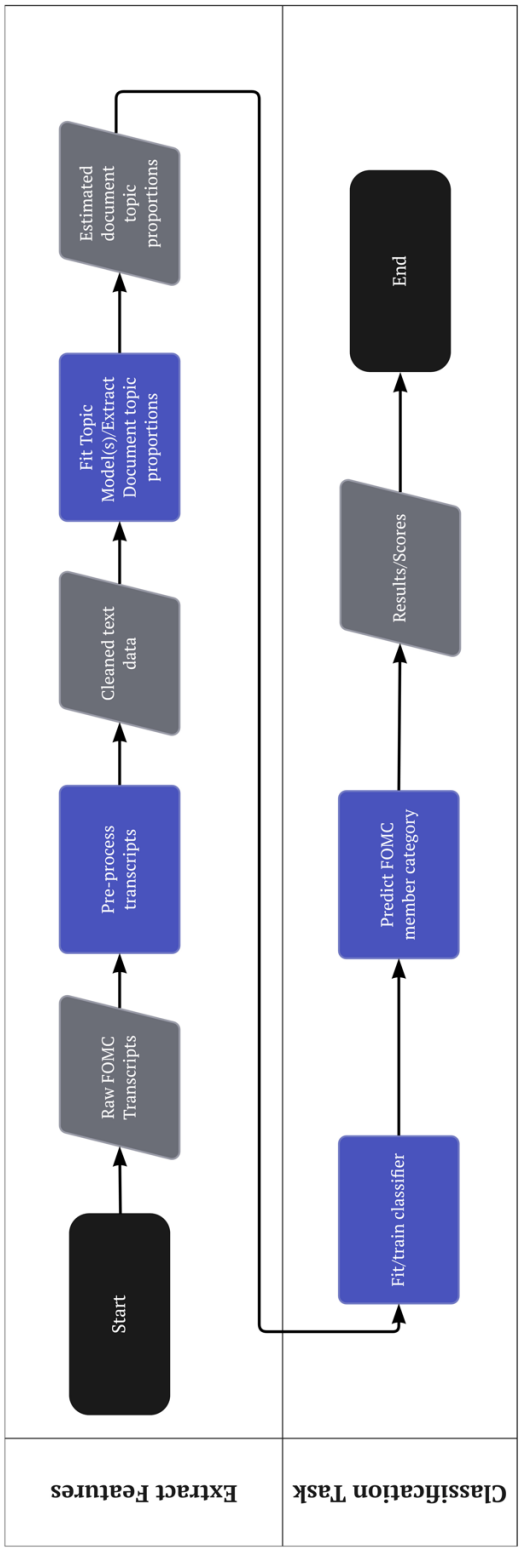


Figure 4.1: End-to-end process for one cycle of the classification task. This process is repeated for each topic model.

Figure 4.1 is a flow-chart representing one cycle of the classification process end-to-end. The grey shapes represent the input/output data. The blue boxes represent processes. The top lane depicts the process for extracting the input features for the logistic regression classifier, while the bottom lane maps out the downstream classification task which I look to evaluate.

4.3.1 Estimating FOMC members' experience

Before running the logistic regression classifier above, I need to estimate the number of years' experience an FOMC member had in the year a given policy meeting took place. The original paper indicates that the measure of an FOMC member's experience is:

> *[T]he number of years a given member has spent working in the [US Federal Reserve] system through a given meeting* (Hansen et al., 2018, p 838).

I interpret this to be the difference between the year of a given meeting and the year that a given member started serving on the committee (year of appointment) *plus* the number of years that same member had worked in the US Federal Reserve prior to their appointment (Equation 4.3).[1]

$$\text{TotalFedExp}_{i,t} = \text{TimeServed}_{i,t} + \text{PriorFedExp}_{i} \tag{4.3}$$

Figure 4.2 plots the distribution of the estimated total number of years' experience of each FOMC member over the classification window,[2] while Table 4.1 presents some summary statistics. A member has around nine years' of total Fed experience

[1] There are three speakers within the sample window where their status changed to *member* from non-member. That is, their appointment happened after the beginning of the sample window, but before November 1993. In these instances, I set the value for their time served on the committee to zero where meetings occurred prior to November 1993.

[2] This counted at the individual statement level.

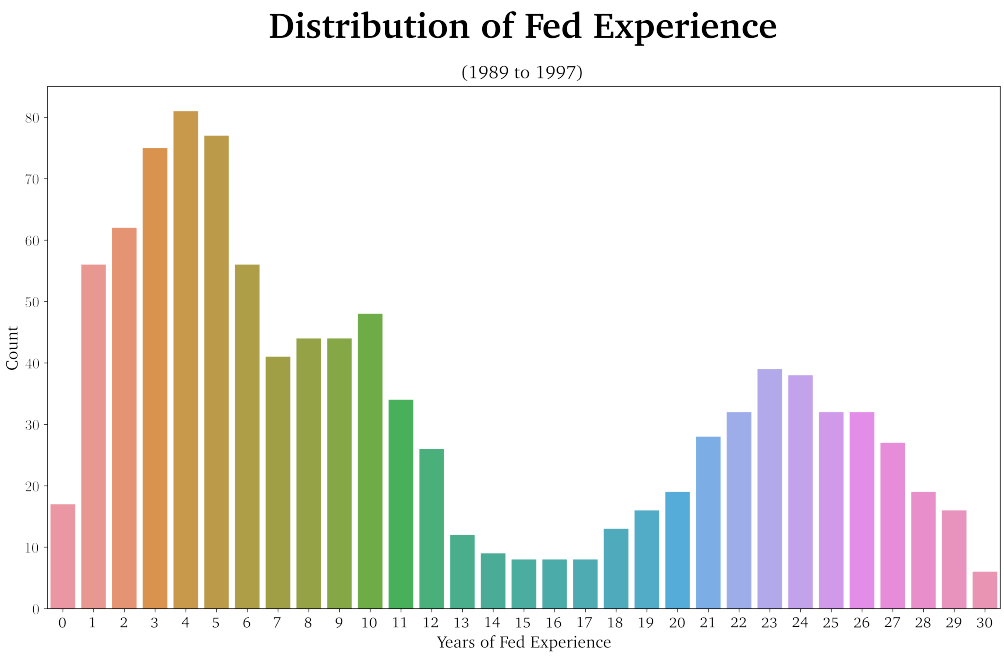


Figure 4.2: Distribution of FOMC members' experience in years over the sample (counted at the individual statement level).

Estimated years of Fed Experience				
mean	std	min	max	No. of meetings
9	7	0	30	72

Table 4.1: Summary statistics for FOMC members' total Fed experience from 1989 to 1997.

on a given meeting, on average.

4.3.2 Defining the *Rookies*

Hansen et al. (2018) loosely defines rookies as those members who have about 20 years less Fed experience than their colleagues. Looking at Figure 4.2 this is roughly the distance between the two peaks of the distribution. Using this as a general guide, I take the following approach when labelling rookie members at a given meeting:

1. For a given meeting, I compute the maximum years of Fed experience. I then take the difference between the maximum experience and the experience of

a given member. I call this value the *distance to maximum*.

2. If the distance to maximum for a given member in a given meeting is more than two standard deviations from the maximum, then I classify that member's status with a 1 (for rookie) for that meeting and apply this label to all respective statements for the same meeting, or 0 otherwise (for veteran).

3. **Tie-breaker**. However, the logic outlined in steps 1 and 2 above does not ensure that a member remains in one category or the other. Indeed, there are two members who switch categories from a rookie to a veteran, then back. In these instances, I use a tie-breaker logic which is as follows. If their average *rookiness* over the sample window is greater than 50 per cent, then they are defaulted to a rookie.[3, 4]

4. The Fed Chair is ***not*** a rookie. Greenspan had only been appointed as Fed Chair two years before the beginning of the sample window, and is recorded as having no prior Fed experience. Despite this, the Fed Chair sets the policy agenda and it is arguably the other members who have to convince the Chair to think otherwise. As such, a Fed Chair – Greenspan in this instance – is not a rookie.

After completing steps 1 to 4 above, I get an average proportion of rookies in a given meeting of around 55%, though this value fluctuated between 45% and 61% over the sample. Despite this near-even split between rookies and non-rookies (veterans), Table 4.2 shows that the proportion of statements (documents) that were uttered by rookie members was noticeably less than 50%. This imbalance was particularly more pronounced within the test (holdout) sample. However, looking at the average ***length*** of each document (Table 4.3), it appears that rookie

[3] Committee members Kheen (defaulted to Rookie) and Syron (defaulted to Veteran) were those that switched categories within the sample. Their average rookiness measures were 63 per cent and 31 per cent respectively.

[4] No measure is perfect, and I have to draw the line somewhere.

Share of Rookie Statements		
	%	No obs.
Full sample	42.76	9,783
Training sample	43.67	8,315
Heldout sample	37.60	1,468

Table 4.2: Proportion of sample statements uttered by ***rookie*** FOMC members.

Average Document Length Over Sample				
	Rookie		Veteran	
	Average length	No obs.	Average length	No obs.
Full sample	52	4,183	34	5,600
Training sample	47	3,631	32	4,684
Heldout sample	79	552	49	916

Table 4.3: Average number of terms per document made by rookies and veterans over the sample.

members tended to deliver statements that were around 1.5 times longer than those delivered by veterans. Based on this observation, one could neatly summarise the characteristics of the sample with the following anecdote: *On average when compared to veterans, rookies only spoke half as often, but for nearly twice as long.*

4.4 Summary

This chapter described the task of *classifying and FOMC member*, the chosen classifier, a logreg and its required inputs. I also described the classification window and its significance and defined what a rookie is in terms of their years of experience. The next chapter will outline how I fit the newer competing models to the classification window data as well as a brief qualitative description of those newer models.

Chapter 5

Competing models: ETM and CatE

This chapter describes estimation process for fitting the two remaining topic models to the pre-processed FOMC data described in Chapter 3.[1] The first competing model is Dieng et al. (2019b)'s Embedded Topic Model (ETM) with the second being Yu et al. (2020)'s Category-Name Guided Text Embedding model (CatE). I then give a qualitative description of the ETM and CatE models' output to address the question stated in **RQ1**.

5.1 LDA vs ETM

5.1.1 Fitting the ETM

To fit the ETM I use Terragni et al. (2021)'s OCTIS package because it was simple to use and provided an easy way to obtain document-topic-proportions.[2] The model setup is fairly straight forward where I only specify two parameters:

1. I set the number of topics k=40,
2. I set the number of epochs to 300, and
3. I set the model to train on the *full* dataset.

[1] I have already discussed the data preparation and how I estimated the inputs (document topic proportions θs) to the logistic regression classifier for the benchmark LDA model in Chapter 3 and will not repeat them in this chapter.

[2] The OCITS source code can be found here: https://github.com/MIND-Lab/OCTIS. Dieng et al. (2019b) original source code can be found here: https://github.com/adjidieng/ETM.

I leave the remaining hyperparameters as their OCTIS default values.

Next, I fit two versions of the ETM over the corpus. The first version learns the word embeddings from a randomly initialised starting point and finds topics in the embedding space *simultaneously*. I call this version "ETM-Trained". The second version is trained using a previously fitted embedding and call it "ETM-Pretrained". I use skip-gram embeddings (Mikolov et al., 2013a) for the ETM-Pretrained model.[3]

5.1.2 Qualitative results: ETM

The panels on Figure 5.1 show heat maps of the topics estimated by the ETM-Trained (top) and ETM-Pretrained (bottom) models ranked by the PI measure. Looking at the ETM-Trained output, one immediately notices the number of what appear to be '*stop-topics*' (analogous to stop words) or *speaking topics*.[4] These are word distributions are focused on pure dialogue where relatively large probabilities are placed on terms such as *dont, im* and *weve* (Topics 31 and 8, for example). Fifteen out of the 40 estimated topics are represented by stop topics. Of the remaining non-stop topics, only a handful seem to be consistently ranked according to the PI measure. For example, Topic 33 *inflation* appears to be a correctly placed topic being seventh from the top. Similarly, Topic 35 euro-system debt referring to the *European sovereign debt crisis* and Topic 25 *financial markets* rank fifth and second from the bottom respectively. The rest of the topics appear to be related to issues that are discussed in every meeting regardless of the business cycle such as staff forecasts (Topics 22 and 38) and reports on regional economic activity (Topics 19 and 15) (Hansen et al., 2018, p.882).

[3] I use a skip-gram embedding with 300 dimensions. The file can be found here `https://bitbucket.org/diengadji/embeddings/src/master/`.

[4] Technically, all topics estimated by each model are *speaking* topics because the documents are raw transcripts of *utterances* made by people.

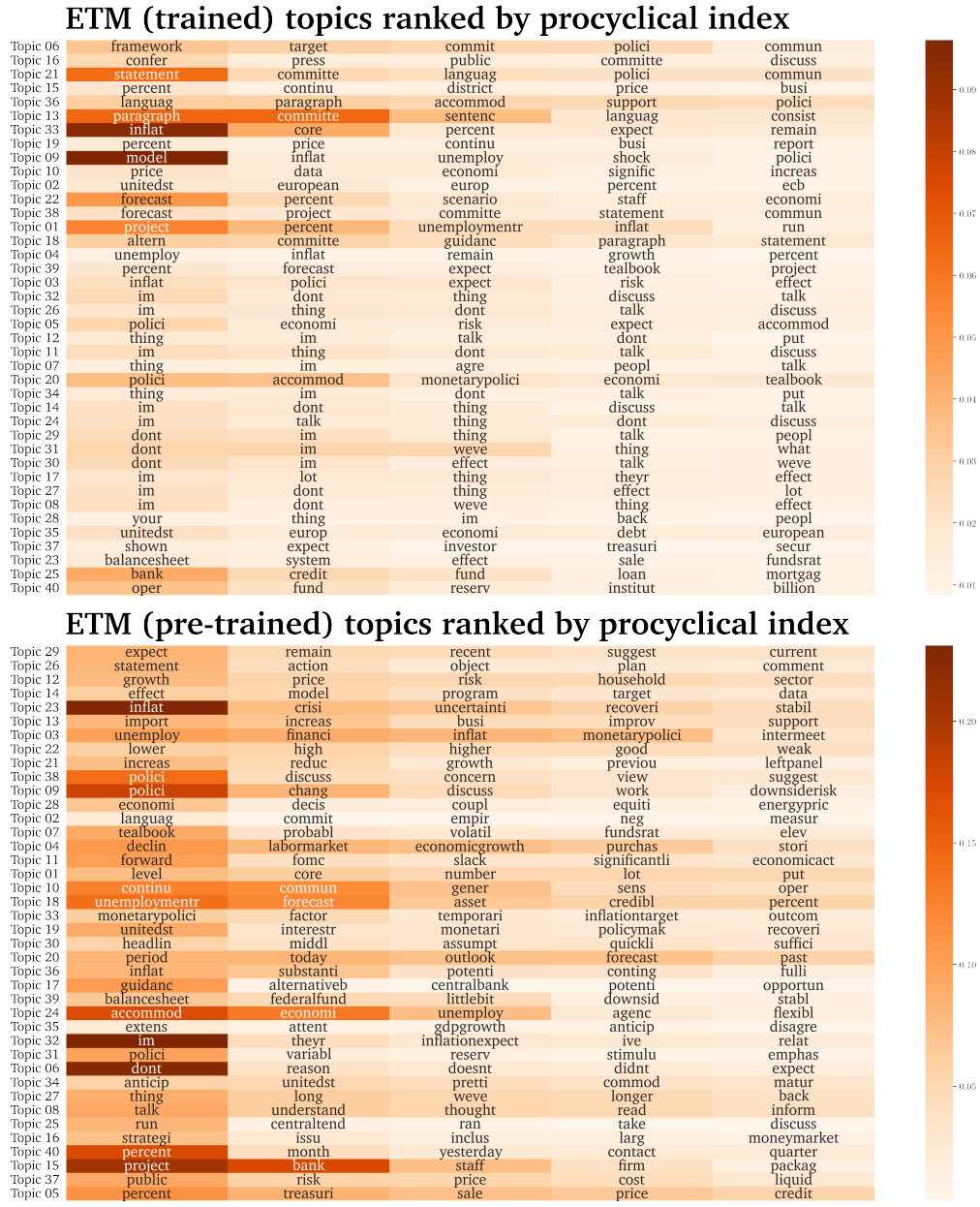


Figure 5.1: Comparison between the ETM-Trained term-topic distributions (top panel) and the ETM-Pretrained term-topic distributions (bottom panel). The layout is the same as Figure 3.3

Turning to the ETM-Pretrained model output on the bottom panel. There are noticeably less stop-topics, however, it appears that only Topic 12 *growth* (procyclical) is consistently ranked as per the PI measure. The ordering for the remaining topics

seem to be jumbled or inconsistent. Take Topic 23 for example, which ranks fifth from the top as a procyclical topic. Its top-weighted term is *inflat* which makes sense. However, the next two highest-weighted terms are *crisi* and *uncertainti* which clearly do not belong in the procyclical (top) end of the PI measure. Compare this to its corresponding topic — Topic 34 — estimated by the gold standard LDA model in Chapter 3 (Figure 3.3, bottom panel). The terms *crisi* and *uncertainti* do not appear in the distribution which is correct.

Perhaps one reason for the ETM-Pretrained model for producing such mixed rankings is that it was unable to distinguish between those topics that were genuinely related to the business cycle, which resulted in multiple topics being inadvertently blended together. This is evident in Figure 5.2, where a given ETM-Pretrained topic exhibits a high correlation across multiple LDA topics on either end of the PI spectrum (unlike what is seen in Figure 3.4). The ETM-Trained model's heatmap paints a similar picture (see Figure A.1).

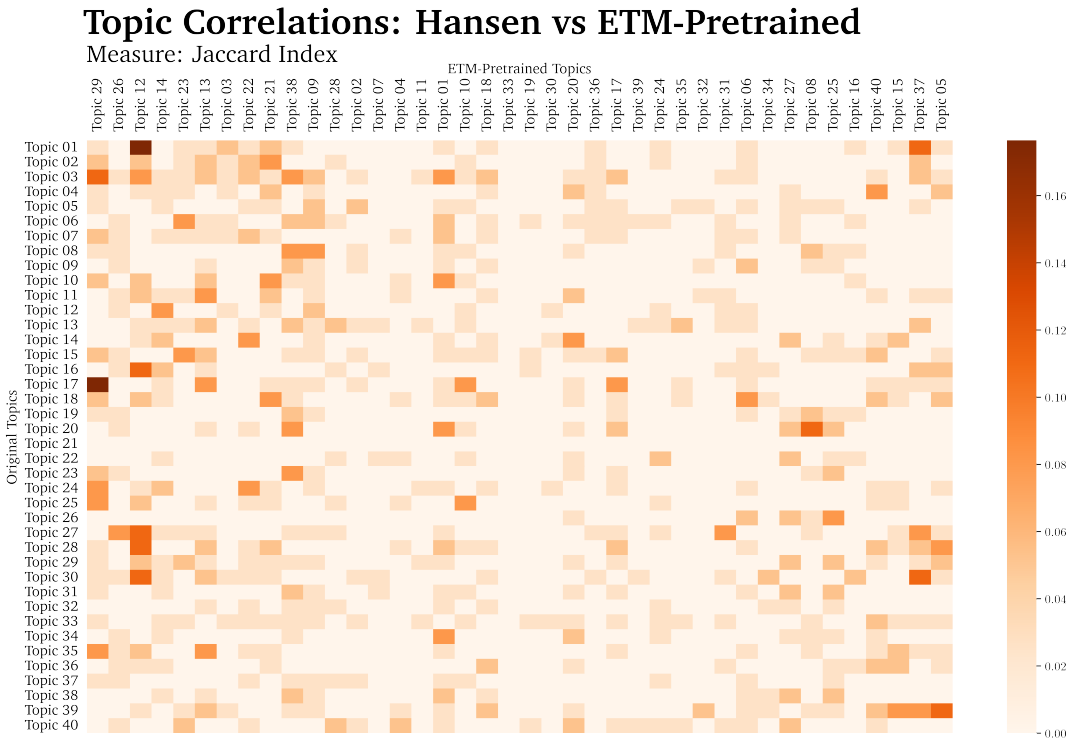


Figure 5.2: Jaccard Index measures between benchmark LDA and ETM-Pretrained topics.

5.2 LDA vs CatE

5.2.1 Fitting the CatE

As noted in Section 2.5, Yu et al. (2020)'s CatE model takes in a set of terms, or categories as inputs to the topic mining process. I fit four version of the CatE model using the package provided by the authors (see `https://github.com/yumeng5/CatE`). For all versions I use the following parameter settings:

1. the number of terms to be returned per topic to 20,
2. the minimum term count to 2,
3. the embedding dimensions set to 100,
4. the local context window size, $h = 5$.

Other model settings are set to default values in the original package.[5] I split the four versions into two groups. Group one is where the word embeddings are learned simultaneously alongside the estimation process. Models that fall within this category are labelled with "Trained". Group two is estimated using pre-trained word embeddings. Models that fall within this category are labelled with "Pretrained".[6] For each group I train two versions. The first uses the top-weighted terms estimated by the original LDA model as the input categories to the CatE. The reason for choosing the set inputs in this way is that top term from Hansen et al. (2018)'s original topic estimations are likely to be a good representation of the underlying data and, as such, are a suitable choice of categories. These are the terms located on the left-most column of the top panel in Figure 3.3. These models are labelled as "original". Similarly, the second version uses the top-weighted terms estimated by the replication (the left-most terms on the bottom panel of Figure 3.3). Again, I apply the same rational here for choosing the set of categories. These models are labelled as "replica".[7]

5.2.2 Qualitative results: CatE

To save space, I only present the output for the CatE Trained Replica model. Table A.1 shows all CatE Trained Replica model topics ordered by the PI measure. There is little difference between the CatE trained and pretrained models for the replicated seeds and as such, the pretrained output are not provided. The same applies to the CatE models initialised with the original LDA terms and I do not discuss them further. I provide the full table and heatmap for the CatE Trained Original model in Table A.2 and Figure A.2 in Appendix A, respectively.

[5] See https://github.com/yumeng5/CatE for description of all model settings.

[6] I use the default pre-trained embedding file supplied by the authors. The file can be downloaded here: https://github.com/yumeng5/CatE/blob/master/word2vec_100.zip

[7] There are some instances where the top-weighted terms repeat down the order. In these cases I use the next term with the highest weight (the next term to the right).

CatE Trained Replica: Selected Topics

category	*Topic 9* **inflat**	*Topic 37* **spread**	*Topic 39* **financi**
	inflationr	cdx	financialsector
	inflationexpect	widen	financialmarket
	core	yield	stress
	riseininfl	narrowest	intermediari
	disinfl	narrow	fragil
	anchor	treasuryyield	creditmarket
	pceinflat	grade	strain
	measuresofinfl	cd	financialsystem
	declineininfl	bbb	turmoil

Table 5.1: Selected topics from CatE Trained Replica.

Table 5.1 shows the few topics that ranked consistently against the PI measure and a handful of their corresponding terms. The topic seeded with the term *inflat*, ranked ninth (Topic 9) and is a relatively procyclic topic as one expects. Topics seeded with the terms *spread* and *financi* were ranked as relatively countercyclic topics (37^{th} and 39^{th} respectively). The rank of the *spread* topic might not be obvious at first glance, but in times of financial market volatility, market commentators typically focus on the difference or — the *spread* — widening between yields on risky assets like *BBB* (junk) rated bonds and the yields on safe assets like US Treasuries. As such, this topic being placed in the countercyclic end of the spectrum is correct.

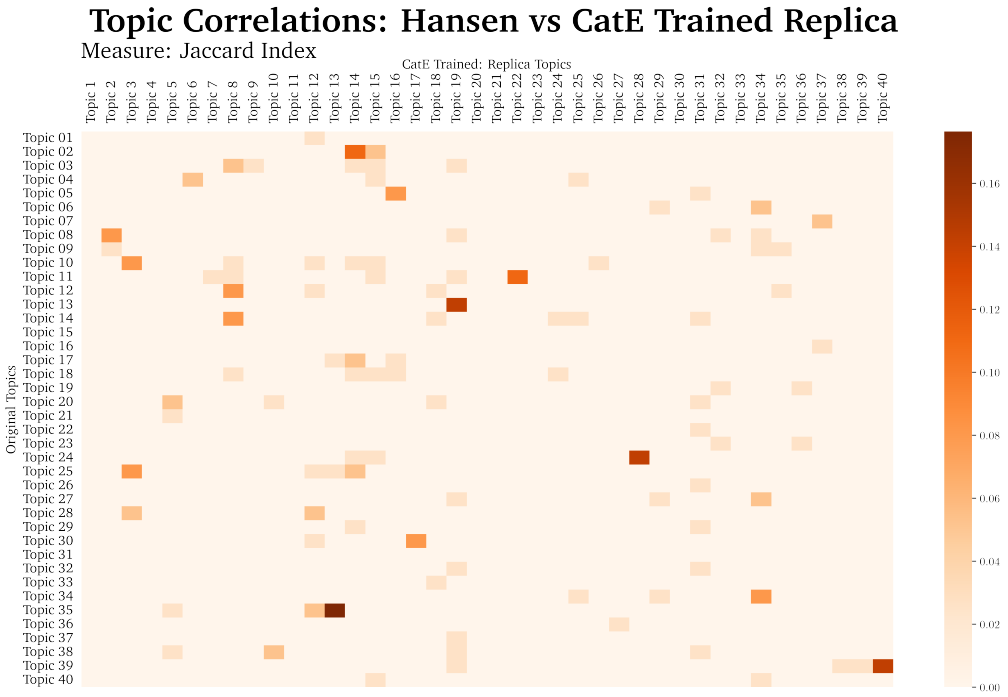


Figure 5.3: Jaccard Index measures between benchmark LDA and CatE Trained Replica topics.

The LDA-CatE correlation heatmap (Figure 5.3) shows that the majority of topic pairs are uncorrelated. But, unlike Figure 3.4, those topic pairs that do exhibit high correlation appear to be scattered more randomly as opposed to lying along the main diagonal (top left to bottom right). One possible explanation for this outcome is that the CatE model primarily focuses on how *similar* the target words are to the seed word and it is likely that some terms that should be opposite sides of the PI spectrum can appear in the same CatE-generated topic. Going back to the topic of *spreads* (Topic 37). The terms *widen* (a countercyclic term) and *narrow* (a procyclic term) are considered by CatE to very similar to the seed word *spread*.

5.3 Summary

This chapter provided an outline of the process for fitting Dieng et al. (2019b)'s ETM and Yu et al. (2020)'s CatE models as well as providing qualitative analysis to address **RQ1**.

Despite *only* some correctly placed topics, the qualitative inspection suggests that both ETM and CatE models do not do a good job (relative to the gold-standard LDA model) of distinguishing topics that are procyclic or countercyclic as defined by the PI measure. The next chapter discusses the results from the downstream classification task and topic quality (**RQ2** and **RQ3**).

Chapter 6

Classification and Topic Quality Results

This chapter presents evaluations of the topic models explored in this thesis (LDA, ETM, CatE) applied to the FOMC transcripts. It first discusses the results of the classification task described in Chapter 4 for each model. I present two versions of the results. The first version shows results based on model estimates conducted at the statement (document) level. These are called the *ungrouped* results. The second version is based on model estimates conducted at the meeting-speaker-section level as in Hansen et al. (2018). These are called the *grouped* results.

I then report the *quality* of the estimated topics from all models using the three measures in outlined in Chapter 2: **pmi**, **npmi** and **lcp**.

Lastly, I answer the questions stated in **RQ2** and **RQ3** regarding whether the new topic models are better than LDA in terms of downstream classification task and standard evaluation measures, respectively; and I discuss whether they come to the same conclusion.

An important note

This chapter is littered with statements that read like:

The LDA model outperformed the ETM-Trained model...

This is taken to mean:

The logistic regression classifier whose input features were generated by the LDA topic model, outperformed the logistic regression classifier whose input features were generated by ETM-Trained model...

It is much easier to write it the first way.

6.1 Classifier results

Here, I define *rookies* as the positive class and *veterans* as the negative class. That said, I report the following five standard measures to evaluate the performance of each model:

1. **Heldout Accuracy**, measures the proportion of correct predictions over the heldout sample.
2. **True Rookie Rate**, measures the proportion of true positives (tp) to the sum of the false negatives (fn) and true positives (tp): $tp/(fn + tp)$. This is also known as Recall.
3. **True Veteran Rate**, measures the proportion of true negatives to the sum of false positives and true negatives: $tn/(fp + tn)$.
4. **Rookie Precision**, measures the proportion of true positives to the total number of positive predictions: $tp/(fp + tp)$.
5. **Veteran Precision**, measures the proportions of true negatives to the total number of negative predictions: $tn/(tn + fn)$.

These measures can be obtained from each model's confusion matrix.

6.1.1 Heldout sample results: ungrouped

Table 6.1 presents the results of the FOMC member classification task based on the ungrouped heldout sample.

Heldout Sample Results: Ungrouped

Model	Heldout Accuracy	True Rookie Rate	True Veteran Rate	Rookie Precision	Veteran Precision
LDA	**0.7037**	**0.5417**	0.8013	0.6216	**0.7437**
ETM-Trained	0.7010	0.4511	0.8515	**0.6468**	0.7202
ETM-Pretrained	0.6628	0.3370	**0.8592**	0.5905	0.6826
CatE Trained Replica	0.6764	0.4692	0.8013	0.5873	0.7147
CatE Pretrained Replica	0.6805	0.4855	0.7980	0.5916	0.7202
CatE Trained Original	0.6567	0.4149	0.8024	0.5585	0.6947
CatE Pretrained Original	0.6601	0.4493	0.5585	0.5598	0.7034
No. obs	*1,468*				
Heldout size (%)	*15*				
Proportion of Rookies	*0.3760*				
Proportion of Veterans	*0.6240*				

Table 6.1: Classification results based on heldout sample.

At first glance, the held accuracy for all models is higher than the majority class baseline which in this case is the veteran category at 62% of the heldout size. This lends support to Hansen et al. (2018)'s original analysis. In this instance, by using a classification task version of the natural experiment, I have been able to find evidence which distinguishes between rookies and veterans based on their utterances.

Overall, the (replicated) benchmark LDA model outperformed the competing models in Heldout Accuracy, True Rookie Rate and Veteran Precision measures (green text). Also, the LDA model was the only model to score above 0.5 in the True Rookie Rate category and was not the worst performer in the True Veteran Rate and Rookie Precision measures.

One standout observation, however, is how poorly the ETM-Pretrained model per-

formed in the True Rookie Rate category, scoring 0.3370. It underperformed the LDA model by 60% and underperformed by 25% compared to the average of all the models in the same category. It was also ranked third worst by the Heldout Accuracy measure, underperforming LDA by around 6%. Meanwhile, the ETM-Trained model outperformed ETM-Pretrained in four out of the five measures and was much closer in performance to the LDA model.

To better understand what is happening, I take a closer look at the ETM-Trained and Pretrained results to see how each model is behaving relative to the benchmark LDA. I start by filtering each model's results by the instances where the benchmark LDA model made a *correct* prediction and where the ETM models made an *incorrect* prediction. I then group the filtered results by speaker.

Looking at the distributions of the misclassified FOMC members (Figure 6.1) it is clear that the bulk of incorrect predictions for both ETM models are centred around five members: Greenspan (whose true classification is a veteran), Jordan, Lindsey, McDonough and Parry (whose true classifications are rookies). Based on this observation, it appears that the ETM models find it relatively difficult to correctly classify rookies.

Next I examine the estimated coefficients from the logistic regression classifier to see if there are any noticeable features in which each model giving large weights.

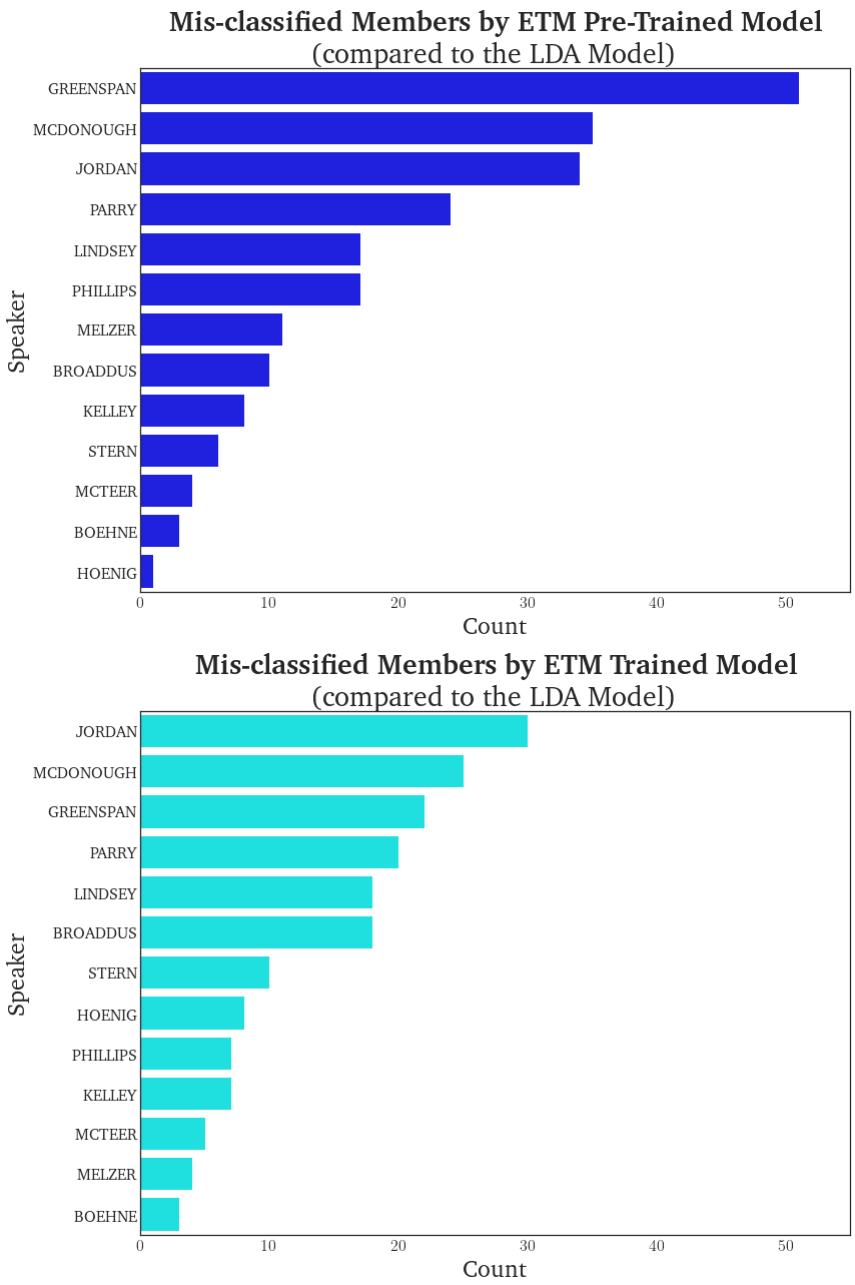


Figure 6.1: Misclassified FOMC Members by ETM-Pretrained model (top) and ETM-Trained model (bottom) relative to LDA model. The x-axis is labelled "Count" for both panels and represents number of times a given speaker was incorrectly classified.

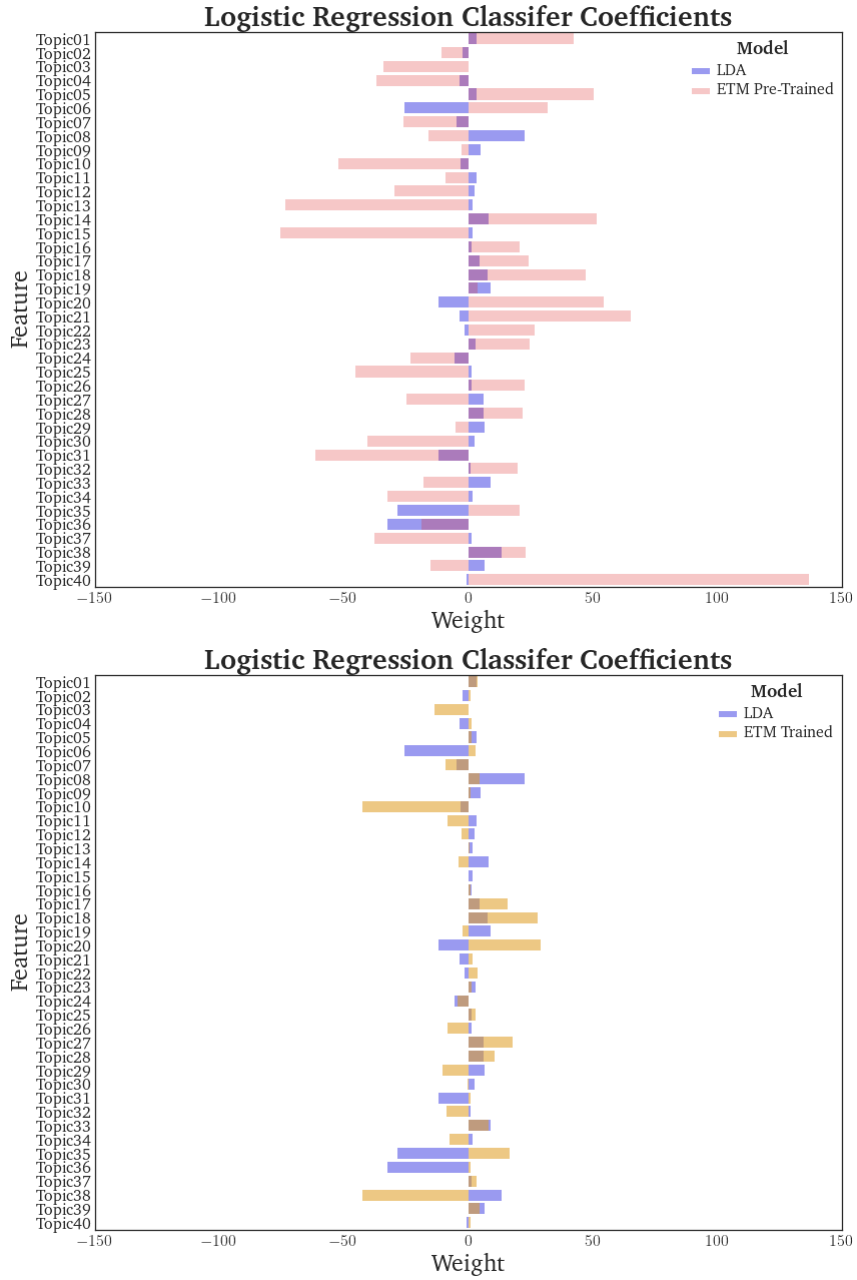


Figure 6.2: Estimated coefficients of Logistic Regression Classifier. The top panel shows the comparison between the LDA model and the ETM-Pretrained model. The comparison between the LDA and ETM-Trained model is on the bottom.

Figure 6.2 shows a comparison between the different sets of logistic regression classifier coefficients generated by each topic model. The top panel shows that the estimated coefficients the ETM-Pretrained model (pink bars) are much larger in magnitude compared to the LDA model (purple bars). Moreover, it appears that the ETM-Pretrained model places large weights on many features indicating that many features are important. In contrast, the LDA model appears to be more selective in which feature it deems important.

Looking at the features themselves, the ETM-Pretrained model gave the biggest weights to Topics 40, 15 and 13 which all relate to general staff reports and updates (see the bottom panel of Figure 5.1 for the ETM-Pretrained word distributions). In contrast, the LDA model identified *speaking* topics, such as 36 ⟨*vote*, *start*, *object*, *propos*, *morn*⟩ and 35 ⟨*dont*, *im*, *weve*, *id*, *thing*⟩, as the most important features. This suggests that LDA model found cleaner signals from the natural dialogue among members when it came to classifying those members, as opposed to speech that was pre-written beforehand (like pre-prepared staff reports and regional updates).

The ETM-Trained model (bottom panel; gold bars) produced coefficient weights that are closer in magnitude to that of the LDA, while also placing large coefficient weights on only a few features. However, unlike the LDA, the ETM-Trained model tended to place more weight on topics that were less related to natural dialogue. And while the Heldout Accuracy scores between the LDA and ETM-Trained model are quite similar, the ETM-Trained model's True Rookie rate is still below 0.5 (much like its ETM-Pretrained counterpart).

Lastly, it is worth briefly noting CatE's performance. Like the ETM, the CatE models were unable to surpass the LDA model's performance in any category. The CatE model's performance relative to the ETM models was not noticeably different on average either, although the CatE Trained model with Hansen et al. (2018)'s original input categories scored the lowest in the Heldout Accuracy and Rookie Precision

measures at 0.6567 and 0.5585 respectively.

6.1.2 Class imbalance

Given that the True Rookie Rates are low particularly for the ETM models, while True Veteran Rates are the highest, it could be the case that either ETM models are generally poor for this downstream task, or that they produce classification models that are more sensitive to class imbalance; that is, here they are skewed towards the majority veteran class. As noted earlier, rookies represent less than half of the training sample (and only slightly more than a third of the heldout set). As such, I equalise the proportions of rookies and veterans within the *training* set to check whether the True Rookie Rate improves. To achieve the desired equal class proportions, I remove documents that are classified as veterans at random from the training set. This leaves a training set with a reduced size of 7,261 observations.

Next I create another training set with the same *size* as the balance-adjusted training set, but this time keeping the *same* class proportions as the original imbalanced training set (a roughly 60/40 split in favour of the veterans). I apply the same procedure of randomly removing observations until the original class proportions are achieved. This scaled training set will serve as the *control* set. I then re-run the logistic regression classifier over the new training samples (both balanced and scaled) each model and make predictions over the *same* heldout set.

6.1.2.1 Revised classifier results

The revised Heldout Accuracy results for the LDA and ETM models were slightly worse when compared compared to the unadjusted training set (Table 6.2). This is the result of randomly throwing away information in the test set and is to be expected (*ceteris paribus*).

What is important though, is that the three models showed a marked improvement to

their True Rookie Rates to being much better than a coin-flip (with scores improving by 46% on average when compared to their unadjusted scores). In particular, the ETM-Pretrained model nearly doubled its True Rookie Rate to a little under 0.60 (compared to its original score of 0.34).

Heldout Sample Results: Balanced

	Heldout Accuracy		True Rookie Rate		True Veteran Rate	
Model	Balanced	*Scaled*	Balanced	*Scaled*	Balanced	*Scaled*
LDA	0.6689	*0.7044*	**0.7065**	*0.5380*	0.6463	*0.8046*
ETM-Trained	0.6680	*0.7037*	**0.5942**	*0.4565*	0.7445	*0.8526*
ETM-Pretrained	0.6144	*0.6669*	**0.5888**	*0.3460*	0.6299	*0.8603*
Reduced training size	*7,261*					

Table 6.2: Heldout results based on a balanced *training* sample.

In terms of its *relative* performance, the ETM-Pretrained model's True Rookie Rate originally underperformed the LDA model by around 40%. However, this margin deficit more than halved to 17% after the balancing adjustment. On the other hand, the ETM-Trained model's underperformance gap was little changed (before and after balancing at 17% and 16% respectively).

Meanwhile, the measures under the *scaled* columns in Table 6.2 are nearly identical to the original values reported in Table 6.1 above.

Overall, adjusting for class imbalance in the training set resulted in the following: 1) the True Rookie Rate for the LDA and both ETM models improved significantly in absolute terms; 2) the relative difference in the True Rookie Rate only improved for the ETM-Pretrained model; 3) despite its marked improvement, the ETM-Pretrained model's performance deficit to LDA model is still wide. This suggests that the poor performance of the ETM models is a mix of being affected by class imbalance and generally lower performance; understanding exactly why this is the case could benefit from further investigation.

6.1.3 Heldout sample results: grouped

Table 6.3 shows the classification results based on the following grouping criteria: group the sample by *meeting*, then *speaker* and then *section*. The grouping process reduces the size of the classification window to 2,560. The grouping process also causes the majority class to flip, in favour of the rookies in both train (54.5%) and heldout (52.1%) sets. After grouping the sample, I take the average of the document-topic-proportions (θs) and these $\bar{\theta}s$ become the new input features to the logistic regression classifier. Grouping the data by speaker brings the analysis more in line with Hansen et al. (2018)'s original setup.

Like the results reported in Table 6.1, I observe that all the models have an overall Heldout Accuracy score well above the majority class baseline (Table 6.3). These grouped-adjusted results further reinforce Hansen et al. (2018)'s findings of being able to distinguish between rookie and veteran speakers.

Looking at Table 6.3 more closely, the group-adjusted results show a marked improvement in the True Rookie Rate as more information is aggregated into each $\bar{\theta}$ for a given class (recall from Chapter 4, that the average proportion of Rookies within a given *meeting* was slightly over half). Moreover, the gains in performance are such that the models were picking rookies at better than a coin-flip (a near 20% improvement over the majority-class baseline). Conversely, we observed marked declines in each model's True Veteran Rate, for the same reason.

However, despite these *intra*-model dynamics, where one measure improves and another worsens, the same pattern emerges *between* models as with the ungrouped results. The baseline LDA model still performs the best on average, while the ETM-Pretrained model performs the worst.

Heldout Sample Results (grouped)

Model	Heldout Accuracy	True Rookie Rate	True Veteran Rate	Rookie Precision	Veteran Precision
LDA	**0.6927**	0.7950	0.5815	**0.6737**	0.7230
ETM-Trained	0.6745	**0.8650**	0.4674	0.6384	**0.7611**
ETM-Pretrained	0.6094	0.7500	0.4565	0.6000	0.6269
CatE Trained Replica	0.6719	0.7400	**0.5978**	0.6667	0.6790
CatE Pretrained Replica	0.6458	0.7650	0.5163	0.6322	0.6690
CatE Trained Original	0.6354	0.7800	0.4783	0.6190	0.6667
CatE Pretrained Originlal	0.6198	0.7750	0.4511	0.6055	0.6484
No. obs	*384*				
Heldout size (%)	*15*				
Proportion of Rookies	*0.5208*				
Proportion of Veterans	*0.4792*				

Table 6.3: Classification results based on the grouped heldout sample. This sample is grouped by meeting, speaker then section

6.2 Topic interpretability measures

This section counterposes the classification results above against measures of topic quality. For this study, I use the topic coherence metrics found in Lau et al. (2014) because they provide an automated evaluation framework that closely resembles human judgement.[1] Table 6.4 reports the average topic coherence measures for npmi, pmi and lcp. A subsection of Table 6.1: *Heldout Accuracy* and *True Rookie Rate*, are also included to assist the reader in compare each model's quality to its classifier results.

Overall, the LDA model underperformed the CatE models by an average of around 30% on both npmi and mpi measures, but was only slightly worse than the ETM-Trained model under the lcp measure. That said, the most important message in Table 6.4 is:

"The gold standard LDA model performs poorly in terms of topic coher-

[1]The source code can be found here: https://github.com/jhlau/topic_interpretability.

ence measures. But, it is the *best* performing model on the classification task."

Combined Quality and Classifier Scores

Measure / Model	npmi mean	pmi mean	lcp mean	Heldout Accuracy	True Rookie Rate
LDA (Replica)	0.122	0.359	-1.399	**0.7037**	**0.5417**
ETM-Trained	0.100	0.589	**-1.349**	0.7010	0.4511
ETM-Pretrained	0.018	0.042	-1.929	0.6628	0.3370
CatE-Trained Original	0.145	0.581	-2.436	0.6764	0.4692
CatE-Pretrained Original	0.154	0.621	-2.358	0.6805	0.4855
CatE-Trained Replica	0.161	0.641	-2.311	0.6567	0.4149
CatE-Pretrained Replica	**0.168**	**0.671**	-2.262	0.6601	0.4493

Table 6.4: Topic quality measures represent the average across the 40 topics estimated by each model. The classifier results are an excerpt from the ungrouped classifier results presented in Table 6.1.

The coherence results also highlight these other notable findings. First. The CatE models performed better than the LDA and ETM models under the average npmi and pmi measures. However, it was the opposite under the average lcp measure. This suggests that the LDA and ETM models seem to be better fitting language models, given the data, when compared to CatE. However, this does not guarantee that topical output of such models will be *more* coherent to a human reader. These results echo that of Chang et al. (2009), where those *'models with higher predictive likelihoods do not lead to improved model interpretability'*[p.8].

Second. The ETM-Pretrained model is again a standout poor performer. Its average npmi (0.018) and pmi (0.042) scores are between six and 11 times smaller compared with the scores of the LDA and ETM-Trained models, while its lcp score was not much better than those of the CatE models. It is not clear why the ETM-Pretrained model is doing poorly in terms of topic coherence. This is unexpected and requires

further investigation.

6.3 Summary

With regards to **RQ2**, I find that the ETM and CatE models were unable to beat Hansen et al. (2018)'s gold standard LDA model in terms of Heldout Accuracy results. I find that class imbalance in the training set only tells part of the story in that it does not address the ETM and CatE's *relative* performance gap to LDA.

Regarding **RQ3**, I find that CatE produces the best mean pmi and npmi measures, but produces the worst lcp measures. This finding aligns with Chang et al. (2009) where topic interpretability and model fit diverge.

Overall, topic models that objectively yielded better quality topics (typically the newer models) did *not* perform better than a model which produced poorer quality topics when it came to assessing a *given* downstream task based on a *specific* dataset. For example, CatE produced superior topics based on coherence, but underperformed in the classification task. In contrast, the LDA model did not yield high coherence measures when compared to other models, but is the **best performer** in the classification task.

Lastly, I highlight how the ETM-Pretrained model performed worst overall ranking last across the majority of the assessment metrics. In particular, its poor showing in the pmi and npmi measures were unexpected and requires further investigation.

Chapter 7

Conclusion and Future Research

Research into central bank communications is important because it is a key tool in implementing policy. Policies which have real effect on real economic activity, which in turn, affects the lives of hundreds of millions of people. These communications should be reflective of what those key policy makers communicate among themselves behind closed doors (Acosta, 2015). For the United States, these doors have been opened given the transparency event in late 1993, when all past and future FOMC meeting transcripts were made available to the public. This gave researchers the opportunity to analyse the raw dialogue of the (typically un-elected) key decision makers for arguably the world's largest economy. The importance of understanding how these policy makers think and how they form policy decisions cannot be stressed enough.

One of Hansen et al. (2018)'s original findings was that rookie members and veteran members of the FOMC behave differently, in particular in changing their behaviour during policy meetings after some exogenous shock. One of the ways in which this manifested was through the topics which those members talked about, as well as the way these topics were talked about. This thesis extended the *topic-modelling-central-bank-text* literature, by framing some of Hansen et al. (2018)'s original findings as a machine learning classification task. I trained a machine learning model (a logistic regression classifier) to predict whether an FOMC member is a rookie (or a veteran), by using features derived from their raw utterances as inputs.

Having successfully replicated the original paper, I was able to design an evaluation framework (in the spirit of Doogan and Buntine (2021)) which allowed me to compare the performance of Hansen et al. (2018)'s gold standard model against Dieng et al. (2019b)'s ETM model and Yu et al. (2020)'s CatE model in the bespoke classification task. This extended the literature by evaluating LDA Dieng et al. (2019b)'s ETM and Yu et al. (2020)'s CatE in an applied setting outside of the NLP domain (Doogan and Buntine, 2021).

In answering **RQ2**: "Do the new topic models in Dieng et al. (2019b) and Yu et al. (2020) perform better compared to the gold standard on a bespoke downstream prediction task?" I find that the gold standard model outperformed the newer models overall. I also find that the ETM-Pretrained model performed worst in both relative and absolute terms. Class imbalance was able to explain (some of) its poor performance in absolute terms, but was not able to address the relative performance gap to the other models. This is investigation is left for future work.

In addressing the qualitative results, regarding **RQ1**: "Do the new topic models produce topic rankings in line with Hansen et al. (2018) gold standard model?" I find that, despite some correctly placed topics, the ETM and CatE models do not do a good job in identifying topics which are procyclical or countercyclical as ranked by Hansen et al. (2018)'s procyclicality index. The ETM-Trained model struggled because it returned a large proportion of stop topics, which in their own right, do not belong to either end of the procyclicality index. The ETM-Pretrained yielded less stop topics, but seemed to have jumbled topic rankings. Similarly, CatE only placed a few topics in the correct regions of the spectrum, with a majority topics being out of place or not belonging to a particular end of the business cycle.

In terms of topic interpretability, **RQ3**: "Do the newer topic models produce better quality topics as measured by the metrics found in Lau et al. (2014)?" I find that the CatE models produced the best mean pmi and npmi measures. However, they

produce the worst lcp measures. These findings align with Chang et al. (2009)'s observation that topic interpretability does not necessarily lead to a better fitting model. Furthermore, I find that the ETM-Pretrained does the worst in terms of average pmi and npmi scores, while also being just as bad as the CatE model under the lcp measure. This suggests that the ETM-Pretrained model has poor interpretability *and* is also a poor model fit. The ETM-Pretrained model's interpretability results were unexpected and requires further investigation in future work.

To tie it all together, I find the gold standard LDA model does not perform well with respect to topic coherence measures when compared with Yu et al. (2020)'s CatE and Dieng et al. (2019b)'s ETM. However, the LDA model is the best performer at the classification task. The take-home message from this study is that despite the newer topic models having made improvements (generally) over LDA in terms of measures that NLP researchers are interested in, it does not necessarily translate to better performance in downstream tasks, which arguably non-NLP researchers care more about.

Appendix

A.1 ETM-Trained Correlations

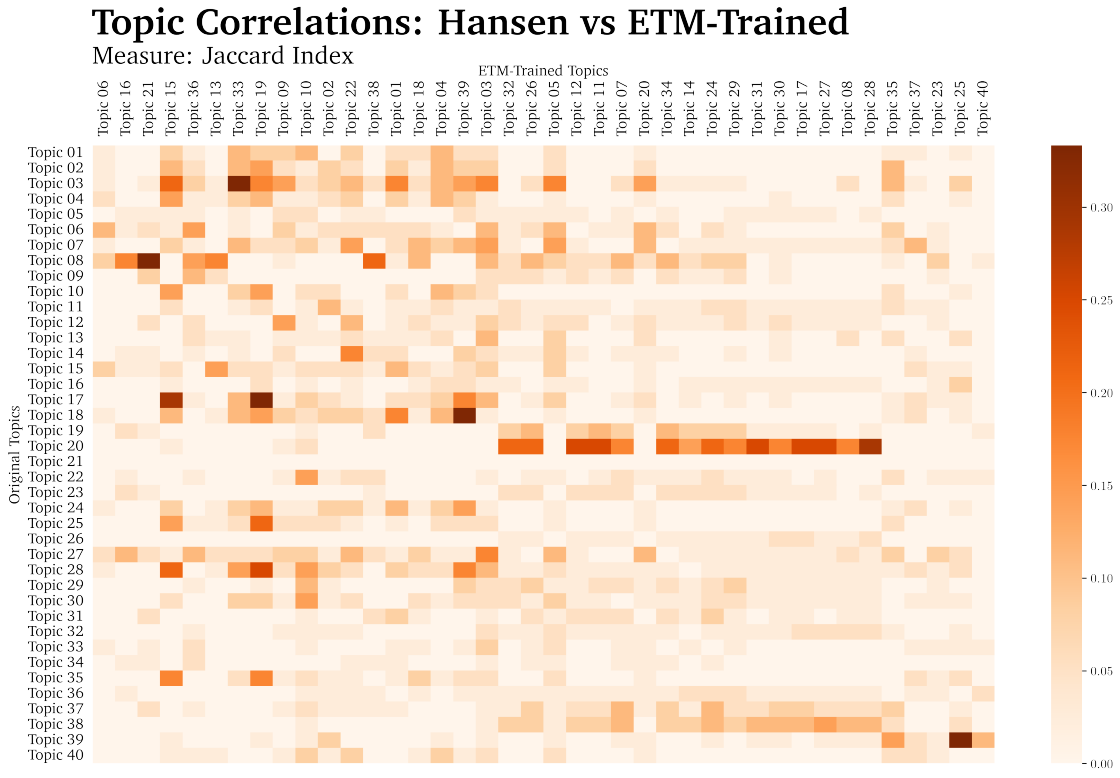


Figure A.1: Jaccard Index measures between benchmark LDA and ETM-Trained topics.

A.2 CatE Trained Replica Full

CatE Trained Replica topics ranked by procyclical index

Topic	Category	Stem 01	Stem 02	Stem 03	Stem 04	Stem 05
Topic 1	vote	abstain	abstent	agreeabl	helen	propos
Topic 2	statement	languag	word	draft	lastsent	sentenc
Topic 3	district	region	nationalaverag	nation	districteconomi	nationallevel
Topic 4	rang	midpoint	lowerend	provision	width	tent
Topic 5	comment	whod	minehan	cathi	echo	remark
Topic 6	percent	percentr	percentagepoint	annualr	percentannualr	half
Topic 7	dollar	greenback	yen	exchangemarket	rupiah	exchanger
Topic 8	forecast	sidestep	greenbookforecast	baselineforecast	inflationforecast	greenbook
Topic 9	inflat	inflationr	inflationexpect	core	riseininfl	disinfl
Topic 10	ve	isn	haven	doesn	wouldn	didn
Topic 11	committe	imprimatur	fomc	explicit	committeememb	public
Topic 12	product	feedstock	stoppag	produc	eaton	capac
Topic 13	report	contact	inbound	broadcast	depot	told
Topic 14	continu	remain	notwithstand	final	moder	sametim
Topic 15	economi	economicact	aggregatedemand	domesticdemand	economicexpans	growth
Topic 16	number	involuntarili	phoni	figur	fluki	seri
Topic 17	price	foodstuff	zinc	commod	nonfood	commoditypric
Topic 18	chang	diddl	inopportun	alter	innocu	differ
Topic 19	risk	downsiderisk	forecastsand	upsiderisk	inflationrisk	balanc
Topic 20	world	codepend	communist	corrupt	saxon	heaven
Topic 21	peopl	infeas	housew	hutt	spous	piano
Topic 22	unitedst	shaghil	abroad	foreign	european	countri
Topic 23	model	workhors	misspecif	equat	frb	econometr
Topic 24	project	prelud	projectionsdetail	projectionperiod	ital	centraltend
Topic 25	period	underw	intermeet	maintenanceperiod	unev	end
Topic 26	signific	substanti	larg	sizabl	diminish	occur
Topic 27	tax	fica	childcar	receipt	medicaid	revenu
Topic 28	shown	rightpanel	leftpanel	panel	middlepanel	middl
Topic 29	polici	toolbox	monetarypolici	policyact	binari	action
Topic 30	oper	fedtrad	openmarket	marketoper	deral	system
Topic 31	thought	fess	toni	wasnt	lindsey	didnt
Topic 32	issu	unstructur	rollout	scrupul	nontranspar	discuss
Topic 33	hous	housingmarket	home	residenti	residentialconstruct	housepric
Topic 34	basispoint	fundsrat	fedfund	ratetarget	ior	today
Topic 35	altern	stylist	alternativeb	section	variant	bluebook
Topic 36	dont	im	shouldnt	doesnt	wouldnt	id
Topic 37	spread	cdx	widen	yield	narrowest	narrow
Topic 38	purchas	program	mb	lsap	taper	asset
Topic 39	financi	financialsector	financialmarket	stress	intermediari	fragil
Topic 40	bank	inelig	uncollater	lend	banksand	fund

Table A.1: CatE Trained Replica topics ranked by the procyclical index (top 5 terms). The category column refers to the seeds that were used as inputs to generate the CatE word groupings.

A.3 CatE Trained Original Full

CatE Trained Original topics ranked by procyclical index

Topic	Category	Stem 01	Stem 02	Stem 03	Stem 04	Stem 05
Topic 1	might	theologian	would	direct	asymmetri	asymmetr
Topic 2	district	districteconomi	manufacturingact	region	nationallevel	nation
Topic 3	forecast	inflationforecast	greenbookforecast	baselineforecast	greenbookproject	outlook
Topic 4	growth	growthrat	gdpgrowth	outputgrowth	realgrowth	gdp
Topic 5	percent	percentagepoint	xa	averag	percentr	roughli
Topic 6	statement	languag	arethat	sentenc	paragraph	word
Topic 7	inflat	core	inflationexpect	inflationr	pceinflat	headlin
Topic 8	dollar	yen	exchanger	depreci	exchangemarket	currenc
Topic 9	product	feedstock	stoppag	capac	labor	produc
Topic 10	employ	jobgrowth	payrol	job	employmentgrowth	payrollemploy
Topic 11	presid	freder	abstain	scheld	eisenbei	governor
Topic 12	know	woven	irrefut	anotherdirect	helluva	pedagog
Topic 13	can	microphon	duct	piano	proctor	decor
Topic 14	chairman	aei	testimoni	semiannu	fra	speech
Topic 15	continu	nonwag	remain	moder	subdu	final
Topic 16	number	consumptionthat	outi	phoni	redbook	data
Topic 17	inventori	inventoryinvest	inventoryaccumul	overhang	finalsal	finaldemand
Topic 18	money	ml	depositori	moneygrowth	opportunitycost	aggreg
Topic 19	month	fewmonth	severalmonth	fewweek	fewday	past
Topic 20	expect	anticip	marketexpect	appar	undoubtedli	impli
Topic 21	report	pult	contact	widespread	preval	told
Topic 22	panel	leftpanel	rightpanel	shown	lowerleft	exhibit
Topic 23	look	fess	lindner	found	studi	analysi
Topic 24	period	perioda	underw	intermeet	protract	mid
Topic 25	project	projectionsdetail	energypricesth	projectionperiod	ital	nextyear
Topic 26	price	foodstuff	zinc	priceswhich	nonfood	commoditypric
Topic 27	will	inept	overturn	betteri	corrupt	polit
Topic 28	weak	weaken	strength	weaker	soften	stronger
Topic 29	risk	downsiderisk	forecastsand	downsid	inflationrisk	upsiderisk
Topic 30	peopl	irration	ould	hutt	ha	manypeopl
Topic 31	invest	softwar	equip	businessinvest	investmentspend	spend
Topic 32	ask	nabe	queri	answer	questionnair	respons
Topic 33	polici	monetarypolici	policyact	toolbox	toolkit	action
Topic 34	sens	unstuck	straitjacket	sort	dogmat	thing
Topic 35	model	workhors	misspecif	equat	frb	econometr
Topic 36	economi	globaleconomi	interdepend	economicact	foreigneconomi	emerg
Topic 37	fundsrat	federalfund	fedfund	nominalfund	ratetarget	target
Topic 38	may	eke	hemorrhag	rebound	recov	gain
Topic 39	thank	recircul	peek	monti	ahin	compliment
Topic 40	bank	inelig	lend	clearinghous	cse	credit

Table A.2: CatE Trained Original topics ranked by the procyclical index (top 5 terms). The category column refers to the seeds that were used as inputs to generate the CatE word groupings.

A.4 CatE Trained Original Correlations

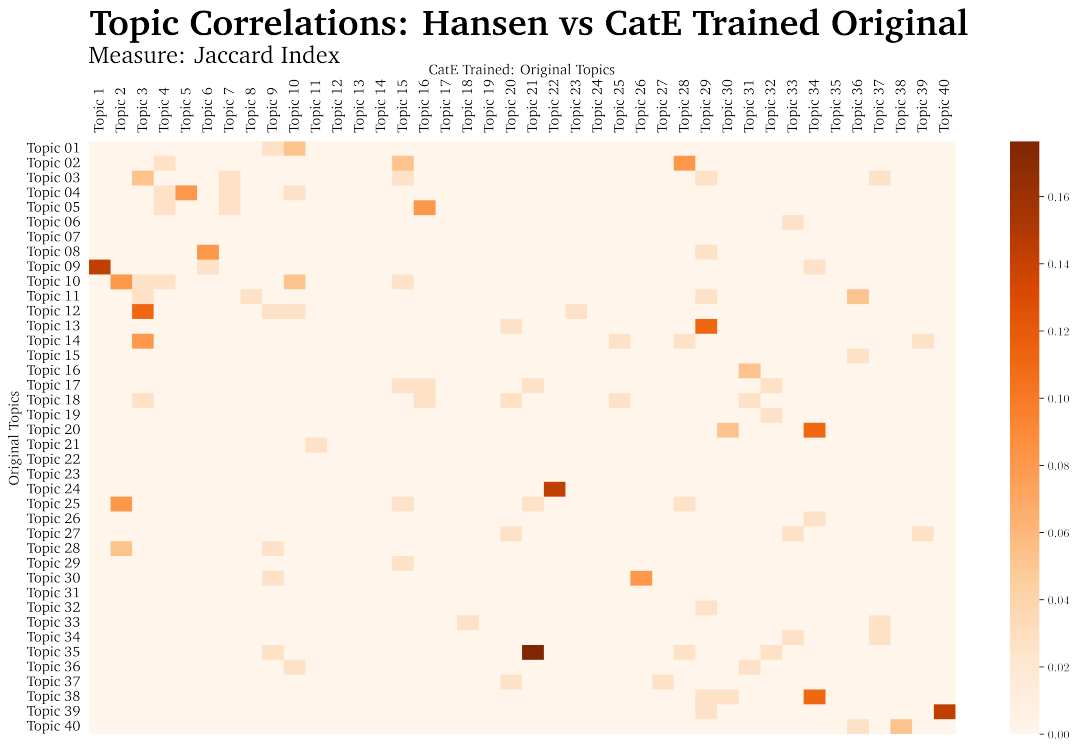


Figure A.2: Jaccard Index measures between benchmark LDA and Cate Trained Original topics.

A.5 Procyclicality Index values

LDA (replica)		ETM-Trained		ETM-Pretrained		CatE Trained Oiginal		CatE Trained Replica	
Topic	procyclicality index	Topic	procyclicality index	Topic	procyclicality index	Topic	procyclicality index	Topic	procyclicality index
Topic 01	0.007814	Topic 06	0.004994	Topic 29	0.000445	Topic 1	0.023828	Topic 1	0.021976
Topic 37	0.003360	Topic 16	0.004008	Topic 26	0.000336	Topic 2	0.008074	Topic 2	0.018804
Topic 09	0.002711	Topic 21	0.003482	Topic 12	0.000279	Topic 3	0.007091	Topic 3	0.012670
Topic 23	0.002434	Topic 15	0.003399	Topic 14	0.000244	Topic 4	0.006530	Topic 4	0.009128
Topic 34	0.001915	Topic 36	0.002842	Topic 23	0.000236	Topic 5	0.005392	Topic 5	0.007691
Topic 06	0.001858	Topic 13	0.002691	Topic 13	0.000204	Topic 6	0.004157	Topic 6	0.006677
Topic 07	0.001678	Topic 33	0.002560	Topic 03	0.000201	Topic 7	0.003046	Topic 7	0.006492
Topic 14	0.001370	Topic 19	0.002158	Topic 22	0.000198	Topic 8	0.003034	Topic 8	0.005904
Topic 21	0.001341	Topic 09	0.002059	Topic 21	0.000188	Topic 9	0.002800	Topic 9	0.005231
Topic 18	0.001267	Topic 10	0.001843	Topic 38	0.000188	Topic 10	0.002786	Topic 10	0.004819
Topic 08	0.001222	Topic 02	0.001842	Topic 09	0.000135	Topic 11	0.002718	Topic 11	0.004596
Topic 13	0.000477	Topic 22	0.001507	Topic 28	0.000133	Topic 12	0.002245	Topic 12	0.003832
Topic 16	0.000472	Topic 38	0.001163	Topic 02	0.000112	Topic 13	0.002127	Topic 13	0.002750
Topic 25	0.000393	Topic 01	0.000542	Topic 07	0.000100	Topic 14	0.001949	Topic 14	0.002224
Topic 28	0.000351	Topic 18	0.000409	Topic 04	0.000091	Topic 15	0.001724	Topic 15	0.001539
Topic 05	0.000311	Topic 04	-0.000097	Topic 11	0.000068	Topic 16	0.001101	Topic 16	0.001096
Topic 03	0.000275	Topic 39	-0.000136	Topic 01	0.000065	Topic 17	0.000547	Topic 17	0.000864
Topic 04	0.000266	Topic 03	-0.000176	Topic 10	0.000057	Topic 18	0.000473	Topic 18	0.000698
Topic 22	0.000230	Topic 32	-0.000307	Topic 18	0.000038	Topic 19	0.000011	Topic 19	0.000689
Topic 39	0.000218	Topic 26	-0.000341	Topic 33	0.000037	Topic 20	-0.000136	Topic 20	0.000669
Topic 20	0.000111	Topic 05	-0.000347	Topic 19	0.000034	Topic 21	-0.000155	Topic 21	0.000468
Topic 36	-0.000030	Topic 12	-0.000358	Topic 30	0.000033	Topic 22	-0.000215	Topic 22	0.000369
Topic 29	-0.000067	Topic 11	-0.000365	Topic 20	0.000014	Topic 23	-0.000435	Topic 23	0.000293
Topic 30	-0.000102	Topic 07	-0.000366	Topic 36	0.000001	Topic 24	-0.000481	Topic 24	-0.000141
Topic 32	-0.000160	Topic 20	-0.000392	Topic 17	-0.000041	Topic 25	-0.000582	Topic 25	-0.000195
Topic 31	-0.000334	Topic 34	-0.000396	Topic 39	-0.000069	Topic 26	-0.000620	Topic 26	-0.000451
Topic 19	-0.000374	Topic 14	-0.000410	Topic 24	-0.000070	Topic 27	-0.000639	Topic 27	-0.001157
Topic 12	-0.000425	Topic 24	-0.000432	Topic 35	-0.000097	Topic 28	-0.001499	Topic 28	-0.001460
Topic 27	-0.000508	Topic 29	-0.000447	Topic 32	-0.000110	Topic 29	-0.001836	Topic 29	-0.001723
Topic 15	-0.000512	Topic 31	-0.000450	Topic 31	-0.000121	Topic 30	-0.002238	Topic 30	-0.002407
Topic 38	-0.000517	Topic 30	-0.000455	Topic 06	-0.000125	Topic 31	-0.002244	Topic 31	-0.002534
Topic 10	-0.000620	Topic 17	-0.000487	Topic 34	-0.000143	Topic 32	-0.002272	Topic 32	-0.003110
Topic 33	-0.001603	Topic 27	-0.000510	Topic 27	-0.000147	Topic 33	-0.002321	Topic 33	-0.003319
Topic 24	-0.001633	Topic 08	-0.000531	Topic 08	-0.000148	Topic 34	-0.003023	Topic 34	-0.003521
Topic 11	-0.002382	Topic 28	-0.000557	Topic 25	-0.000155	Topic 35	-0.003178	Topic 35	-0.004517
Topic 17	-0.002713	Topic 35	-0.000738	Topic 16	-0.000232	Topic 36	-0.003233	Topic 36	-0.008842
Topic 40	-0.002723	Topic 37	-0.002530	Topic 40	-0.000310	Topic 37	-0.005262	Topic 37	-0.014474
Topic 35	-0.003830	Topic 23	-0.003005	Topic 15	-0.000342	Topic 38	-0.005404	Topic 38	-0.017344
Topic 26	-0.004981	Topic 25	-0.008183	Topic 37	-0.000426	Topic 39	-0.005517	Topic 39	-0.020755
Topic 02	-0.006558	Topic 40	-0.013481	Topic 05	-0.000901	Topic 40	-0.038341	Topic 40	-0.033529

Table A.3: This table shows the topic numbers and their corresponding procyclicality index. CatE Pretrained values are omitted.

Bibliography

Acosta, M. (2015). FOMC Responses to Calls for Transparency. Working Paper.

Ahern, K. R. and Sosyura, D. (2015). Rumor has it: Sensationalism in Financial Media. *The Review of Financial Studies*, 28(7):2050–2093.

Antweiler, W. and Frank, M. Z. (2004). *The Journal of Finance*, 59(3):1259–1294.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(6):1137–1155.

Bernanke, B. (2020). The new tools of monetary policy. American Economic Association Presidential Address, January 4.

Bholat, D. M., Hansen, S., Santos, P., and Schonhardt-Bailey, C. (2015). Text mining for central banks.

Blanchard, O. and Sheen, J. (2013). *Macroeconomics*. Pearson Education Australia, Melbourne.

Blei, D. and Lafferty, J. (2006). Dynamic Topic Models. In *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, volume 2006, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blinder, A., Goodhart, C., Hildebrand, P., Lipton, D., Wyplosz, C., and Lawler, P. (2002). How do central banks talk? *World economy*, 25(7):1040–1041.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence. Working Paper.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-K Text to Gauge Financial Constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Boukus, E. and Rosenberg, J. (2006). The Information Content of FOMC minutes. Working Paper.

Bowles, S. and Carlin, W. (2020). What students learn in economics 101: Time for a change. *Journal of Economic Literature*, 58(1):176–214.

Boyd-Graber, J., Hu, Y., and Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296.

Calomiris, C. W. and Mamaysky, H. (2019a). How News and its Context Drive Risk and Returns Around the World. *Journal of Financial Economics*, 133(2):299–336.

Calomiris, C. W. and Mamaysky, H. (2019b). Monetary policy and exchange rate returns: Time-varying risk regimes. Technical report, Cambridge, Mass., USA.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, pages 288–296.

Chen, X., Hu, X., Shen, X., and Rosen, G. (2010). Probabilistic topic modeling for genomic data interpretation. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 149–152.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy Prices. *Journal of Finance*, 75(3):1371–1415.

Cross, J. P. and Greene, D. (2020). Talk is not cheap: Policy agendas, information

processing, and the unusually proportional nature of European Central Bank communications policy responses. *Governance*, 33(2):425–444.

Deerwester, S., Dumais, S., Furnas, G., and Landauer, T. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2019a). The Dynamic Embedded Topic Model. *arXiv preprint arXiv:1907.05545*.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019b). Topic Modeling in Embedding Spaces. *CoRR*, abs/1907.04907.

Doogan, C. and Buntine, W. (2021). Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848.

Fang, L. and Peress, J. (2009). Media Coverage and the Cross-section of Stock Returns. *The Journal of Finance*, 64(5):2023–2052.

Fligstein, N., Brundage, J. S., and Schultz, M. (2014). Why the federal reserve failed to see the financial crisis of 2008: The role of macroeconomics as a sense making and cultural frame. Institute for research on labor and employment, working paper series, Institute of Industrial Relations, UC Berkeley.

Girolami, M. and Kabán, A. (2003). On an equivalence between plsi and lda. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, page 433434, New York, NY, USA.

Association for Computing Machinery.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences - PNAS*, 101(Suppl 1):5228–5235.

Hansen, S. and McMahon, M. (2016). Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. *Journal of International Economics*, 99(S1):S114–S133.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hansen, S., McMahon, M., and Tong, M. (2019). The Long-Run Information Effect of Central Bank Communication. *Journal of Monetary Economics*, 108:185 – 202.

He, C., Zhuo, T., Ou, D., Liu, M., and Liao, M. (2014). Nonlinear compressed sensing-based lda topic model for polarimetric sar image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(3):972–982.

Hendry, S. (2012). Central Bank Communication or the Media' s Interpretation : What Moves Markets ? *Bank of Canada*, (February):1–34.

Hendry, S. and Madeley, A. (2010). Text mining and the information content of bank of canada communications. *Available at SSRN 1722829*.

Hoberg, G. and Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 5057, New York, NY, USA. Association for Computing Machinery.

Holmström, B. (1999). Managerial Incentive Problems: A Dynamic Perspective. *The Review of Economic Studies*, 66(1):169–182.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):927.

La Rosa, M., Fiannaca, A., Rizzo, R., and Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC bioinformatics*, 16(6):1–9.

Lau, J. H., Baldwin, T., and Newman, D. (2013). On collocations and topic models. *ACM Trans. Speech Lang. Process.*, 10(3).

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 530–539.

Lindsey, D. E. (2003). A modern history of fomc communication: 1975-2002.

Lindstedt, N. C. (2019). Structural topic modeling for social scientists: A brief case study with social movement studies literature, 20052017. *Social Currents*, 6(4):307–318.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22.

Loughran, T. and McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Meade, E. E. and Stasavage, D. (2008). Publicity of debate and the incentive to dissent: Evidence from the us federal reserve*. *The Economic Journal*, 118(528):695–717.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262272, USA. Association for Computational Linguistics.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 362369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Moniz, A. and de Jong, F. (2014). Predicting the impact of central bank communications on financial market investors interest rate expectations. In *The Semantic Web: ESWC 2014 Satellite Events*, page 144155, Berlin, Heidelberg. Springer-Verlag.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June):100–108.

Pan, Z., Liu, Y., Liu, G., Guo, M., and Li, Y. (2015). Topic network: topic model with deep learning for image classification. In *International Conference on Knowledge Science, Engineering and Management*, pages 525–534. Springer.

Park, Y.-E. (2020). Uncovering trend-based research insights on teaching and learning in big data. *Journal of Big Data*, 7(1):1–17.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Srivastava, A. and Sutton, C. (2018). Variational inference in pachinko allocation machines. *CoRR*, abs/1804.07944.

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., and Candelieri, A. (2021). OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of finance*, 62(3):1139–1168.

Tetlock, P. C. (2011). All the News That's Fit to Reprint: Do Investors React to Stale Information? *The Review of Financial Studies*, 24(5):1481–1512.

Tetlock, P. C. (2014). Information Transmission in Finance. *Annual Review of Financial Economics*, 6(1):365–384.

Tetlock, P. C., SaarTsechansky, M., and Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance*, 63(3):1437–1467.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the*

2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, page 173180, USA. Association for Computational Linguistics.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Windsor, C. (2021). the Intellectual Ideas Inside Central Banks: What'S Changed (or Not) Since the Crisis? *Journal of Economic Surveys*, 35(2):539–565.

Yu, M., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., and Han, J. (2020). Discriminative Topic Mining via Category-Name Guided Text Embedding. *arXiv.org*.