Focused Concept Miner (FCM): Interpretable Deep Learning for Text Exploration

Dokyun "DK" Lee, Emaad Ahmed Manzoor, Zhaoqi Cheng
{Dokyun, Emaad, Zhaoqi}@cmu.edu
Carnegie Mellon University*

(Working Draft - Please Do Not Distribute)
Previous Versions June 2018, October 2018, October 2019.
This Version Dec 2019.

Abstract

We introduce the Focused Concept Miner (FCM), an interpretable deep learning text mining algorithm to (1) automatically extract interpretable high-level *concepts* from text data, (2) *focus* the mined concepts to explain user-specified business outcomes, such as conversion (linked to read-reviews) or crowdfunding success (linked to project descriptions), and (3) quantify the correlational relative importance of each concept for business outcomes against one another and to other explanatory variables. Compared to 4 interpretable and 4 prediction-focused baselines that partially achieve FCM's goals, FCM attains higher interpretability, as measured by a variety of metrics (e.g., automated, human-judged), while achieving competitive predictive performance even when compared to prediction-focused blackbox algorithms.

The relative importance of discovered concepts provides managers and researchers with easy ways to gauge potential impact and to augment hypotheses development. We present FCM as a complimentary technique to explore and understand unstructured textual data before applying standard causal inference techniques.

Applications can be found in any setting with text and structured data tied to a business outcome. We evaluate FCM's performance on 3 datasets in e-commerce, crowdfunding, and 20-NewsGroup. Plus, 2 experiments investigate the accuracy-interpretability relationship to provide empirical observations for interpretable machine learning literature along with the impact of focusing variables on extracted concepts. The paper concludes with ideas for future development, potential applications, and managerial implications.

Keywords: Interpretable Machine Learning, Deep Learning, Text Mining, Automatic Concept Extraction, Coherence, Transparent Algorithm, Managerial Exploratory Tool, XAI.

*Equal contributions by all authors. We thank Olivier Toubia, Jey Han Lau, David Blei, Arun Rai, Sudhir K., Carl Mela, Alex Burnap, Panos Ipeirotis, Alan Montgomery, Lan Luo, Gedas Adomavicius, Dinesh Puranam, Joy Lu Tong, George Chen, Mingfeng Lin, Daehwan Ahn, Lizhen Xu, DJ Wu, Saby Mitra, Sri Narasimhan, Doug Chung, Elie Ofek, John A. Deighton, Navid Mojir, and participants of Marketing Science Conference 2018 and 2019, Conference on Information Systems and Technologies 2018, Conference on Digital Marketing and Machine Learning 2018, Advanced Computing in Social Sciences and Businesses Symposium 2019, Choice Symposium 2019, INFORMS 2019, Conference on AI, ML, and Digital Analytics 2019, KRAIS ICIS 2019, seminar audiences at McGill University, Korea Advanced Institute of Science and Technology, Seoul National University, University of Pittsburgh, University of Southern California, University of Minnesota, HEC Paris, and University of Maryland, Georgia Institute of Technology, Harvard University for very helpful comments or conversations that shaped the paper. We gladly acknowledge generous grants by Nvidia for research, Marketing Science Institute Grant # 4000562. All errors are our own. www.fcminer.com will contain project details in the future.

1 Introduction

It is becoming imperative for businesses to efficiently process and understand text data, given that more than 90% of data is estimated to be unstructured (Gantz and Reinsel, 2011), 68% of which is consumer generated (Mindtree, 2017). Content creators like Netflix and Amazon are collecting user content consumption data and feedback reviews to create new shows tailored for success (Wernicke, 2015). Companies like C&A Marketing have entire teams of people who read through reviews on Amazon and eBay to identify consumer needs and use them to create new products (Feifer, 2013). After the decades-long surge of unstructured data, retailers and researchers are getting better at utilizing text data to obtain actionable insights, whether it is consumer preferences extracted from reviews or default probability gleaned from crowdfunding descriptions (Netzer et al., 2019).

Yet, despite the deluge of potentially insightful data, studies show that nearly 80% of enterprises don't know how to manage unstructured data (Rizkallah, 2017), and $3 trillion in value goes uncaptured globally in text data alone (McKinsey, 2016). We believe this is due in part to the lack of text mining methodologies that (1) emphasize the high-level interpretability of automatically extracted concepts (by concept, we mean a singular coherent construct), (2) extract concepts that are directly linked to an existing Y of business importance to explain it, and (3) provide an importance weight to mined concepts in relation to one another and in the context of other structured exploratory variables to gauge their potential economic significance. For a text mining method to be useful for managers, all three requirements must be met. In this paper, we introduce the Focused Concept Miner (FCM), a novel deep learning-based text mining algorithm that (1) inherently increases interpretability of mined concepts as quantified by both human judgements and a coherence metric (Mimno et al., 2011) (a measure from topic modeling literature that measures how singular and coherent a set of keywords is—also called topic interpretability) compared to existing techniques, (2) is focused on mining concepts guided by Y specified by the user, and (3) can provide the *correlational* relative importance of mined concepts compared to user-specified referential X (predictor variables). We outline the algorithm in Algorithm 1 and visualize it in Figure 1. By applying FCM, managers should be able to quickly make sense of and extract insights from a large amount of textual data tied to a business outcome *before* launching a more involved causal inference study for prescriptive policy.

We discuss existing approaches to achieve the partial output described in Algorithm 1 in Section 2, as well as the novel aspects of FCM. Figure 2 presents a flow chart for concept mining and when to use FCM over other methods. We evaluate FCM's performance on a unique dataset that tracks individual-level review reading, searching, and purchasing behaviors on an e-commerce site. Applying FCM to this dataset can extract the product review content read by consumers that is correlated to higher conversion rates and illustrate how this content compares to other consumer browsing and click behaviors, as is commonly used in the industry and research. The results and ex-

Algorithm 1 Focused Concept Miner: Algorithm Overview

Input	(1) *Optional* Structured X (e.g., numerical, categorial) (2) Textual X (corpus) (3) Y of business importance (numerical, categorical) linked to corpus and X
Output	(1) Focus-mined concepts predictive of Y (2) Correlational relative importance of concepts against one another and against structured X (3) predictive model
Features	(1) Improved interpretability of mined concepts (2) Potential new concept finding (3) Focused concepts based on Y (4) Joint estimation of structured X and text that is end-to-end (in one pipeline) (5) No need for pre-defined human-tagged training data (6) Inductive inference for new unseen data (7) Online learning via mini-batch gradient descent

periments show that FCM does indeed provide more human-interpretable concepts, as measured by human-judged metrics and coherence metric in comparison to similar, yet feature-lacking methods. In addition, FCM automatically uncovered concepts from consumer-read product reviews that the consumer behavior literature found to be important for conversion—this provides an instance of external validity. Furthermore, FCM performs better than existing techniques in predicting conversion using both structured X and text data. More importantly, FCM provides the relative importance of mined concepts compared to other X, so that managers may understand the economic importance of the extracted concepts. Lastly, we enhance the model to show at what position in the review-reading process these mined concepts matter more in predicting consumers' conversion decisions. Taken together, we demonstrate how FCM can be applied to extract valuable insights from text to inform managerial practices. We also run FCM on a crowdfunding dataset to demonstrate robustness in Section 5.6 and 20-newsgroup (Appendix E). Additionally, we run series of experiments to investigate the accuracy-interpretability relationship to provide empirical observations. Our contribution is thus a new method for text exploration with a focus on demonstrating method potentials, business use-cases, and experimental results on accuracy-interpretability relationship to add to the interpretable machine learning literature.

In comparison to related methods, FCM excels in extracting coherent focused concepts and predictions due to the following conceptual reasons:

1. **Focused:** Concept mining is guided by a user-specified Y of business importance to extract Y relevant concepts. This seems counterintuitive since it seems to add an additional constraint. On the flipside, text is high dimensional data and providing Y effectively reduces hypotheses space. Focusing by Y refines the task more accurately for the algorithm.
2. **Semantic Similarity Knowledge**: Word representation that learns semantic relationships is used. It also uses both local and global contextual information to focus-mine concepts.
3. **Concept Diversity & Sparsity**: The model forces discovered concepts to be distinct from each other (diversity) and to be pithy (sparsity).
4. **End-to-End**: Focus-mined concepts and other explanatory X are jointly estimated to predict outcome variable in an end-to-end (one pipeline optimization) fashion. This ensures that the model shares information from the beginning to the end and is more efficient.

The end result is a deep learning-based exploratory method specifically constructed for deriving value from textual data in business settings with managers in mind. It performs in one optimization step what may have taken managers many steps worth of text mining and processing tasks, often filled with ad-hoc feature engineering and unclear methods for defining and constructing coherent and interpretable concepts. FCM demonstrates that deep learning approaches, normally associated with a lack of interpretability and considered blackbox, could be utilized to help businesses better understand textual data.

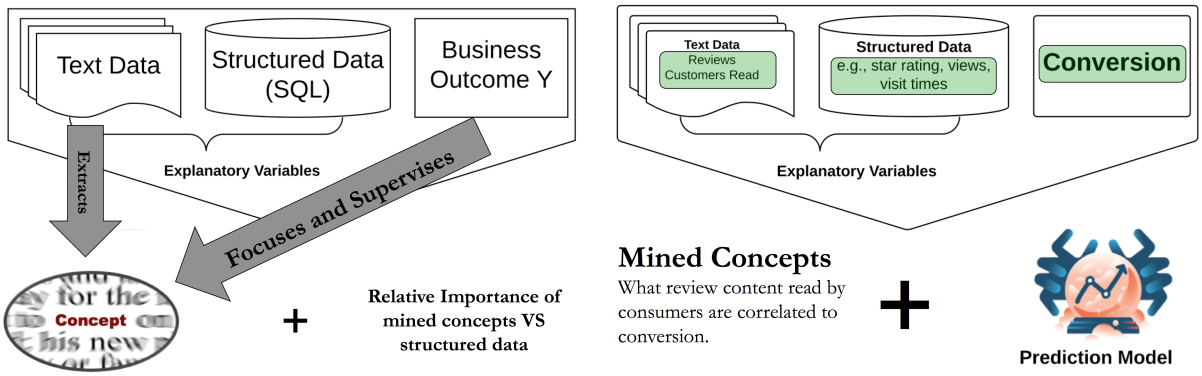


Figure 1: FCM Features Visualized with Use-case Example.

We end this introduction with caveats on what FCM can and cannot do, its envisioned usage, and a description of the use-case presented in this paper. Firstly, FCM is an exploratory and non-causal technique that is predictive in nature[1]. Rather, we see FCM as a complimentary technique to explore and understand a large amount of textual data before zeroing-in on a hypothesis-driven causal study. New concepts extracted from FCM can inform further in-depth causal investigations. FCM could serve as an essential tool to empirically explore and refine exploding volumes of unstructured text data in an empirical-theoretical-empirical-theoretical (ETET)[2] approach to management science, as discussed by Ehrenberg (1994), or as a tool to recover Empirical Generalization[3] in management science, as discussed by Bass (1995). As we will demonstrate, two use cases of FCM include:

- **Consumer-read Content Insights:** Figuring out what content consumers care about in product reviews before making purchases. Reviews read are the text data, product-level and consumer-level data are the referential X, and conversion is the Y. FCM discerns what content consumers care about and the quantifies concept correlation to conversion.

- **Successful Crowdfunding Content**: Exploring the type of well-funded crowdfunding project, such as on Kickstarter or Donorchoose. Project descriptions are the text data, project details are X, and the funding's success is the Y. FCM can discern which projects are funded more successfully.

Section 2 introduces the interpretable machine learning literature then discusses existing relevant text mining techniques in relation to FCM. Section 3 unfolds the model. Then, we describe the proprietary review dataset used to demonstrate FCM in Section 4. Section 5 touches on different aspects of the algorithm's performance in relation to existing baseline models and other extensions. Accuracy-interpretability experiments are presented in Section 5.5 and Section 5.6 runs FCM on crowdfunding data while-exploring the impact of Y on mined concepts. We discuss limitations and future ideas for FCM in Section 6 and conclude with managerial implications and several hypothetical use-cases of FCM in management science.

[1] Extracted concepts and estimations were robust for +500 different runs. However, causality is beyond the scope of this method. Instead, we discuss the challenges and potential ways to use FCM for causal inference in Section 6.

[2] Ehrenberg (1994) posits that management science needs an empirical-theoretical-empirical-theoretical (ETET) approach in which a researcher would "1) Establish some empirically well-grounded theory and 2) Test the theory more widely, deduce new conjectural theory, test that widely, and continue".

[3] Bass (1995) describe Empirical Generalization as "a pattern or regularity that repeats over different circumstances and that can be described simply by mathematical, graphic, or symbolic methods".

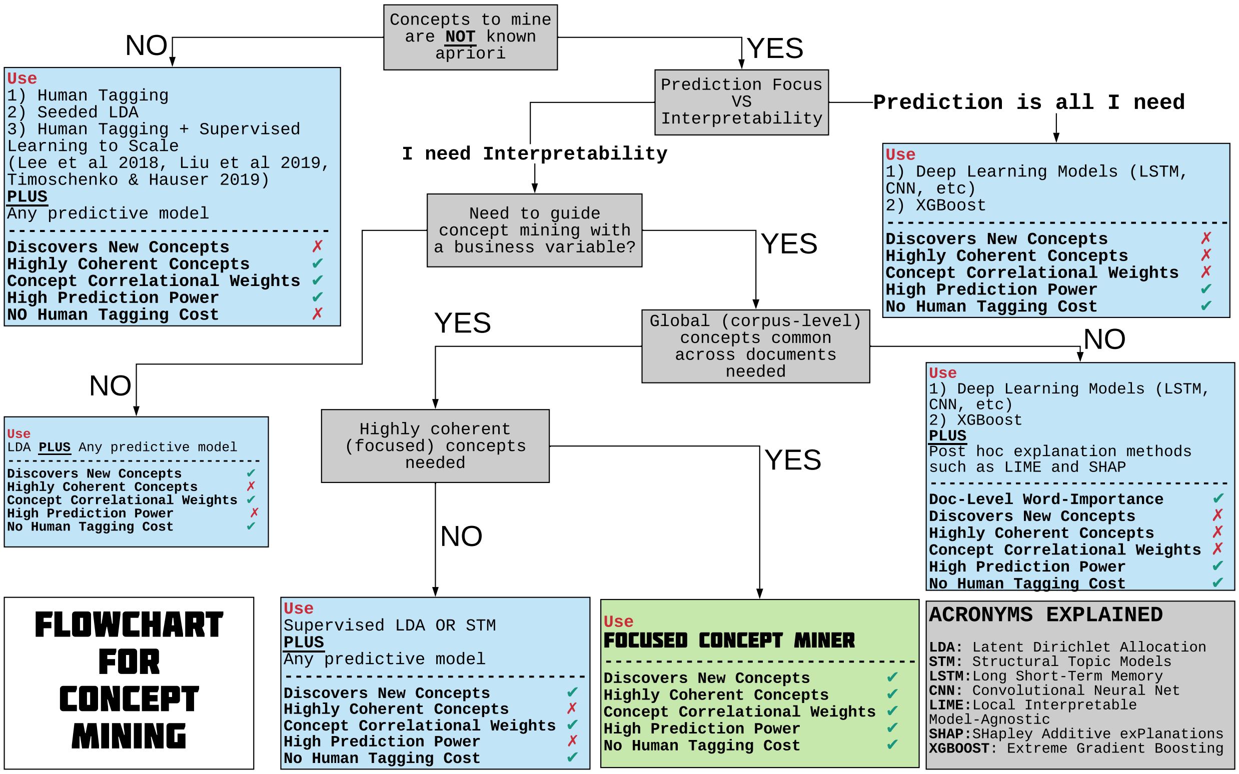

Concepts to mine are NOT known apriori
NO
YES
Use
1) Human Tagging
2) Seeded LDA
3) Human Tagging + Supervised Learning to Scale
(Lee et al 2018, Liu et al 2019, Timoschenko & Hauser 2019)
PLUS
Any predictive model
Discovers New Concepts ✗
Highly Coherent Concepts ✓
Concept Correlational Weights ✓
High Prediction Power ✓
NO Human Tagging Cost ✗
Prediction Focus VS Interpretability
Prediction is all I need
I need Interpretability
Use
1) Deep Learning Models (LSTM, CNN, etc)
2) XGBoost
Discovers New Concepts ✗
Highly Coherent Concepts ✗
Concept Correlational Weights ✗
High Prediction Power ✓
No Human Tagging Cost ✓
Need to guide concept mining with a business variable?
NO
YES
Use
LDA PLUS Any predictive model
Discovers New Concepts ✓
Highly Coherent Concepts ✗
Concept Correlational Weights ✓
High Prediction Power ✗
No Human Tagging Cost ✓
Global (corpus-level) concepts common across documents needed
YES
NO
Use
1) Deep Learning Models (LSTM, CNN, etc)
2) XGBoost
PLUS
Post hoc explanation methods such as LIME and SHAP
Doc-Level Word-Importance ✓
Discovers New Concepts ✗
Highly Coherent Concepts ✗
Concept Correlational Weights ✗
High Prediction Power ✓
No Human Tagging Cost ✓
Highly coherent (focused) concepts needed
NO
YES
FLOWCHART FOR CONCEPT MINING
Use
Supervised LDA OR STM
PLUS
Any predictive model
Discovers New Concepts ✓
Highly Coherent Concepts ✗
Concept Correlational Weights ✓
High Prediction Power ✗
No Human Tagging Cost ✓
Use
FOCUSED CONCEPT MINER
Discovers New Concepts ✓
Highly Coherent Concepts ✓
Concept Correlational Weights ✓
High Prediction Power ✓
No Human Tagging Cost ✓
ACRONYMS EXPLAINED
LDA: Latent Dirichlet Allocation
STM: Structural Topic Models
LSTM:Long Short-Term Memory
CNN: Convolutional Neural Net
LIME:Local Interpretable Model-Agnostic
SHAP:SHapley Additive exPlanations
XGBOOST: Extreme Gradient Boosting

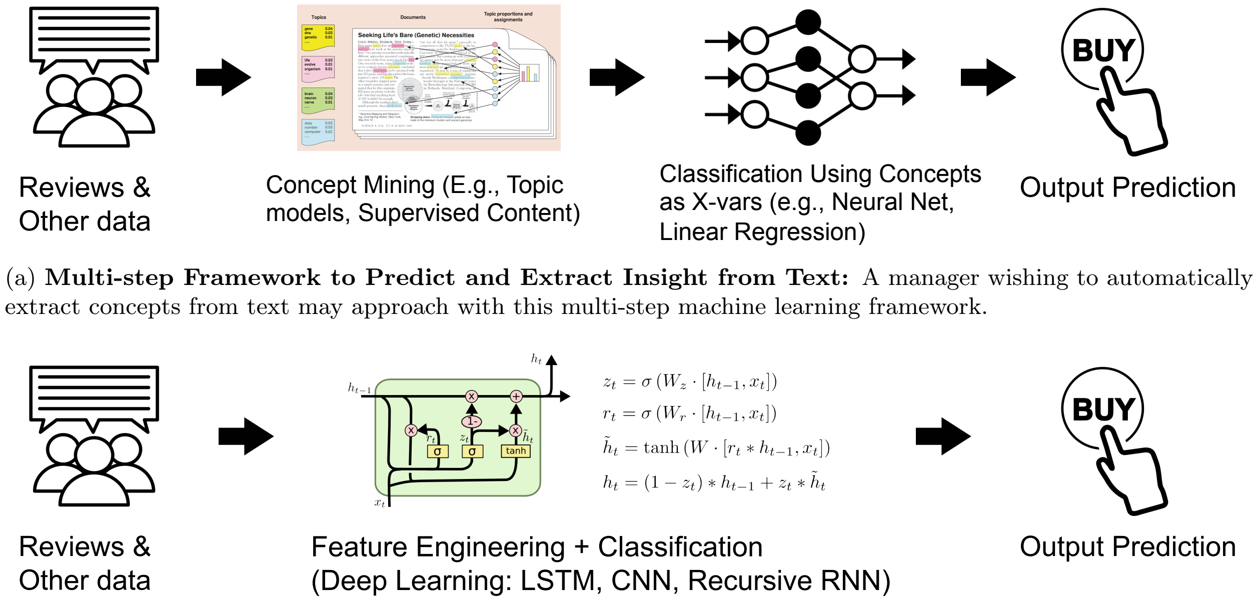
Figure 2: Flowchart for Concept Mining and When to Use FCM. See Section 2 for Literature.

2 Literature

Broadly, the task described in Algorithm 1 is applicable in many settings where text data is linked to a business outcome. For clarity, take for example a consumer purchase (Y-var) related to consumer behavior on the web (X-var) and the reviews the consumers read as a text input. The manager may want to 1) predict conversion from user behavior and reviews read, 2) investigate what behavior (X-vars) may predict conversion, and 3) delve deeper into the review text, to understand what content may be highly correlated to conversion. FCM can be used for all of these tasks, but its focus is primarily on the 3rd problem. Normally, deriving insights from text is tackled with a multi-step approach involving several different techniques, as discussed in the next section. We conceptually outline a popular supervised learning framework to extract the economic importance of text in business data as shown in Lee et al. (2018) and Liu et al. (2019). Figures 3 show two approaches. The first is a multi-step approach, Figure 3a, which emphasizes text insight extraction—the 3rd problem. One would first apply concept mining algorithms such as Latent Dirichlet Allocation or aspect mining (to be discussed in the next subsection), which aim to reduce and extract topics from text either supervised or unsupervised. Mined concepts can then enter any classification or regression framework as X for prediction and correlational relative importance. As discussed next and in the results section, this framework suffers from poor interpretability, lack of focus by the Y, and subpar prediction accuracy when compared to FCM. Next, a manager that prioritizes high predictive performance may apply recent advances in deep learning (Figure 3b), such as long short-term memory (LSTM) which excel at processing sequential data or convolutional neural nets (CNN) which recover local-level feature patterns to aid prediction (for details, see Goldberg (2016)).

However, this approach offers no insight from the text on why conversion might have happened. For further interpretable insights, post hoc processing to open up the blackbox must be applied such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017). We defer the discussion of this methods to Section 2.1, but Guidotti et al. (2018) offers a great survey. However, this multi-step approach often results in fragmented local-level concepts specific to individual data point (i.e., specific to a document and not corpus), unlike FCM—that is, users cannot get *high-level global concepts* from any post hoc explanation method on deep learning.

In Section 2.1, we discuss what it means for an algorithm to be more "interpretable" by introducing the XAI (eXplainable Artificial Intelligence) and interpretable machine learning literature. We discuss several definitions of interpretability and choose operationalizable ones to come up with metrics to compare FCM against existing techniques. Next, we discuss two relevant literatures that provide competing techniques from computer science (pertaining to extracting content and concepts from text data) and similar applications (content extraction from text) from business. We point out the conceptual differences of FCM compared to other techniques. We defer the technical details to Section 3 and interpretability measurement details to Section 3.7.


Reviews & Other data
Concept Mining (E.g., Topic models, Supervised Content)
Classification Using Concepts as X-vars (e.g., Neural Net, Linear Regression)
BUY
Output Prediction
$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$
$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$
$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$
$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$
Reviews & Other data
Feature Engineering + Classification (Deep Learning: LSTM, CNN, Recursive RNN)
BUY
Output Prediction

(a) **Multi-step Framework to Predict and Extract Insight from Text:** A manager wishing to automatically extract concepts from text may approach with this multi-step machine learning framework.

(b) **Deep Learning Framework to Predict Conversion:** A manager who wants high predictive performance may apply recent advances in deep learning. However, this offers no insight from the text on why conversion might have happened. For further interpretable insights, post hoc processing to open up the blackbox must be applied. Guidotti et al. (2018) surveys different methods to do so. However, this approach results in fragmented concepts, unlike FCM. Middle LSTM graphic taken from https://colah.github.io/

Figure 3: **Conceptual Frameworks to Predict Outcome and Derive Insights from Text.**

2.1 Interpretable Machine Learning (ML)

The success of high-performing blackbox algorithms such as deep neural networks (LeCun et al., 2015) and boosted trees (Chen and Guestrin, 2016) is well documented. For example, top-performing algorithms in data science competitions, such as those on Kaggle.com, are almost always one of the mentioned two.[4] The current challenge, however, is to help humans to fully understand these algorithms. These blackbox algorithms are too complicated for humans to fully discern why certain predictions were made. These algorithms do not give any rationale for predictions and attempting to figure them out is prohibitively expensive, if not impossible, due to their sheer intricacy. This is an especially significant issue when deployed on business intelligence systems that deal with consumer data, such as in automated credit approval and investments where auditability, liability, privacy, and other high-stake issues are entangled. In fact, the EU implemented GDPR (General Data Protection Regulation) as of May 2018 to protect consumers and to encourage algorithmic transparency among firms while in that same year DARPA announced $2 billion in initiatives for building the next generation of XAI (eXplainable Artificial Intelligence).

In most business settings or in non-trivial situations, understanding why algorithms made certain predictions is critical to prevent egregious failures, justify usage, improve efficiency, and to ultimately use for decision making. Blackbox failures have been well documented. Angwin et al. (2016) and Wexler (2017) report cases of algorithmic racial bias in bailout decisions (stemming from biased training data obfuscated by the opaqueness of the algorithm), and even instances where the algorithms incorrectly denied parole. Egele et al. (2012) report training a deep net to distinguish regular code from malware. The system ended up picking up a signal of badly written comments in the code rather than actual code content. Zech et al. (2018) report training a deep vision net in the context of medical disease prediction based on x-rays. The system keyed on the meta-tagged word "portable"—reflective of where the samples came from—instead of a valid signal for disease. Additional consequences continue to pile up as blackbox algorithms are utilized without interpretability.

In response to the need for interpretability in machine learning algorithms, several sub-streams of research have boomed since mid-2010s (Please see Guidotti et al. (2018); Gilpin et al. (2018) for surveys). The stream most related to our work is the XAI literature, which broadly defines (Rudin, 2019) two different algorithm families for interpretability.

Definition 1 (Explainable Machine Learning): Given a blackbox predictor B and a training dataset $D = \{X, Y\}$, the explainable machine learning algorithm takes as an input a blackbox B and a dataset D, and returns a transparent predictor T with requirements that 1) T replicates the prediction of blackbox predictor B with high fidelity, and 2) T offers human-understandable rationale for each prediction either at the instance-level or model-average level. T may be a shallow tree, small set of rules, or linear regression with not too many explanatory variables.

Definition 2 (Interpretable Machine Learning): Interpretable machine learning algorithms refer to inherently transparent algorithms that provide human-understandable rationale for predictions yet still offer competitive performances compared to prediction-focused blackbox algorithms.

In this framework, our paper falls under the category of interpretable machine learning algorithms.

While the XAI literature has grown significantly for the last five years and will continue to do so, the definition of "interpretability" still remains an illusive, fragmented, and domain-specific notion (Rudin, 2019; Lu et al., 2020) left to the researcher and user to define. Lipton (2016) states "Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept." A recent comprehensive survey of XAI literature, Guidotti et al. (2018), concludes by stating that "One of the most important open problems is that, until now, there is no agreement on what an explanation is." There have been several attempts to define this. To briefly paraphrase a few sampled works, Doshi-Velez and Kim (2017) state "Interpretability is the degree to which a human can consistently

[4] According to Kaggle.com co-founder Anthony Goldbloom. https://www.linkedin.com/pulse/lessons-from-2mm-machine-learning-models-kagglecom-data-harasymiv/

predict the model's result", Miller (2018) state "Interpretability is the degree to which a human can understand the cause of a decision", and Dhurandhar et al. (2017) state "AI is interpretable to the extent that the produced interpretation I is able to maximize a user's target performance". Few papers also tackle desiderata for interpretability conceptually, such as unambiguity (input and outputs are clear), selectiveness (a parsimonious explanation that does not overwhelm the user), contrastiveness (a "had input been x, output would have been y" type of explanation), factative (has to be highly truthful), etc. (Lipton, 2016; Doshi-Velez and Kim, 2017; Miller, 2018; Ras et al., 2018; Murdoch et al., 2019; Lu et al., 2020).

In this paper, we incorporate insights from XAI literature, as well as interpretability in topic modeling literature, to propose to measure the "interpretability" by 1) tapping into existing interpretability measurements in topic modeling called coherence, 2) confirming the coherence measure with human-judgement directly from mechanical turk, and 3) operationalizing the definition by Dhurandhar et al. (2017) into our problem-specific metric and measuring it directly with human subjects from mechanical turk. We discuss the measures in Section 3.7.

2.2 Finding Concepts in Text: Topic Modeling & Others

The algorithms of interest related to our method include any method which *partially* achieves the goal outlined in Algorithm 1.

Within machine learning (ML) literature, natural language processing (NLP) is primarily concerned with extracting meaningful concepts from a given text. NLP literature offers several ways to achieve the outcome described in Algorithm 1. An initial approach might be to apply any supervised machine learning algorithm on Input (as defined in Algorithm 1) to treat individual words or n-grams (collections of n contiguous words) as X to predict the Y of business interest. Using any combination of feature selection methods (Chandrashekar and Sahin, 2014), a data scientist may be able to extract several keywords or n-grams that could potentially explain the business outcome. However, these methods usually provide a fragmented list of words without enough coherence to be effective in extracting prescriptive, policy-worthy concepts. Once they have a list of salient and informative words from these analyses, managers must drill down further to manually draw out several coherent concepts, which is a subjective rather than objective procedure. After this, it is still unclear how a manager may be able to gauge the economic impact of singular concepts that can consist of several different key words.

The sub-area of sentiment analysis called aspect-based sentiment analysis may offer tools to partially achieve the goal of Algorithm 1. This analysis is concerned with mining opinions from text about 1) specific aspects of focal products or subjects and 2) sentiment valence of these aspects (Liu, 2012). Specifically, an aspect extraction sub-task can be used to reveal key concepts of the text (please see Pontiki et al. (2016) for task descriptions and relevant papers). Briefly summarized, aspect extraction in product opinion mining utilizes a series of smaller techniques and heuristics to figure out concepts that describe product aspects. For example, most algorithms first identify adjectives in text using part-of-speech tagging and then conduct additional filtering based on informativeness and importance. This set of techniques lacks the features to achieve the goal of Algorithm 1 because 1) the discovered aspects are usually very simple and specific to a product or subject in the text, further requiring managers to manually identify more complex concepts by combining collections of words; 2) aspect extraction cannot be supervised by Y of business importance; 3) these techniques often require domain knowledge to feature engineer and extract out aspects, which defeats the exploratory purpose of identifying new concepts in unexplored text data as outlined in Algorithm 1; and 4) while aspect-based sentiment analysis is usually concerned with product reviews and is thus unclear in how it extends to other texts, FCM goes beyond product reviews and can be applied to any text data.

One of the most relevant and influential bodies of work that focuses on automatically summarizing and extracting concepts from text data is topic modeling literature. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a probabilistic generative model that seeks to identify latent topics that comprise a collection of textual data (corpus). Very briefly described, LDA assumes that a topic is a probabilistic distribution over a finite set of vocabularies and that a document consists of several different topics (specified by the user). Then, the method estimates a hierarchical Bayesian architecture with an expectation-maximization algorithm to converge on document-topic distribution and topic-word distribution. The end result is that a user gets a probability vector of topics for each document (document-topic vector) and a bag-of-words with loadings to describe topics (topic-word distribution).

While LDA is a seminal work, it cannot achieve the goal outlined in Algorithm 1 because it is an unsupervised algorithm. Given prior domain knowledge, a user cannot guide the algorithm to discover certain topics or provide a variable for algorithms to adjust the topics. Seeded LDA by Jagarlamudi et al. (2012) extends the original LDA to enable users to guide the topics based on user-inputted topic words while supervised versions of LDA (Blei and Mcauliffe, 2008; Zhu et al., 2012) modified the original model to guide topics discovered with user-specified variables. Lastly, Structural Topic Model (STM (Roberts et al., 2014)) can both incorporate X and be supervised. Yet these algorithms (1) cannot provide relative importance compared to other explanatory X, (2) do not jointly optimize with the given explanatory X and discovered topics, and, most importantly, (3) lack learning interpretable topics or concepts, as discussed by Chang et al. (2009). In other literature streams, several papers tackle how to increase the semantic-coherence of topic models (see Mimno et al. (2011) for a good introduction). However, these models lack the ability to focus-mine topics from a variable of interest, as mentioned previously. Finally, a handful of recent papers explore deep learning-based models that combine word embedding (Mikolov et al., 2013a) and LDA to collaboratively improve the quality of latent topics and word embedding (Xun et al., 2017), to improve discovered topics and word embeddings via topic-specific word embeddings (Shi et al., 2017), and to jointly learn topic and word embeddings (Moody, 2016). However, these papers are again missing several features, such as 1) the supervision of topic discovery guided by the Y, 2) the joint estimation of structured and unstructured data to predict the Y, 3) inductive inference that enables prediction given a new unseen data, and 4) fall behind FCM in predictive performance as well as interpretability. In summary, we are not aware of any methodologies that achieve the same output as our model.

2.3 Content Extraction Via NLP in Business

Natural language processing has been applied to a variety of different textual data for business insights and applications. Some studies are dedicated to extracting and measuring brand perception, market trends, and marketing efforts from social media data (Netzer et al., 2012; Culotta and Cutler, 2016; Lee et al., 2018), while many are dedicated to extracting content and signals out of customer-generated product review data (Decker and Trusov, 2010; Archak et al., 2011; Lee and Bradlow, 2011). In the context of crowdfunding platforms, Netzer et al. (2019) apply text mining to identify signals for loan default.

Here, we mention papers concerned with methodology or the applications of *automatically* extracting concepts or topics from textual business data. To the best of our knowledge, most papers in business research that seek to automatically extract concepts and topics from text data involve some variation of LDAs. Tirunillai and Tellis (2014) apply LDA to consumer-generated review data to extract latent dimensions of consumer satisfaction. Buschken and Allenby (2016) extend the traditional LDA model to restrict one topic per sentence and achieve better extraction of concept from user-generated product reviews. Puranam et al. (2017) apply LDA in a restaurant review setting to see the impact of health regulation on what consumers talk about and on word-of-mouth. Liu and Toubia (2018) extends LDA and develop Hierarchically Dual Latent Dirichlet Allocation, then apply it in a search setting to infer consumers' content preferences based on their queries on the fly. Toubia et al. (2018) apply positive psychology and a seeded LDA model by Jagarlamudi et al. (2012) in a consumer entertainment consumption setting to predict movie watching behavior.

The extant approaches in the literature—echoing the discussion in Section 2.2— are missing many of the features that we propose with FCM. While LDA methodology is a very useful and seminal work with many benefits, these models are not built to extract concepts from text data guided by user-inputted variables nor to discern the relative importance of X in relation to mined concepts, all of which are important in making sense and extracting actionable insights. More importantly, as we show in Section 5.2, results from these models often yield incoherent and intruded (diffusion of concepts) topics with unclear ways to understand and utilize discovered topics as X.

On the other hand, some papers report success with multi-stage supervised machine learning approaches in which pre-defined key content is first defined and tagged by humans and then used to train NLP algorithms to scale to larger unseen data. In this stream, Lee et al. (2018) utilize a traditional NLP approach to study what content (pre-defined and human-tagged content in informative and brand-personality advertising) companies should post on social media to increase user-engagement (like, comment, share, and click). Timoshenko and Hauser (2018) utilize a deep learning NLP approach to extractively reduce user review data and identify content (pre-defined and human-tagged on informativeness) related to consumer needs. Liu et al. (2019) utilize a deep learning NLP approach to investigate what content (pre-defined and human-tagged content in dimensions of product price and quality) in user reviews influence consumer conversion.[5] These papers require 1) ex-ante knowledge of what content to examine and 2) human-tagged labels on text data to answer particular business questions. In contrast, FCM needs neither and identifies concepts automatically. This is essential when managers do not have strong ex-ante knowledge and want to discover concepts in text highly correlated to a specified Y for exploratory purposes.

3 Model

This section formalizes the notion of a *focused concept* in a neural network model and estimation details.

3.1 Definition of Focused Concept

We begin by defining focused concept. Broadly, a concept is an idea or a construct. Our model is designed to optimize the coherence of the uncovered concepts by associating with each concept a collection of semantically similar keywords that describe the central idea of the concept; for example, a concept associated with the words "beautiful, large, lovely" is essentially one that embodies "aesthetics". Intuitively, individual words form the most basic concept – an atomic concept. Several words together form a more complex concept. The key idea is that concept and word, abstractly speaking, can live in the same space. This particular idea has been successfully utilized to quantify

[5] They also propose a second convolutional neural net-based approach called the "full deep learning model". The goal of this model is purely prediction, and while a post-hoc salient n-gram study was used to visualize what n-grams may have influenced the CNN model prediction, this does not 1) extract coherent concepts guided by a Y-var and 2) cannot provide the relative importance of concept correlation to a Y-var.

Dimensions		**Intermediate Elements**	
D	Number of documents	d	Document index
T	Number of concepts	b_d	Bag-of-words document vector
V	Vocabulary size	w, c	Pivot, context word indices
E	Embedding size	v_w, v_c	Pivot, context word embeddings
k	Window size	$C_k(w)$	Set of context word indices
m	Number of negative samples	$N_m(w)$	Set of negative samples
Learned Parameters		**Loss Weights**	
$\mathbf{E_w}$	Word embedding matrix ($V \times E$)	λ	Dirichlet sparsity strengths
$\mathbf{E_t}$	Concept embedding matrix ($T \times E$)	η	Diversity regularizer strength
$\mathbf{W}$	Document-concept weights ($D \times T$)	ρ	Classification loss strength
θ	Concept-classification weights ($1 \times T$)		
CAN	Concept Allocator Network		

Table 1: Notation

complex concepts such as gender and ethnic stereotypes (Garg et al., 2018) and cultural connotations (Kozlowski et al., 2018) directly from text data using a technique called Word2Vec, which comprise the first layer of FCM and will be elaborated next. FCM builds on this proof-of-concept via a carefully constructed novel architecture, loss function, and regularization terms, to ensure that extracted concepts are both focused by the outcomes in the data, and diverse in the sense that no two concepts have a significant overlap (i.e., the concepts are segregated). The proposed model can be trained on any large corpus of documents and their associated outcomes, and the concepts can be recovered from the model parameters.

Relating to the underlying mathematics of the model, concept is defined as:

Concept a vector in a semantic-similarity-aware vector space. Similar vectors in this space have similar role and meaning in natural language. A concept can be represented by a word or collection of words local to each other.

Focused-Concept a representative vector for a collection of words, where the words in the collection have similar semantics and high correlation to Y. Simply, a concept highly correlated to Y.

Connection to topics in topic modeling literature is simple. Topics are defined as distribution over words and a topic consists of words that co-occur frequently in documents. In connection to topics:

Concept Topic + additional constraint that all words describing this topic are semantically similar.

In the rest of this section, we describe the model by introducing each of its components and their motivation, and finally tie them all together and discuss various forms of data that the model may be trained on. Our notation is summarized in Table 1.

3.2 Embedding Words, Documents, and Concepts

Embedding words. We begin by modeling the distribution of words in each document. We rely on the distributional hypothesis in linguistics (Rubenstein and Goodenough, 1965; Sahlgren, 2008), which states that words used together in similar contexts tend to have similar meanings.[6] Recent

[6] In the words of linguist John Rupert Firth: "You shall know a word by the company it keeps."

models based on this hypothesis have demonstrated state-of-the-art performance on various natural language processing tasks (Mikolov et al., 2013a,b; Pennington et al., 2014). We follow the example of word2vec (Mikolov et al., 2013a), which encodes the distributional hypothesis as a neural network to find vector representations of words such that semantic-spatial relationships are preserved—that is, "similar" words lie nearby in the embedding space. For clarity, we adopt the model derivation and notation from Goldberg and Levy (2014).

To this end, we represent each word w by a real-valued vector v_w of length E, called its *embedding*. If we denote the vocabulary of all words in our corpus of documents by $\mathcal{V}$, the embeddings are stored as rows in a matrix $\mathbf{E_w}$ of dimension $|\mathcal{V}| \times E$. We denote by $\mathcal{C}_k(w)$ the set of *context* words around the *pivot* word w within a symmetric k window size. The *pivot* word w is the center word used to predict the surrounding *context* words $c \in \mathcal{C}_k(w)$. We define the likelihood of a corpus of documents in terms of their words and contexts. Given a corpus (a document d, or collection of documents), its likelihood is defined as:

$$\prod_{d \in \text{corpus}} \prod_{w \in \text{document d}} \prod_{c \in \mathcal{C}_k(w)} p(c|w; \mathbf{E_w}) \tag{1}$$

This is essentially a mathematical formulation of the distributional hypothesis in linguistics: the probability of a document is defined in terms of the conditional probability of each context word c given its corresponding pivot word w. Given a context word c and its pivot word w, we would like to capture the fact that words occurring in the same context often should have similar embeddings. Hence, we parameterize the conditional probability $p(c|w; \mathbf{E_w})$ as follows:

$$P(c|w; \mathbf{E_w}) \quad = \quad \frac{e^{v_c \cdot v_w}}{\sum_{c' \in \mathcal{V}} e^{v_{c'} \cdot v_w}} \tag{2}$$

where v_c and v_w are the embeddings of c and w. Our goal is to learn the embeddings $\mathbf{E_w}$ that maximize the likelihood of the observed data in eq.(1).

However, computing the conditional probability term in eq. (2) involves a computationally expensive summation in the denominator over all possible words in the vocabulary, of which there may be hundreds of thousands. Hence, we approximate this objective via *skip-gram negative sampling* (Mikolov et al., 2013b).

Let $\mathcal{D}$ be the set of all observed pivot-context word pairs in the corpus, and $\mathcal{D}'$ be the set of all pivot-context pairs that do not occur in the corpus. For a given a pivot-context word pair (w, c), let $P((w, c) \in \mathcal{D}|\mathbf{E_w})$ be the probability that this pair occurs in the corpus, and let $P((w, c) \notin \mathcal{D}|\mathbf{E_w}) = 1 - P((w, c) \in \mathcal{D}|\mathbf{E_w})$ be the probability that the pair does not occur in the training corpus. We would like to find $\mathbf{E_w}$ such that the likelihood of the observed pivot-context pairs is maximized, while the likelihood of the unobserved pivot-context pairs is minimized. This is captured by the following objective:

$$\max_{\mathbf{E_w}} \prod_{(w,c) \in \mathcal{D}} P((w, c) \in \mathcal{D}|\mathbf{E_w}) \prod_{(w,c) \in \mathcal{D}'} P((w, c) \notin \mathcal{D}|E_w) \tag{3}$$

We can parameterize the probability $P((w, c) \in \mathcal{D}|\mathbf{E_w})$ using the logistic-sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$ that scales its argument to lie in $(0, 1)$:

$$P((w, c) \in \mathcal{D}|E_w) = \sigma(v_w \cdot v_c) = \frac{1}{1 + e^{-v_c \cdot v_w}} \tag{4}$$

Plugging eq. (4) into eq. (3) and taking logarithms leads to the following objective:

$$
\begin{aligned}
&\max_{\mathbf{E_w}} \log\Big(\prod_{(w,c)\in\mathcal{D}} P((w,c)\in\mathcal{D}|E_w) \prod_{(w,c)\in\mathcal{D}'} P((w,c)\notin\mathcal{D}|\mathbf{E_w})\Big)\\
&= \max_{\mathbf{E_w}} \sum_{(w,c)\in\mathcal{D}} \log(P((w,c)\in\mathcal{D}|\mathbf{E_w})) + \sum_{(w,c)\in\mathcal{D}'} \log(P((w,c)\notin\mathcal{D}|\mathbf{E_w}))\\
&= \max_{\mathbf{E_w}} \sum_{(w,c)\in\mathcal{D}} \log\frac{1}{1+e^{-v_c\cdot v_w}} + \sum_{(w,c)\in\mathcal{D}'} \log\Big(1-\frac{1}{1+e^{-v_c\cdot v_w}}\Big)\\
&= \max_{\mathbf{E_w}} \sum_{(w,c)\in\mathcal{D}} \log\frac{1}{1+e^{-v_c\cdot v_w}} + \sum_{(w,c)\in\mathcal{D}'} \log\frac{1}{1+e^{v_c\cdot v_w}}\\
&= \max_{\mathbf{E_w}} \sum_{(w,c)\in\mathcal{D}} \log(\sigma(v_c\cdot v_w)) + \sum_{(w,c)\in\mathcal{D}'} \log(\sigma(-v_c\cdot v_w)) \qquad (5)
\end{aligned}
$$

The computationally expensive summation over all possible $(w,c)\in\mathcal{D}'$ in the second term can be approximated by summing over m *negatively-sampled* pivot-context pairs $\mathcal{D}_m$. The sampling is performed as follows for every $(w,c)\in\mathcal{D}$: sample $(w,c'_1),(w,c'_2),\ldots,(w,c'_m)$ such that $(w,c'_i)\notin\mathcal{D}$ and each c'_i is drawn with probability proportional to its frequency[7] in the corpus, $P(c'_i)=n(c'_i)/N$ where $n(w)$ is the frequency of word w in the corpus and N is the number of words in the corpus.

Converting the maximization to a minimization problem yields the first component of the *loss function*[8] that our method seeks to minimize:

$$
\mathcal{L}_{\text{neg}} = -\sum_{(w,c)\in\mathcal{D}} \log(\sigma(v_c\cdot v_w)) - \sum_{(w,c)\in\mathcal{D}'} \log(\sigma(-v_c\cdot v_w)) \qquad (6)
$$

where σ is the logistic sigmoid function $\sigma(x)=(1+\exp(-x))^{-1}$ as defined earlier, m is the number of negative samples and k is the window-size. Taking a closer look at this loss function, we observe that the first summation operates over all pivot-context pairs in the corpus to ensure that words occurring often in the same context have similar embeddings. The second term operates over each pivot-context pair that does *not* occur in the corpus, to encourage them to have *dissimilar* embeddings.

Embedding documents and concepts. We now describe an extension of the model to capture "concepts" (collections of coherent words relating to easily interpretable ideas) by combining ideas from Mikolov et al. (2013a) and Moody (2016). We assume the existence of a fixed number of concepts T, and assume that the words of each document are drawn from a distribution over these T concepts. We store the unnormalized form of this distribution (the "concept weights" for each document) in matrix $\mathbf{W}$ of dimension $D\times T$, where D is the number of documents in the corpus. Each concept is represented by its own embedding of length E, stored in a matrix $\mathbf{E_t}$ of dimension $T\times E$; note that the concept embeddings lie in the same space as the word embeddings, which is crucial for each concept to be interpreted by a collection of keywords. Given the concept embeddings $\mathbf{E_t}$ and concept weights $\mathbf{W}$, the embedding of a document v_d can be derived as a weighted linear combination of its concept embeddings, in line with our earlier assumption. We first transform the

[7] Note that Mikolov et al. (2013b) use the frequency exponentiated to 3/4 which provided superior empirical performance.

[8] An alternative method to obtain word embeddings is via factorization of the shifted pairwise mutual information (PMI) matrix Levy and Goldberg (2014). However, formulating the objective this way eliminates the flexibility to extend the model to incorporate several objectives such as concepts, supervision, and diversity, to be discussed later in this section.

document-concept weights $\mathbf{W}[d]$ to a probability distribution p_d, and then use this to weight each of the concept embeddings:

$$\begin{aligned} p_d[i] &= \frac{e^{\mathbf{W}[d][i]}}{\sum_{j=1}^{T} e^{\mathbf{W}[d][j]}} \quad \forall i = 1, \ldots, T && (7) \\ v_d &= \sum_{i=1}^{T} p_d[i] \times \mathbf{E_t}[i] = p_d \times \mathbf{E_t} && (8) \end{aligned}$$

We now need a way to link concepts and words in order to jointly learn their embeddings in the same space, using the loss function given in eq. (6). We do this by linking words with concepts via the documents they occur in. Specifically, we define a "document-specific" word embedding v_{dw} that represents the word w as it appears in the context of document d. For example, the word "ball" in an article about tennis could have a different meaning (and hence, a different embedding) than that in an article about soccer. We define this document-specific word embedding as a *translation* of the original word embedding v_w by the document embedding v_d:

$$v_{dw} = v_d + v_w$$

While any general function $v_{dw} = f(v_d, v_w)$ could have been used to define a document-specific word embedding, simple translation ensures that the loss function remains differentiable and efficient to compute, which are necessary for efficient minimization. The skip-gram negative sampling loss in eq.(6) can now be modified as:

$$\mathcal{L}_{\text{neg}} = -\sum_{(w,c)\in\mathcal{D}} \log(\sigma(v_c \cdot v_{dw})) - \sum_{(w,c)\in\mathcal{D}'} \log(\sigma(-v_c \cdot v_{dw})) \quad (9)$$

where $d \in 1, \ldots, D$ is the index of the document containing w and c. Minimizing this loss now enables us to learn both v_w and v_d (and hence, $\mathbf{E_t}$) from the training corpus.

Predicting concepts for unseen documents. The concept weight matrix $\mathbf{W}$ defined thus far enables learning the concept weights for documents in the available corpus. In some scenarios, we foresee a trained FCM model being used to *predict* the concepts for unseen documents, which were previously unavailable in the corpus. Hence, we now propose an alternative to the fixed concept weight matrix $\mathbf{W}$ to predict concepts for unseen documents. A document d in its raw form can be represented as a bag-of-words vector $b_d \in \mathbb{R}^V$, containing its word-counts or TF-IDF scores (for example). We introduce a new FCM component, the Concept Allocator Network **(CAN),** that takes as input a bag-of-words vector b_d and generates its concept probability distribution p_d. **CAN** is a fully-connected multilayer neural network with H hidden layers, each of size h and *tanh* non-linear activations between its hidden-layers. A *softmax* activation after its final layer transforms its output to be valid probability distribution. Formally, **CAN** is defined in terms of the input layer matrix $\mathbf{M}^{(0)} \in \mathbb{R}^{V\times h}$, hidden layer matrices $\mathbf{M}^{(1)} \ldots \mathbf{M}^{(h-1)} \subset \mathbb{R}^{h\times h}$, and output layer matrix $\mathbf{M}^{(h)} \in \mathbb{R}^{h\times T}$. Each matrix is also associated with a bias vector, $m^{(0)}, \ldots, m^{(h-1)} \subset \mathbb{R}^h$ and $m^{(h)} \in \mathbb{R}^T$. The process of mapping a bag-of-words vector b_d to its concept probabilities p_d is given by the following:

$$\begin{aligned} x^{(0)} &= \tanh(b_d\mathbf{M}^{(0)} + m^{(0)}) && (10) \\ x^{(j)} &= \tanh(x^{(j-1)}\mathbf{M}^{(j)} + m^{(j)}) \quad \text{for}\sim j = 1, \ldots, h-1 && (11) \\ p^d &= \text{softmax}(x^{(h-1)}\mathbf{M}^{(h)} + m^{(h)}) && (12) \end{aligned}$$

where the *softmax* function transforms its vector-valued input into a probability distribution:

$$\text{softmax}(x)[i] = \frac{e^{x[i]}}{\sum_{j=1}^{T} e^{x[j]}} \quad \forall i = 1, \ldots, T$$

Thus, **CAN** replaces the concept weight matrix $\mathbf{W}$ to generate the concept probabilities p_d from the document's bag-of-words vector b_d. Like $\mathbf{W}$, **CAN** is jointly trained with the other FCM parameters. The hidden-layer size and number of hidden layers in **CAN** are hyperparameters that we tune via cross-validation.

Encouraging concept sparsity. In reality, we expect each document to embody only a few concepts, with the others being barely present or completely absent. This *sparsity* of the document-concept distribution is also easier to interpret and inspired by LDA. To enforce sparsity on the document-concept distribution, we append the product of the document-concept probabilities p_d to the loss function, transformed logarithmically to prevent numerical underflow (since the product of many probabilities will be a very small floating point number). This leads to the following "Dirichlet loss" term weighted by hyper-parameter λ, which approximately penalizes the document-concept distributions for having too many non-zero probability values:

$$\mathcal{L}_{dir} = \lambda \log(\prod_{d=1}^{D} \prod_{k=1}^{T} p_d[k]) = \lambda \sum_{d=1}^{D} \sum_{k=1}^{T} \log(p_d[k]) \tag{13}$$

Note that, while penalizing the L_0 norm $\|p_d\|_0 \; \forall d = 1, \ldots, D$ would enforce sparsity exactly, it is non-differentiable, leading to issues when minimizing the loss function using gradient descent. Penalizing the product of the probabilities (or the summation of the log-probabilities) as above serves to approximate the sparsity objective while remaining differentiable and efficient to compute.

Encouraging concept diversity. The model described so far tends to learn concepts that are highly overlapping, especially when a few concepts are significantly more prevalent in the corpus than others. To better capture less prominent but potentially important concepts, we introduce a novel extension that we term the "diversity regularizer" on the model. This regularizer encourages every pair of concept embeddings $\mathbf{E_t}[i], \mathbf{E_t}[j]$ to be dissimilar in terms of their dot-product. This is formulated as the following extension to the loss function:

$$\mathcal{L}_{\text{div}} = \eta \sum_{i=1}^{T} \sum_{j=i+1}^{T} \log \sigma(\mathbf{E_t}[i] \cdot \mathbf{E_t}[j]) \tag{14}$$

where η is a hyper-parameter that controls the strength of the prior, and the logσ log-sigmoid transformation ensures that this term and its gradient lie on the same scale as the other terms in the loss function.

3.3 Focusing Concepts on Target Outcomes

In practice, the concepts embodied by documents may fall into several different descriptive modes. For example, the set of concepts "furniture", "technology", and "kitchen" describe the *category* of product being sold, whereas the set of concepts "aesthetics", "functionality", and "reliability" describe *characteristics* of the product being sold; both these descriptive modes may exist simultaneously in the corpus, and our goal is to uncover the one that best explains the given *outcome* associated with each document.

Hence, we introduce a loss component that "focuses" the concepts toward explaining these outcomes. We assume that the target outcomes are binary, $y_d \in \{0, 1\} \forall d = 1, \ldots, D$, though extensions to real-valued outcomes are straightforward. We introduce a parameter vector $\theta \in \mathbb{R}^T$ that assigns an *explanation-weight* to each concept, that is shared across all documents. Given the explanation weights θ and the document-concept distribution p_d, define $\hat{y_d}$ for a document d as a weighted combination of its concept probabilities:

$$\hat{y_d} = \theta \cdot p_d \tag{15}$$

Given the observed outcome y_d, we would like $\hat{y_d}$ to be large if $y = 1$ and small if $y = 0$. This requirement is captured by the following *cross-entropy loss* term that we append to the overall loss function weighted by hyper-parameter ρ:

$$\mathcal{L}_{\text{clf}} = \rho(y_d \log \sigma(\hat{y_d}) + (1 - y_d) \log(1 - \sigma(\hat{y_d})))$$

Note that we could also add any user-specified X in Equation 15. This 1) increases prediction power and 2) allows managers to compare the correlational relative importance of mined concepts to key-referential X. We discuss this extension in Section 5.4 and in particular Equation 17.

3.4 Model Summary

In each training iteration, the input for the model is the pivot word w, the set of context words from the size-k window $C_k(w)$, the index d of the document and the outcome y. The complete model incorporates all the losses defined in the previous sections, leading to the following loss function for each input iteration:

$$\mathcal{L}(w, C_k(w), d, y) = \mathcal{L}_{\text{neg}}(w, C_k(w)) + \mathcal{L}_{\text{dir}}(d) + \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{clf}}(d, y)$$

To prevent longer documents (those with more words) from contributing disproportionately more to the loss, we also scale it by the length $l_d \in (0, 1)$ of each document which is inversely proportional:

$$\mathcal{L}_{\text{corpus}} = \sum_{d,y \in \text{Corpus}} \sum_{w \in \text{doc } d} \mathcal{L}(w, C_k(w), d, y) \times \frac{1}{l_d}$$

The model architecture is visualized in Figure 4. This diagram describes the raw input to the model, how the input constitutes the matrices to be estimated by the FCM, further processing by the model using the matrices, and final output that enters the loss function to be minimized. This diagram describes how the data flows through the neural network model.

We construct train, validation, and test sets using 70%, 15%, and 15% of the full data, respectively. To improve generalizability, we regularize $\mathbf{W}$ and θ with their L_2 norm, perform dropout on $\mathbf{E}$ and gradient clipping to prevent exploding gradients. We initialize our algorithm with pre-trained word2vec word-vectors, trained in an unsupervised fashion on a corpus of 100 billion words from Google News. We train the model using mini-batch stochastic gradient descent with a batch-size of 10,240 on an Nvidia Titan X Pascal with 12GB of GPU memory. The estimation roughly took 2 hours on this hardware specification to get to 200 epochs.

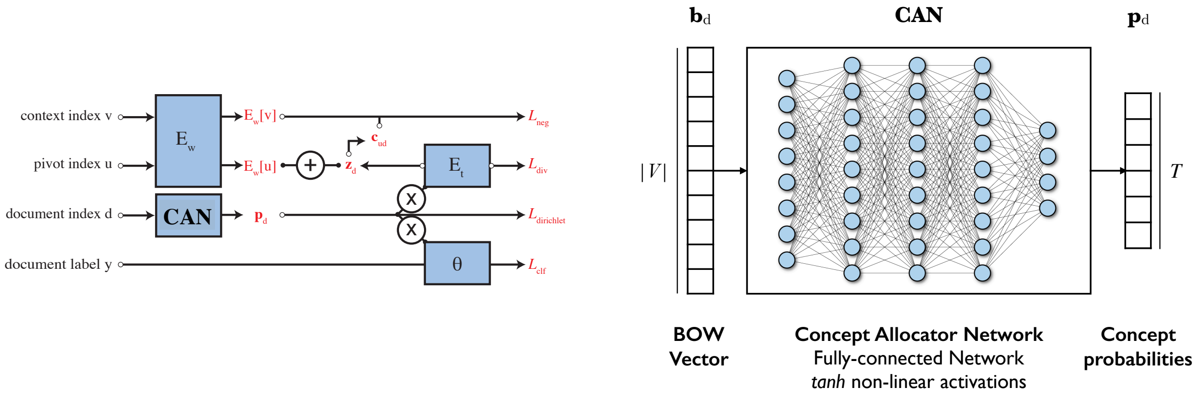


Figure 4: FCM Model Architecture

3.5 Describing Concepts

Given the trained model, one way to describe concepts mined is to 1) get concept embedding vectors and 2) find the closest word vectors for each concept vectors. We take this approach and use the dot product distance. Other generative models such as recurrent neural networks could be utilized to generate sentence-level description of concepts, but this is beyond the scope of this paper.

3.6 Measure of Predictive Performance

To measure the performance of different models, we use a receiver operating characteristics (ROC) curve, which compares the true positive rate (TPR) against the false positive rate (FPR) at different discrimination thresholds. The Area Under the ROC curve is called AUC, which captures the probability that a model ranks a randomly chosen positive sample higher than a randomly chosen negative sample. In general, a classifier with higher AUC tends to have better predictive performance. For example, a random guess classifier yields an AUC of 0.5, while a perfect model yields an AUC of 1. Additionally, we show simple accuracy, precision, recall, and F1 score.[9]

3.7 Measure of Interpretability

To measure the interpretability of model output, we define and use three different metrics as discussed in Section 2.1

1. **Coherence:** This is a measure as defined by Mimno et al. (2011) from the topic modeling literature. This measure computes the sum of a pairwise score function on the top n words $w_1, w_2.., w_n$ used to describe each topic:

$$Coherence = \sum_{i<j} \log \frac{D(w_i, w_j) + 1}{D(w_i)} \tag{16}$$

where $D(w_i)$ is the count of documents containing the word w_i, and $D(w_i, w_j)$ is the count of documents containing both words w_i and w_j. Simply put, the coherence metric measures

[9] Measures are defined as accuracy (the total % correctly classified), precision (out of predicted positives, how many are actually positive), recall (out of actual positives, how many are predicted as positives), and $F1 = \frac{2*Precision*Recall}{Precision+Recall}$ (the harmonic average of precision and recall)(Jurafsky, 2000).

how well-focused the group of top topic key words are in describing a singular concept. A higher topic coherence means that the key words within one topic dimension are more coherent with each other in concept. Note that while the measure of coherence originates from topic modeling literature, it is directly applicable to any set of keywords—it simply measures how coherent a set of keywords is in describing a singular concept. The topic modeling literature first came up with ways to detect "interpretability" by human judgement through mechanical turk (Chang et al., 2009), then the "coherence" construct was validated with domain expert taggers who "annotated [topics] as 'good' if they contained words that could be grouped together as a single coherent concept" (Mimno et al., 2011). Next, automated measures were constructed that seem to perform as well as or better than humans (Mimno et al., 2011; Newman et al., 2010; Lau et al., 2014). For example, metrics based on word co-occurrences and mutual information based on an external corpus such as Wikipedia are more representative of how humans would evaluate a topic as interpretable (Newman et al., 2010). From the XAI literature perspective, this measure of interpretability fits the desiderata of unambiguity (Ras et al., 2018) and selectivity of explanation (Lipton, 2016; Miller, 2018).

2. **1st Human-judged - Number of Distinct Concepts Found**: Using Amazon Mechanical Turk, we directly obtain the number of distinct concepts found in algorithm outputs. Each topic or concept would be more useful and interpretable if it describes fewer distinct concepts. This parallels the coherence measure but is more direct.

3. **2nd Human-judged - Usefulness of Algorithm Output for the Particular Target Task**: We use the definition of interpretability from Dhurandhar et al. (2017)—which states "AI is interpretable to the extent that the produced interpretation I is able to maximize a user's target performance"—and operationalize it in our case. As the next section elaborates, the data context is the consumer purchase journey and reviews read. Thus, the target goal here is making a purchase decision. Therefore, if we apply the definition of Dhurandhar et al. (2017), the algorithm output that extracts concepts from reviews that are more helpful for making a purchase decision should be considered more interpretable. We ask Amazon Mechanical Turkers to provide the usefulness of algorithm outputs for making a purchase decision.

4 Demonstration of FCM on a Novel Data

To demonstrate the efficacy of FCM, we apply it to a proprietary novel dataset from a top consumer review platform. We explain the data here.

4.1 Raw Data

Our data comes from an online retailer in the United Kingdom through a top review platform company. They track 243,000 consumers[10] over the course of two months in February and March of 2015 in the electronics and home & garden categories. There are 41 different subcategories, as shown in Appendix A. The data tracks consumer page views, review-reading behavior, clicks, and transactions. That is, the data includes the typical clickstream and conversion data plus consumers' review-reading behaviors, which is essential for FCM application. The data also records (1) when a

[10] The users are assigned an anonymous unique identifier, which enables us to effectively analyze customer browsing habits and transaction activity at the individual level. The user identifier persists in the third-party network cookie and lasts for up to 18 months. Even if a user has cleared their cookies or is browsing on another device, we can still identify them through an encrypted IP address.

user clicks on review pages, (2) whether each review has appeared on a user's browser, and (3) for how long the content was viewed on the user's browser, measured accurately down to milliseconds. With these data, we assume that if a review appeared on a user's browser, the user has read the review.

4.2 Processed Data for FCM

From the perspective of a user shopping procedure, we can imagine a "decision-making journey" that characterizes how a user purchases or abandons a particular product. In such a journey, a user will first visit the product page to read the description, then read reviews of the product, and finally decide whether to buy the product or not. Accordingly, our dataset is at the "decision-making journey" or at the `UserID-ProductID` level. A data sample contains (1) the product review texts read by the user, (2) the explanatory variables shown to matter in predicting the purchase conversion in business literature (e.g., product price and review star ratings), and (3) a binary label indicating conversion. Next, we discuss selection criteria and data construction.

Selection Criteria & Data Frame Construction

We first define the scope of `UserID-ProductID` pairs ("journey"). These pairs are used to identify the journeys and serve as the primary key for our constructed data frame. We keep all existing `UserID-ProductID` pairs in the raw dataset, except we remove users who do not read any reviews and products that have no viewed reviews, since FCM requires meaningful text linked to business outcome variables.

The next step is to process the review texts. On the website, reviews are presented in groups of five. Consumers can read more reviews by clicking on the review page numbers. For each journey, we collect all the reviews that the consumer has read, sort them by browsing time, and concatenate them into a single document. This constitutes the text data of interest. As 88% of the journeys have less than 10 reviews read, we take the final 10 reviews read by consumers before they purchase or abandon.

Lastly, as the total conversion rate is 1.37%, there are many more journeys that ended in abandon (negative label) than conversion (positive label). Considering that the imbalance might negatively affect the performance of the trained FCM, we under-sample the negative pairs to reduce the class imbalance, achieving an abandon-to-purchase ratio of roughly 77:23. Finally, our constructed data frame is left with $58,229$ journeys of $30,218$ unique consumers and $6,612$ unique products. Of these journeys, $13,094$ yield user purchases.

Table 2 presents the summary statistics of the explanatory variables at product, user, or journey levels.

Variable	**Variable-Level**	**Definition**	**Count**			
Product ID	Product	Total number of products	6612			
User ID	User	Total number of unique users who have read reviews	30218			
Content ID	Product	Total number of reviews	87593			
Variable	**Variable-Level**	**Definition**	**Mean**	**Std Dev**	**Min**	**Max**
Price	Product	Price of the product	63.28	84.41	0.24	1049
Rating Average	Product	Average rating of reviews available for the product	4.28	0.49	1	5
Prod Total # Page views	Product	Total number of page views for a particular product	141.79	178.02	2	3280
Prod Total # Reviews	Product	Total number of reviews available for the product	66.37	117.27	1	1463
User Total Wallet size	User	Total dollar value of transaction made by a user	104.68	171.76	0	3968.79
User Total # of Reviews Read	User	Total Number of Reviews read by a user	48.73	75.22	1	2675
User-Prod Page views	Journey	Total number of page views for a particular user-product pair	3.25	2.64	1	100
User-Prod # of Reviews Read	Journey	Total number of reviews read in a journey	10.90	13.85	1	376

Table 2: Variable Summary Statistics

5 Results

The predictive performance of FCM in comparison to existing methods is presented first. Then, we discuss the main goal of FCM to extract coherent and human interpretable concepts, first through a coherence metric and then through human-judged metrics. To gauge the economic impact of mined concepts, we then present FCM's ability to compare the correlational relative importance of mined concepts to other structured explanatory variables. A series of experiments to investigate the accuracy-interpretability relationship based on different hyperparameters are presented. Lastly, results on an additional dataset, Donorschoose, are presented along with the impact of Y focusing on extracted concepts.

5.1 Predictive Performance

For predictive performance measurement, data is split into 70% training, 15% validation, and 15% test sets. The FCM takes the explanatory variables and the read-reviews as inputs to predict a probability between 0 and 1 indicating how likely it is that the user will purchase or abandon the product. We set the number of concepts to recover at 5, which we chose via a perplexity measure on running a separate vanilla LDA model. The adaptive moment estimation (Adam) optimizer was used for training the model with a learning rate of 0.001 and weight decay of 0.01. To prevent overfitting and gradient explosion, we set the dropout rate at 0.25 and clipped the gradients into the range $[-5.0, 5.0]$. The model is trained on up to 500 different parameter configurations and our model gives stable results across these sets. For brevity, all results hereafter are produced by the model trained under the configuration $\lambda = 10, \rho = 1000, \eta = 1000$.

The predictive performance is measured on 15% of the test sample using the accuracy, precision, recall, F1-score, and ROC AUC, as discussed in Section 3.6 in Table 3. Figure 5 shows the ROC curve. We compare FCM's predictive performance against two set of baseline models—interpretable models vs. uninterpretable prediction-focused models. For all models, unless stated otherwise, we include all X and text information. For brevity, we provide a brief description of the model with the citation for readers interested in more details.

Interpretable Models

- **LDA + LR:** Plain LDA + logistic regression classifier.
- **SLDA:** Supervised LDA proposed by Blei and Mcauliffe (2008); Zhu et al. (2012) (model does not allow other X).
- **SeedLDA + LR:** Seeded LDA proposed by Jagarlamudi et al. (2012) + logistic regression classifier. We seed the topics based on dimensions of price and quality as defined in the literature and discussed in Table 4 to maximize its performance.
- **Structural Topic Model:** Structural Topic Model proposed by Roberts et al. (2014). Incorporates both X and natively handles Y.

Uninterpretable or Prediction-Focused Models

- **BOW+LR:** Bag-of-words approach + logistic regression classifier.
- **Sentiment:** Review-level sentiment-labelled classification.
- **CNN with GLoVe:** Deep Learning models excel at prediction tasks. Convolutional Neural Net (CNN) was chosen since it was empirically shown to be superior to LSTM and Recursive Neural Net for *this particular dataset* (Liu et al., 2019). We also utilize GLoVe word embedding (Pennington et al., 2014) as the first layer.
- **XGB:** eXtreme Gradient Boosting (Chen and Guestrin, 2016) is a boosted tree model often known to achieve the best predictive performance in a wide variety of ML competitions off-the-shelf. Please see Footnote 4.

Figure 5 presents the ROC curves and the AUC values for the FCM (blue), 4 interpretable baselines (red), and 4 prediction-focused baselines (black). Table 3 provides accuracy, precision, recall, F1, and AUC. First note that all interpretable models (red) significantly fall behind FCM (blue). While there are two uninterpretable algorithms (black) that surpass FCM, the difference is rather small at less than 0.03 in AUC. Two uninterpretable algorithms perform worse than FCM.

Going into specific algorithms, as noted in Footnote 4, the top two performing algorithms are, unsurprisingly, deep learning and XGB. CNN with GLoVe embedding achieves the highest AUC at 0.9186 with XGB performing nearly the same at 0.9184. FCM follows closely at 0.8885. Given that we could also boost (reduce bias) and bootstrap aggregate (reduce variance) FCM predictions, albeit at a higher computational cost, FCM performance is very much competitive with the cutting-edge prediction-focused methods. Contrastingly, FCM performs significantly better than the traditional bag-of-words approach (0.7773) and the basic sentiment analysis (0.6093).

Among the interpretable competitors, the best performing algorithm is the seeded LDA (0.8087). Our seed words were driven from existing theory in consumer purchase behavior as will be elaborated in Section 5.2 and Table 4. As these are topics that are known and proven in the literature to influence consumer purchase decisions, it is unsurprising that this carefully domain-knowledge driven attributes will have high predictive performance over naive bag-of-words approach. However, seeded LDA still falls short of FCM, suggesting that FCM extracts residual signals above and beyond

	Coherence	Accuracy	Precision	Recall	F1 score	AUC
FCM	-1.86	0.8228	0.8346	0.8475	0.8410	0.8885
Interpretable Model Competitors						
LDA + LR	-2.25	0.5653	0.5624	0.9630	0.7101	0.6086
SLDA	-2.17	0.5857	0.6163	0.6641	0.6393	0.6041
Seeded LDA+LR	-1.95	0.7490	0.7945	0.7365	0.7644	0.8087
STM	-5.13	0.5872	0.5977	0.7745	0.6747	0.6015
Uninterpretable Model Competitors						
BOW+LR	N/A	0.7212	0.7231	0.8033	0.7611	0.7773
Sentiment	N/A	0.5804	0.5896	0.7925	0.6762	0.6093
CNN with GloVe	N/A	0.8421	0.8307	0.8973	0.8627	0.9186
XGB	N/A	0.8475	0.8525	0.8757	0.8639	0.9184

Table 3: **Prediction Performance Against Competing Methods**

theory-driven concepts. Seeded LDA approach can be useful when managers are equipped with domain-knowledge but not feasible for exploratory concept extraction. Other interpretable models such as vanilla LDA (0.6086), supervised LDA (0.6041), and Structural Topic Models (0.6015), all perform worse than FCM even with good effort of parameter tuning. Appendix E presents FCM's predictive performance against baselines for a well-known benchmark dataset called 20-Newsgroup, which show FCM excels even over XGB in some cases.

FCM excels in predictive performance over most baselines while staying competitive with the top uninterpretable prediction-focused algorithms. Next we discuss the main goal performance of FCM: interpretable and coherent concept extraction for managerial insight.

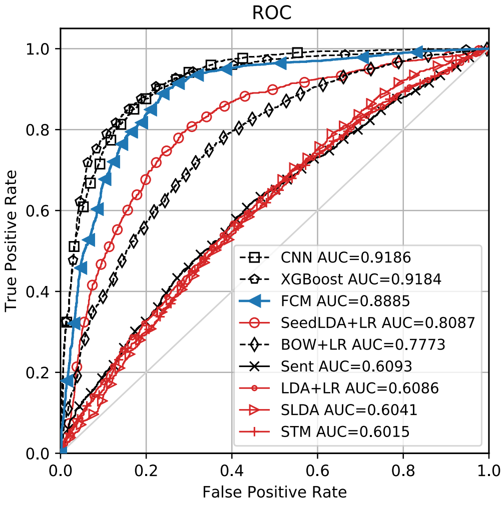


Figure 5: **Receiver Operating Characteristics Curve of FCM vs. Baselines**

5.2 Concept Extraction: Interpretability Comparison Based on Coherence

This section discusses the main ability of FCM to extract concepts that may be insightful and useful for prescriptive business policy. As described in the Model section, FCM is not limited to just generating topic-describing words as LDA or LDA variants are. However, we can generate words to describe the concepts, much like LDA, as described in Section 3.5. We do this here to compare the interpretability of FCM-extracted concepts to those from LDA. Then we use the measure coherence described in Equation 16.

Figure 6 shows five topics (topic N decided by perplexity measure) recovered by plain LDA as the baseline. Out of 50 runs, we present the LDA result with the highest coherence of -2.25. Two experts then manually inspect the topics recovered and color-code similar conceptual words for ease of interpretation. Upon inspecting the LDA-recovered topic, we note the following observations. First, within a given topic, many different concepts appear. For example, Topic 1 has concepts related to product names (orange), features (dark red), and aesthetics (green). All other topics suffer the same diffused conceptual intrusion. Second, the recovered topics are not well separated. Many words related to singular concepts such as aesthetic (green), product name (orange), or features (dark red), appear in most if not all of the five topics. Presented with such outputs, it is *unclear* how managers may then utilize recovered topics as X for further study and insight.

Table 7 displays the concept-describing words for each concept extracted by FCM. We present five concepts (keeping the same number of concepts as LDA) with a word cloud. Compared to the topics found by the standard LDA, which are shown in Table 6, the FCM's concepts are more semantically coherent within each dimension. Concepts are both well focused within as well as separated from one another. Borrowing from the topic modeling literature, we apply the same measure of coherence in equation 16 and obtain an average coherence of -1.86, which is greater (higher coherence) than the -2.25 obtained by LDA. Among the five concepts, the aesthetics concept represented by words such as "nice little big clean look design" achieved the highest coherence of -1.384. Even for the concept of serviceability with the lowest coherence score of -2.197, the words are semantically coherent. We compare the ranges of average 5-topic coherence of 50 distinct estimations (varying starting points and initialization) of LDA vs. FCM. The LDA range lies in (-3.5, -2.25) while the FCM range lies in (-1.92, -1.68). The coherence ranges do not overlap and are statistically significantly different, suggesting that FCM excels in extracting many coherent and singular (and thus more human-interpretable) concepts from review texts.

In fact, the concepts recovered by FCM closely coincide with the dimensions of product price and quality that are shown in the literature to influence consumer purchase. Garvin (Garvin, 1984, 1987) compiled and introduced a set of quality and price dimensions aimed at helping organizations think about product, as shown in Table 4. On multiple FCM runs, we were able to automatically extract all concepts compiled by Garvin. This serves as an evidence of the external validity of FCM.

Table 3 provides the coherence of 4 other interpretable baselines. FCM is superior to all of them. Only seeded LDA was not statistically significantly different from FCM's concept coherence. For actual examples of reviews along with computed concept and case study, please see Appendix C.

Topic 0: phone + iron + lamp + **find** + **feature** + steam + **replace** + **old** + **design**

Topic 1: **sound** + picture + **old** + **brilliant** + battery + son + **feature** + **problem** + **smart** + **fantastic**

Topic 2: kettle + **clean** + microwave + **quick** + **size** + toaster + **design** + **heat** + **brilliant** + **quickly**

Topic 3: **assemble** + **sturdy** + room + curtain + **space** + **size** + **hold** + bin + **perfect** + **design**

Topic 4: bed + **comfortable** + cover + **lovely** + pillow + duvet + floor + mattress + **feel** + **thin**

Figure 6: **Topics Extracted by LDA:** Best coherence out of 50 runs. Average topic coherence is -2.25. We manually color-coded similar concepts for easy visualization.

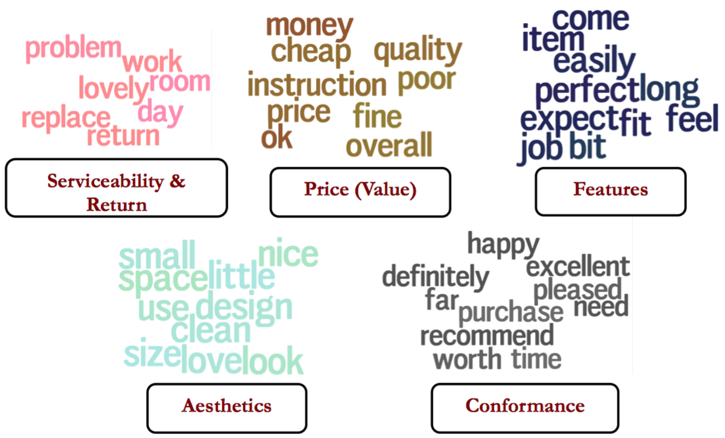


Figure 7: **Concepts Extracted by FCM:** Concept representative words are found to compare to LDA. Average Concept Coherence is -1.86.

Dimension	Description
Aesthetics	The review talks about how a product looks, feels, sounds, tastes, or smells.
Conformance	The review compares the performance of the product with per-existing standards or set expectations.
Durability	The review describes the experience with durability or product malfunctions or failing to work as per the customer's satisfaction.
Feature & Performance	The review talks about presence or absence of product features.
Brand	The review talks about indirect measures of the quality of the product like the reputation of the brand.
Price	The review contains content regarding the price of the product.
Serviceability	The speed, courtesy, competence, and ease of repair in case of any defects with the product.

Table 4: **Literature-defined Key Dimensions of Price and Quality**

5.3 Concept Extraction: Interpretability Comparison Based on Human-Judgement

Two human-judged measures of interpretability (as discussed in Section 3.7) are obtained from two distinct survey instruments (Appendix B). In both surveys, we show the outputs from both FCM and LDA to Amazon Mechanical Turkers.

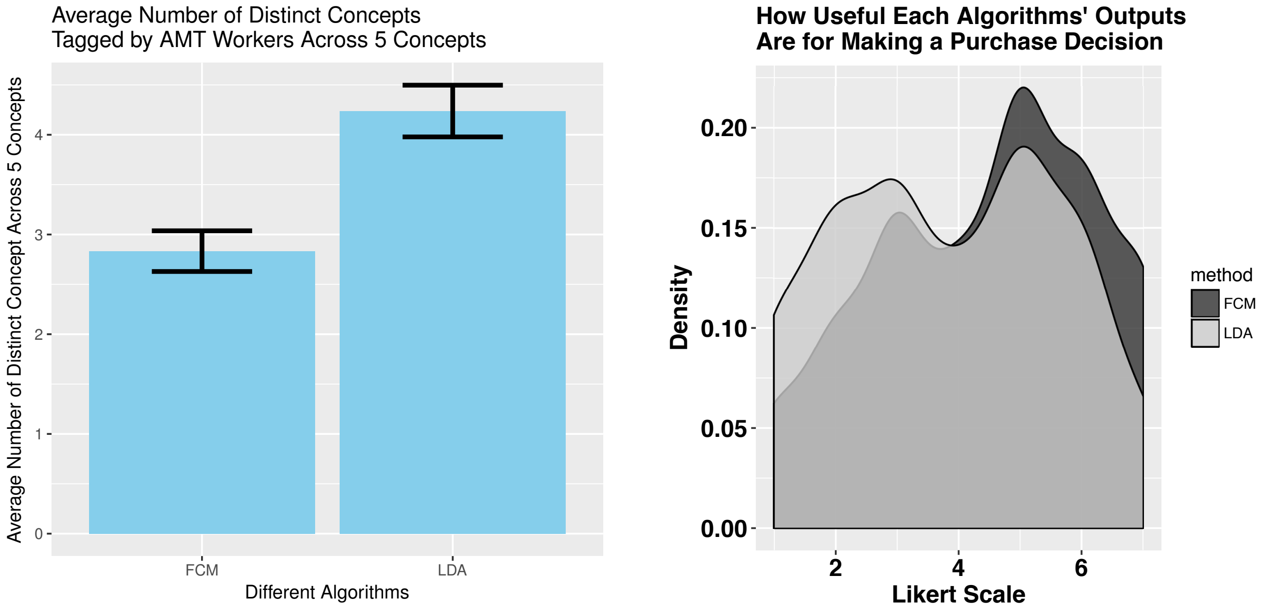
For the *first* human-judged measure, we ask turkers how many distinct concepts they see in algorithm outputs at the topic or concept level. A smaller number signifies that concept or topic is more focused and concentrated in meaning, as well as less ambiguous, and thus more interpretable.

For the *second* measure, turkers are told to imagine a hypothetical situation in which they are shopping online and making a purchase decision. We show them the outputs of FCM and LDA and state that these are topic keywords about the reviews. Next, we ask them to rate on a Likert-like scale how useful these reviews would be in making a purchase decision.

For each survey, we asked 100 turkers who have previously completed at least 100 tasks with 98% or greater accuracy. We embedded a couple of attention questions and also filtered the results to prevent bots. The ordering of questions and topic presentations were randomized.

Figure 8a shows that turkers on average found 2.83 distinct concepts in each FCM concept and 4.23 distinct concepts in each LDA topic. The T-test of difference in mean was statistically significant at the p-value $= 8.19 \times 10^{-8}$. This suggests that FCM is able to produce concepts that are less ambiguous and more focused compared to LDA, and thus more interpretable.

Figure 8b shows the result of the second survey. The X-axis shows the value of the Likert scale where 1 means algorithm's output was "extremely not useful" for making a purchase decision while 7 means "extremely useful". We first note that both distributions are bimodal, suggesting some concepts and topics were useful while others were not. Second, FCM on average scored 4.439, while LDA scored 3.855, suggesting that FCM outputs were more helpful. T-test of difference in mean was statistically significant at the p-value $= 1.472 \times 10^{-5}$. Taken together, human judgement metrics find FCM more interpretable compared to LDA outputs.



(a) **Human-Judged Number of Concepts in FCM vs. LDA:** Standard errors are shown.

(b) **Human-Judged Usefulness of Concepts Found by FCM vs LDA for Making a Purchase Decision**

Figure 8: **Human-Judged Measures of Interpretability for FCM vs. LDA**

5.4 Correlational Relative Importance of Mined Concepts vs. Referential X to Gauge Economic Significance

Concepts	Coef	Explanatory variables	Coef
Aesthetics	0.107	Price	-0.003
Conformance	-0.076	Avg Rating (Standardized)	0.288
Features	0.058	User-Page Views	-0.005
Value	0.054		
Serviceability	-0.083		

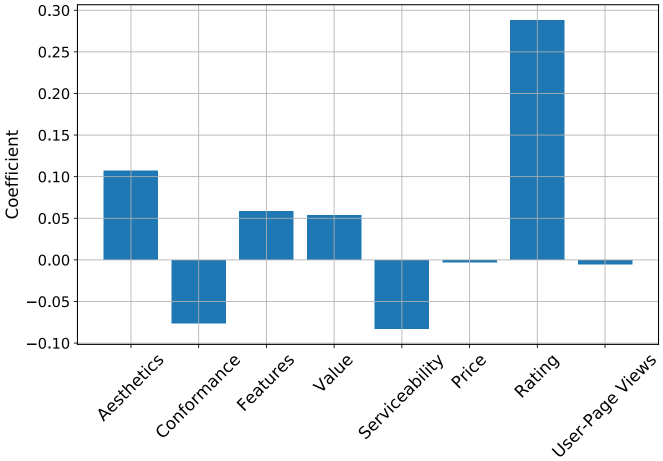


Table 5: **Estimated Coefficients**

For better utilization of the extracted concepts for business decisions and insight, FCM provides a way to compare the correlational relative importance of mined concepts against the user-inputted X. As demonstrated in this section, the idea is to supply FCM with relatively well-understood X of interest along with texts to compare the correlational impact on the Y outcome. For model parsimony, we choose the top three reference-worthy explanatory variables of high theoretical importance by information gain to compare to mined concepts: price, average star rating, and user page views. Results run with more explanatory variables perform similarly. We note that (1) FCM can still be applied on text-only data without any other X, (2) the relative importance provided does not have any causal claims, as they are predictive in nature, and lastly, (3) out of 500 different runs of FCM, the recovered concepts and estimated coefficients are rather stable.

In the last layer, recall that FCM predicts the conversion (Equation 15) of a journey with the mined document-concept distribution, p_d. We modify this prediction layer to include user-specified X. We rename p_d as $DocConceptD$ for clarity and add the explanatory variables $ExpVar$ with a sigmoid (generalized logistic) function:

$$\hat{Conversion} = \sigma(\theta_0 + \sum_i \theta_i DocConceptD_i + \sum_j \theta_j ExpVar_j) \tag{17}$$

where $DocConceptD$ is a probability vector of different concepts that sums up to 1. For this exposition, we have named our concepts as discussed in the last section and in Figure 7. For example, if a document gets a distribution of $(0.2, 0.2, 0.2, 0.2, 0.2)$ over 5 concepts, this means that all 5 concepts are equally represented in this document (product review). This indicates that the coefficients speak to the impact of concept volume present in the review documents. As in generalized linear models, the trained weights, θ, characterize how much the predicted conversion will respond to the change of explanatory variables. For clarification, although the sigmoid layer of FCM follows the formula of a generalized linear regression, we are not aware of any work that could provide the confidence interval of a deep learning-based model. Thus, we do not provide the confidence interval.

Table 5 shows the trained coefficients of explanatory variables and concept weights including aesthetics, conformance, features, value, and serviceability. We standardized the average rating for easier interpretation. The results pass the sanity check: a negative coefficient for price and a high positive coefficient on average ratings. Aesthetic concepts had the highest positive correlation with conversion while the serviceability and return-related concepts had the lowest. For a better interpretation of mined concepts, we will now calculate and present the marginal effects.

Interpretable Correlational Association

Concepts	Marg effects (%)	Explanatory variables	Marg effects (%)
Aesthetics	0.201	Price	-0.022
Conformance	-0.266	Avg Rating (Standardized)	0.336
Features	0.226	User-Page Views	-0.041
Value	0.040		
Serviceability	-0.279		

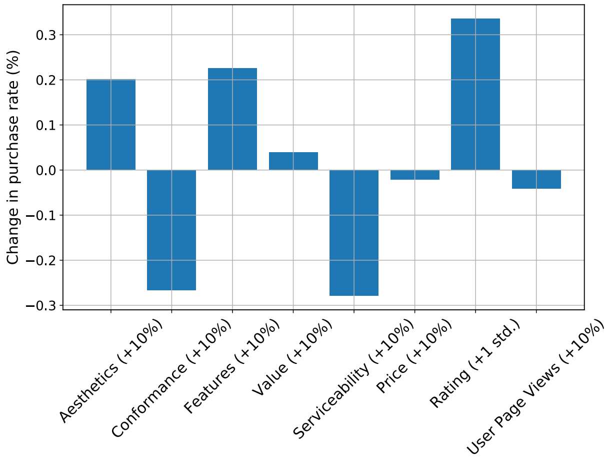


Table 6: **Correlational Association of Review Concepts and Explanatory Variables**

While the last sigmoid layer can provide an odds-ratio interpretation, just as in logistic regression, we calculated the correlational association of the variables on the percentage change of predicted conversion rate for brevity and clarity. To do this, for each journey and for each variable, we calculate the probability of conversion both as is and after adjusting the variable, then average the effects. First, we use the full FCM model to calculate the predicted conversion rate $Conver\hat{s}ion$. Then, we increase the variable by an interpretable amount we define and compute the new predicted conversion rate $Conver\hat{s}ion'$, holding everything else fixed. The difference of the two conversion rates $\Delta = Conver\hat{s}ion' - Conver\hat{s}ion$ will be interpreted as the correlational association of the variable. We apply a similar method to mined concepts. However, keeping in mind that the concept distribution sums up to one within a document, we compute the correlation of a concept dimension by increasing its distribution by 10% (percentage point) and decreasing others by 2.5%. That is, we ask what happens to a predicted conversion rate if a consumer reads 10% more of a given concept within a journey.

The correlation are presented in Table 6 along with visualization. The increase in the review rating by one SD (0.49 star) is associated with the predicted conversion rate increase of 0.336%. Given that the average conversion rate is 1.37%, this is equivalent to an increase of approximately 25% in the predicted average conversion rate. In comparison, different concepts mined change the predicted conversion rate by -0.279%˜0.226% , which equates to the range of a 20% decrease to a 16% increase on the predicted average conversion rate. Comparing the referential X (average product rating) to concepts mined by FCM, we see that the impact of a 10% increase in concept volume in consumer-read reviews amounts to approximately 12%˜83% of the effects (in absolute values) equivalent to increase in the average rating by one SD. The relative importance weight in conjunction with correlation calculations serves to provide mangers with a sense of economic importance of automatically mined concepts in comparison to familiar key referential X such as review ratings.

Our model suggest correlationally that reviews containing a higher volume of aesthetics, features, and price (value) might boost the conversion rate. On average, if the proportion of a document's aesthetical information increases by 10%, the predicted conversion rate increase by 0.201%. On the other hand, the concepts of conformance and serviceability are negatively correlated to consumer purchases. Upon investigating several examples, this is likely because users usually mention negative aspects when conformance (e.g., mismatch in product size) and serviceability (e.g., mentions returning or experience of returns) concepts are mentioned. We further investigate if certain concepts matter more if presented on different review pages. That is, does the order in which certain review concepts are shown to customers correlate with more or less conversion? We present this in Appendix D.

Given the FCM results, an e-commerce manager may next launch a more focused causal study to investigate how to prioritize certain concepts and information in customer reviews—such as including text that better highlights—features and aesthetics, to increase conversion rate. Alternatively, a longer term objective may be to focus on particular set of products to discern aspects of the category that consumers care about most for new product development. This is outside the scope of this paper.

5.5 Experimentation on Interpretability-Accuracy Relationship

This section is dedicated to exploring how interpretability and accuracy are correlated in FCM. In the topic modeling literature, Chang et al. (2009) show that the model fit (perplexity) is negatively correlated to interpretability (coherence). In the context of object classification in computer vision via convolutional neural networks, Bau et al. (2017) report experimental findings using many neural network models and datasets that indicate the interpretability of a model can be decreased without changing its discrimination ability. The authors concluded that the interpretability of a model is not required for high accuracy. Extending this work, Zhang et al. (2017) develop an interpretable convolution neural network and in the process show that there is a trade-off between interpretability and classification accuracy.

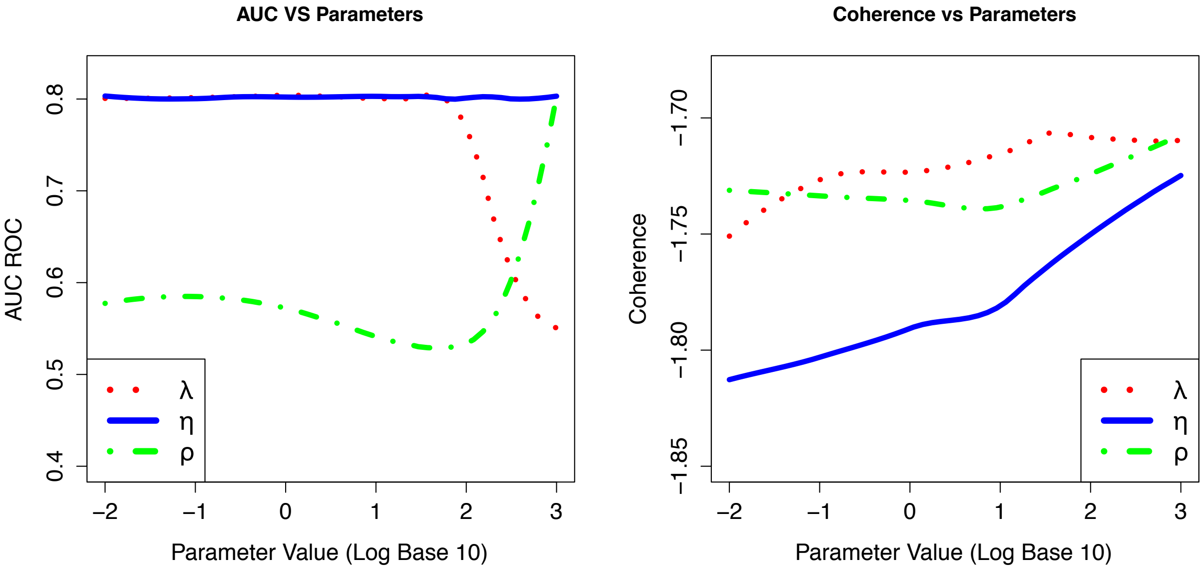
Specifically, as FCM's objective function consists of different components, we can directly see how increasing certain weight on the objective function changes the accuracy vs. coherence. For example, we explore how increasing ρ influences accuracy and coherence. Increasing ρ should also increase the accuracy but it is not clear if it will have the same impact on coherence. In particular, we examine ρ (classification weight), η (concept diversity weight), and λ (Dirichlet sparsity weight). We run experiments and plot the results in Figure 9. The first plot shows the impact on AUC as we vary ρ, η, λ and the second plot shows the impact on coherence. For each parameter, we vary the parameter from 10^{-2} to 10^{3} in 20 equally spaced points while repeating each point three times, giving us a total of 60 points per parameter. While one parameter is varied, the other two are fixed. The results are then smoothed with a LOESS (locally estimated scatterplot smoothing) plot.

For AUC (predictive power), ρ has an expected trend. As ρ increases, AUC increases. On the other hand, λ shows the opposite pattern. As we force sparsity of concept, the predictive accuracy decreases—a clear loss in signal. Interestingly, η, concept diversity, does not seem to influence the AUC. From the geometric point of view, in the concept vector space, there may be clusters of several different concepts. Increasing η does not seem to decrease predictive information, since it forces each concept to cover different regions in the concept space, as opposed to increasing sparsity, which does decrease predictive signals.

For coherence (interpretability), neither λ or ρ seem to have a clear trend. Both have a slight upward trend, but the range is not so large nor the pattens so clear. However, in comparison, η seems to have a clear upward pattern. As we increase the concept diversity, coherence and interpretability increases.

Taken together, we document several interesting findings: 1) The interpretability-accuracy tradeoff correlation depends on different parameters of the model. 2) Increasing ρ increases AUC as intended, but only slightly increases interpretability. 3) Increasing concept sparsity, λ, decreases AUC but doesn't influence interpretability. and 4) Increasing concept diversity, η, does not influence AUC while increasing interpretability. Taken together, the experiments suggest that the accuracy and interpretability is a separate dimension, and that one does not necessarily increase or decrease based on the other. This section echoes findings from (Bau et al., 2017) in computer vision setting with convolutional neural networks.

We speculate on the slight increase of coherence as ρ increases. It is unclear why guiding the concept discovery with classification loss increases coherency (albeit only slightly). One idea is that this is because the text does have useful signals that are correlated to (or even cause) outcome prediction. In this case of conversion, the coherent concepts such as aesthetics and conformance are what may be driving the purchase decisions (the Y). Thus, the concepts embedded in the reviews and the reasons why consumers make decisions are aligned and interpretable.



(a) **Variation in ROC-AUC**

(b) **Variation in Coherence**

Figure 9: Variation in ROC-AUC and Coherence for different hyperparameter settings. Hyperparameters were varied for each while the other two were fixed.

5.6 Results on Different Dataset For Robustness - DonorsChoose.org

We apply FCM on another dataset. We only focus on the main task of concept extraction and comparison. The open data is from DonorsChoose.org, a nonprofit crowdsourcing site that allows individuals to donate directly to public school classroom projects. This dataset spans over 6.2 million donations and 1.2 million fundraising projects from 2002 to 2016 at the project level. Since the project page usually provides little structured information to the individual donors other than

	Concept-Describing Words	Concept Title	Coef
1	camera history photography animal world trip experience picture video life	Photography & Outdoor Learning	0.0311
2	art material child supply color easel chair time pencil center	Art Supplies	0.0868
3	music play drum instrument equipment song sensory fitness physical keyboard	Music & Physical Education	0.0188
4	book read reading novel text library reader level love language	Reading & Literature	0.0597
5	technology math computer ipad lab project able science allow laptop	Tech Equipment	-0.0689

Table 7: **Concepts Extracted by FCM for the DonorsChoose Dataset**

the full text of teacher-written essays, we believe that the textual data should yield a significant signal in predicting the fulfillment of the donation goal—predicting and giving insight into what sort of crowdfunding project for education is successfully funded.

Based on the fundraising status of each donation project (whether the total donation hits the goal before the expiration date), we label the projects as either a success (positive) or failure (negative); we then construct training data consisting of 10,000 positives and 10,000 negatives by randomly selecting projects. Each row contains 1) the project essay text and 2) the binary project status as the label. We also control for the project funding amount. As with our main results, we set five concepts to be discovered. Extracted concept-describing words and estimated relative importance coefficients are found in Table 7.

Again, the concepts are semantically highly coherent and well separated from one another. We manually name each concept with descriptive titles for convenience, and find that the extracted concepts consist of essentially five different curriculum types. The estimated coefficients imply that art supply-related (0.0868) or literature-related (0.0597) concepts are more likely to be successfully funded. FCM also suggests that technology-equipment funding requests are less likely to be successful.

5.7 The Role of Y-Focusing - Examples from DonorsChoose.org

This section investigates the impact of Y-focusing on FCM output—i.e., what concepts are extracted if a different Y is selected to guide the concept discovery? For this, DonorsChoose data is used again due to the availability of a different Y.

Y-Variable: $NotFunded$

	Concept-Describing Words	Concept Title	Coef
1	teach time new allow day tool board lesson special	Special Teaching Tool	-0.0402
2	child create project material provide come like activity education	Project Materials	0.0441
3	music play math science kit experience stem language opportunity	Music & STEM	-0.0224
4	book read reading text novel level library reader love english	Reading & Literature	-0.0275
5	technology skill ipad computer program able grade tablet app access	Tech Equipment	0.0461

Table 8: **DonorsChoose Result Guided by Y-variable** $NotFunded$**.**

One way to obtain a different Y is to simply flip the Y from Section 5.6 to predict $NotFunded$. An FCM run on this data will extract out concepts that are highly correlated to the project being unsuccessfully funded. Intuition suggests that FCM will extract concepts that are repeated and perhaps will also yield new concepts. Table 8 presents the results. Comparing the results for the original (Table 7), we find that concepts "Tech Equipment" and "Reading & Literature" are extracted yet again, with coefficients' directions flipped from the original dataset results. This shows FCM's consistency in both concept extraction and correlational coefficient estimation.

However, FCM also recovers slightly modified concepts or even new concepts not originally found in Table 7. Compare Concept 3 in both tables. It was originally "Music & Physical Education" and positive for successful funding. Now it is changed slightly into "Music & STEM" with a negative value for *NotFunded*. While the directions tell consistent result, extracted concepts are slightly modified. Concept 1 is a new concept seeking to fund some sort of special teaching tool that will enable the teachers. The coefficient is negative, suggesting that it is likely to get funded. Concept 2 refers to student project material requests and is positive, suggesting a lower chance of getting successfully funded.

Y-Variable: *Exciting*

	Concept-Describing Words	**Concept Title**	**Coef**
1	life opportunity experience world garden society culture grow hope	Experiential Learning	-0.0842
2	camera ipad technology video class projector able use tablet allow	Tech Equipment	0.0588
3	drum math calculator science music hand instrument game play teach	Music & STEM	-0.0972
4	art ball provide new material child writing like supply	Art Supplies	0.0049
5	book read reading center love time remember level listen player	Reading & Literature	0.1177

Table 9: **DonorsChoose Result Guided by Y-variable *Exciting*.**

The 2014 KDD (Knowledge Discovery and Data mining) Conference hosted a joint open data science competition with Kaggle.com (https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data) using DonorsChoose data. The competition created a customized binary Y linked to each funding request called *Exciting* as defined and deemed important by DonorsChoose. *Exciting* goes beyond our DonorsChoose Y used in Section 5.6 in that the project must 1) be fully funded (original Y used), 2) have at least one teacher-acquired donor, and 3) have a higher than average percentage of donors leaving an original message, in addition to other stipulations. For more details, please refer to the Kaggle link. In summary, *Exciting* is a domain-expert defined Y that is sufficiently different (correlation between successfully funded and *Exciting* is 0.1484 – note that there are many successfully funded projects that are not *Exciting*) from a simple Y that indicates whether the project was successfully funded or not.

Table 9 shows the results. Two concepts, 4 ("Art Supplies") and 5 (Reading & Literature), are recovered once again and the directions match the original result. Concept 2 ("Tech Equipment") is recovered once again but this time with reversed direction. While the tech equipment was negatively correlated with an unsuccessfully funded project in the original set up, it is positive for the *Exciting* project. Concept 3 ("Music & STEM") is slightly modified from "Music & Physical Education" with reversed direction. Lastly, a new concept related to "Experiential Learning" is discovered to be negatively correlated to *Exciting*.

Summary

Even on the same data, FCM recovers both repeated and new concepts depending on the Y used to focus the concept. This enables creative use of FCM. For example, if a manager had return data connected to review-reading and conversion data, FCM can recover what review content may be instrumental in reducing product returns.

6 Limitations and Future Extensions

We share several short idea overviews for extending the basic FCM models and post-processing FCM outputs for future papers. The scope of difficulties range from feasible to very uncertain. Specifically, four extension ideas are shared in order of increasing difficulty: Automatic Naming of Concepts, Adding Valence to Concepts, Zooming into Concepts, and finally FCM and Causality.

Automatic Naming of Concepts In demonstrating FCM with review data, we have manually interpreted and titled each concept that FCM extracted. Titling concepts can be further automated post-FCM for a faster and more “objective” presentation of the results. One method involves adopting existing techniques from topic modeling literature (e.g., Lau et al. (2011)), which reduces down to generating label candidate sets by tapping into external text—such as Wikipedia—or the corpus itself and ranking the label set by its similarity to topics. However, given that FCM architecture includes semantic-spatial relationship-aware word embedding at its basis, geometric methods within the concept (and word) Hilbert space may be more appropriate and powerful.

Adding Valence to Concepts The current FCM only deals with the volume of concepts present in text. Valence (positive or negative sentiment) could be added to enhance the model and interpretation. Within the model pipeline, before the last softmax layer, the document-concept vector could be further processed. However, it is unclear how to modify the end-to-end architecture to inject valence of the document-concept vector without the need for training datasets or breaking the end-to-end framework.

Zooming into Concepts Model architecture could be extended to be hierarchical in concept representation relations. Given the hierarchical nature of discovered concepts, the model could also include a lever to zoom in or out on concept hierarchy for different levels of abstraction. One potential way might be to tap into existing semantic networks and knowledge graph databases that are aware of concept hierarchy, such as ConceptNet (http://conceptnet.io/) or WordNet (https://wordnet.princeton.edu/).

FCM and Causality While FCM is envisioned and presented as an exploratory tool (and *not causal*) with many caveats, some users may still want to extend it for causal uses. Using deep learning for causal inference is still a nascent field with only a handful of papers (e.g., Hartford et al. (2016), Louizos et al. 2017, Kallus 2018, etc.) and is theoretically undeveloped due to the fact that there are no theoretical asymptotic results on generic deep learning models, which makes it difficult to draw robust inferences.

To speculatively suggest, as the last layer of the FCM architecture resembles classical logistic regression, perhaps it can be extended to inject characteristics of extant causal models. We are unsure where to begin, however. For now, a quick and robust way to utilize FCM is to simply use it as a representation learning algorithm (Bengio et al., 2013) to extract non-trivial representation of input text data by lobotomizing the FCM and using the inner-layer data representations. This could be anything from simple document-concept vectors to a complicated nonlinear combination of document-elements. Given that document-concept vectors are the most interpretable, we suggest starting with these vectors as X inputs to other traditional causal techniques.

7 Conclusions and Managerial Implications

We introduced a new deep learning-based text mining method, the Focused Concept Miner (FCM), to explore, organize, and extract information from large-scale textual data guided by any Y of business importance. Upon applying the algorithm to a unique, comprehensive dataset that tracks individual-level review reading, searching, and purchasing behaviors, we were able to extract content from product reviews that consumers consider before making purchases. The mined concepts overlapped heavily with dimensions of price and quality that existing management theories claim matter for consumer purchase behavior. FCM achieves competitive prediction performance and higher interpretability in comparison to existing techniques and additionally provides relative importance measures for structured X and focus-mined concepts, which in turn provides managers with an easy way to understand and derive value from textual data. We introduced the interpretable machine learning literature and use-case of one such algorithm, FCM, for management applications. FCM also excels in real-time nimble processing of fast incoming text data to provide interpretable feedback since 1) inductive prediction does not require additional training, 2) retraining with incoming data is easy with mini-batch stochastic gradient descent, and 3) Focusing by Y offers different angles to extract insights effortlessly.

We envision FCM as an *exploratory tool* to make sense of the severely untapped textual data in the business world, and as a jumping off point for further policy implementation or causal studies. The review-reading data we focused on illustrated the potentials of FCM, and we now discuss several other possible applications. An immediate application for *review content presentation* on an e-commerce site can be brainstormed based on the FCM output. For example, if the specified Y was return rate instead of conversion, the FCM output could extract the concepts that may be instrumental in preventing returns,[11] and may inform review presentation design to decrease returns. Another immediate application arises naturally due to 1) nimble ability of FCM to focus-mine concept in "one click", 2) one-the-fly prediction of incoming data via inductive prediction, and 3) dynamic model update capability of the model based on incoming data owing to mini-batch stochastic gradient descent algorithm. Specifically, FCM can be used to dynamically monitor consumer feedback and complaints (i.e., *dynamic resonance marketing tool* (Clemons et al., 2006)) on social media and websites as shown in Netzer et al. (2012) and Culotta and Cutler (2016). Managers can benefit from a dashboard of daily or weekly summarization of consumer chatters using FCM in place and even quickly see different aspects by focusing with appropriate Ys. On the *product design side*, producers of a relatively new category of products could quickly collect review data and grasp the consumers' wants and needs by using zoomed-in review data. FCM will recover the concepts that consumers talk about and provide the relative importance of recovered concepts linked to any business variable of importance, such as willingness to recommend and whether a consumer actually makes the purchase. This could provide a more exploratory insight before a costly focus-study group. For example, in the wearable fashion and smart product categories (e.g., health trackers, smart clothing, connected luggage), which are relatively new, running FCM on reviews could provide rank order (relative importance) of product quality dimensions that consumers care about. Perhaps aesthetics is more important than features in product category X, which may well be reversed for product category Y. As a last example, if a manager restricts a review dataset to be of a specific product category (e.g., digital cameras), FCM could be modified to extract product-specific feature concepts and relative importance, which could inform product design. This application would be a quicker but much less accurate alternative to the method described by Timoshenko and Hauser (2018).

[11] To speculate, perhaps it could be about clothing fit information, information regarding experience attributes, etc.

We chose the review-reading data due to the familiarity of the concepts in reviews. However, FCM can also be applied to more esoteric text data (e.g., medical text data, technical text data with jargon.), as the algorithm is oblivious to levels of textual technicality. In this case, managers may potentially find new concepts that they were not aware of. Recovered concepts must then be discussed with domain experts to inform further actions.

In the context of Ehrenberg's approach to management science (ETET), there are two broad use-cases for machine learning algorithms for managers and business researchers: 1) scale hypotheses testing, 2) discovering hypotheses from empirical data. Extant papers already utilize ML to scale theory testing (see end of Section 2.3). Additionally, ML can serve as a navigator to point out interesting patterns that could potentially generate worthwhile hypotheses to be causally explored in more depth. Using ML to augment hypotheses generation is ripe for serious consideration. Interpretable Machine Learning algorithms such as FCM can be utilized to explore structured and unstructured data alike in order to augment hypotheses building. Ultimately, however, FCM is as good as the user. For example, in our data, the reviewers online are self-selected and heterogeneous. FCM captures insight from the text *as is*, albeit while controlling for any variables that are supplied to the model. Only researchers who practice sound logic through domain knowledge can be good judges of what is spurious and what is worth further investigation. We hope managers and researchers can use FCM creatively with any combination of text, structured, and business outcome variables to glean insights, build out new hypotheses, and prepare theory-driven causal analyses.

References

Angwin, J., L. Kirchner, S. Mattu, and J. Larson: 2016, 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks., May 2016'.

Archak, N., A. Ghose, and P. G. Ipeirotis: 2011, 'Deriving the pricing power of product features by mining consumer reviews'. *Management Science* **57**(8), 1485–1509.

Bass, F. M.: 1995, 'Empirical generalizations and marketing science: A personal view'. *Marketing Science* **14**(3_supplement), G6–G19.

Bau, D., B. Zhou, A. Khosla, A. Oliva, and A. Torralba: 2017, 'Network dissection: Quantifying interpretability of deep visual representations'. *arXiv preprint arXiv:1704.05796*.

Bengio, Y., A. Courville, and P. Vincent: 2013, 'Representation learning: A review and new perspectives'. *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1798–1828.

Blei, D. M. and J. D. Mcauliffe: 2008, 'Supervised topic models'. In: *Advances in neural information processing systems*. pp. 121–128.

Blei, D. M., A. Y. Ng, and M. I. Jordan: 2003, 'Latent dirichlet allocation'. *Journal of machine Learning research* **3**(Jan), 993–1022.

Buschken, J. and G. M. Allenby: 2016, 'Sentence-based text analysis for customer reviews'. *Marketing Science* **35**(6), 953–975.

Chandrashekar, G. and F. Sahin: 2014, 'A survey on feature selection methods'. *Computers & Electrical Engineering* **40**(1), 16–28.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei: 2009, 'Reading tea leaves: How humans interpret topic models'. In: *Advances in neural information processing systems*. pp. 288–296.

Chen, T. and C. Guestrin: 2016, 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. pp. 785–794.

Clemons, E. K., G. G. Gao, and L. M. Hitt: 2006, 'When online reviews meet hyperdifferentiation: A study of the craft beer industry'. *Journal of management information systems* **23**(2), 149–171.

Culotta, A. and J. Cutler: 2016, 'Mining brand perceptions from Twitter social networks'. *Marketing science* **35**(3), 343–362.

Decker, R. and M. Trusov: 2010, 'Estimating aggregate consumer preferences from online product reviews'. *International Journal of Research in Marketing* **27**(4), 293–307.

Dhurandhar, A., V. Iyengar, R. Luss, and K. Shanmugam: 2017, 'TIP: Typifying the interpretability of procedures'. *arXiv preprint arXiv:1706.02952*.

Doshi-Velez, F. and B. Kim: 2017, 'Towards a rigorous science of interpretable machine learning'. *arXiv preprint arXiv:1702.08608*.

Egele, M., T. Scholte, E. Kirda, and C. Kruegel: 2012, 'A survey on automated dynamic malware-analysis techniques and tools'. *ACM computing surveys (CSUR)* **44**(2), 6.

Ehrenberg, A. S.: 1994, 'Theory or well-based results: which comes first?'. In: *Research traditions in marketing*. Springer, pp. 79–131.

Feifer, J.: 2013, 'The Amazon Whisperer'.

Gantz, J. and D. Reinsel: 2011, 'Extracting value from chaos'. *IDC iview* **1142**(2011), 1–12.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou: 2018, 'Word embeddings quantify 100 years of gender and ethnic stereotypes'. *Proceedings of the National Academy of Sciences* **115**(16), E3635–E3644.

Garvin, D. A.: 1984, 'What Does Product Quality Really Mean?'. *Sloan management review* p. 25.

Garvin, D. A.: 1987, 'Competing on the 8 dimensions of quality'. *Harvard business review* **65**(6), 101–109.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal: 2018, 'Explaining Explanations: An Overview of Interpretability of Machine Learning'. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. pp. 80–89.

Goldberg, Y.: 2016, 'A primer on neural network models for natural language processing'. *Journal of Artificial Intelligence Research* **57**, 345–420.

Goldberg, Y. and O. Levy: 2014, 'word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method'. *arXiv preprint arXiv:1402.3722*.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi: 2018, 'A survey of methods for explaining black box models'. *ACM Computing Surveys (CSUR)* **51**(5), 93.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy: 2016, 'Counterfactual Prediction with Deep Instrumental Variables Networks'. *arXiv preprint arXiv:1612.09596*.

Jagarlamudi, J., H. Daumé III, and R. Udupa: 2012, 'Incorporating lexical priors into topic models'. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 204–213.

Jurafsky, D.: 2000, 'Speech and language processing: An introduction to natural language processing'. *Computational linguistics, and speech recognition*.

Kallus, N.: 2018, 'DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training'. *arXiv preprint arXiv:1802.05664*.

Kozlowski, A. C., M. Taddy, and J. A. Evans: 2018, 'The geometry of culture: Analyzing meaning through word embeddings'. *arXiv preprint arXiv:1803.09288*.

Lau, J. H., K. Grieser, D. Newman, and T. Baldwin: 2011, 'Automatic labelling of topic models'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. pp. 1536–1545.

Lau, J. H., D. Newman, and T. Baldwin: 2014, 'Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality'. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 530–539.

LeCun, Y., Y. Bengio, and G. Hinton: 2015, 'Deep learning'. *nature* **521**(7553), 436.

Lee, D., K. Hosanagar, and H. Nair: 2018, 'Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook'. *Management Science*.

Lee, T. Y. and E. T. Bradlow: 2011, 'Automated marketing research using online customer reviews'. *Journal of Marketing Research* **48**(5), 881–894.

Levy, O. and Y. Goldberg: 2014, 'Neural word embedding as implicit matrix factorization'. In: *Advances in neural information processing systems*. pp. 2177–2185.

Lipton, Z. C.: 2016, 'The mythos of model interpretability'. *arXiv preprint arXiv:1606.03490*.

Liu, B.: 2012, 'Sentiment analysis and opinion mining'. *Synthesis lectures on human language technologies* **5**(1), 1–167.

Liu, J. and O. Toubia: 2018, 'A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries'. *Marketing Science*.

Liu, X., D. Lee, and K. Srinivasan: 2019, 'Large Scale Cross Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning'. *Journal of Marketing Research - Forthcoming*.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling: 2017, 'Causal effect inference with deep latent-variable models'. In: *Advances in Neural Information Processing Systems*. pp. 6446–6456.

Lu, J., D. Lee, T. Kim, and D. Danks: 2020, 'Good Explanation for Algorithmic Transparency'. In: *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*.

Lundberg, S. M. and S.-I. Lee: 2017, 'A unified approach to interpreting model predictions'. In: *Advances in neural information processing systems*. pp. 4765–4774.

McKinsey: 2016, 'The Age of Analytics: Competing in a Data-Driven World'. Technical report, McKinsey Global Institute.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean: 2013a, 'Distributed representations of words and phrases and their compositionality'. In: *Advances in neural information processing systems*. pp. 3111–3119.

Mikolov, T., W.-t. Yih, and G. Zweig: 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*. pp. 746–751.

Miller, T.: 2018, 'Explanation in artificial intelligence: Insights from the social sciences'. *Artificial Intelligence*.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum: 2011, 'Optimizing semantic coherence in topic models'. In: *Proceedings of the conference on empirical methods in natural language processing*. pp. 262–272.

Mindtree: 2017, 'Integrated Customer Insights'. Technical report, Mindtree.

Moody, C. E.: 2016, 'Mixing dirichlet topic models and word embeddings to make lda2vec'. *arXiv preprint arXiv:1605.02019*.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu: 2019, 'Interpretable machine learning: definitions, methods, and applications'. *arXiv preprint arXiv:1901.04592*.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko: 2012, 'Mine your own business: Market-structure surveillance through text mining'. *Marketing Science* **31**(3), 521–543.

Netzer, O., A. Lemaire, and M. Herzenstein: 2019, 'When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications'. *Journal of Marketing Research* **Forthcoming**.

Newman, D., J. H. Lau, K. Grieser, and T. Baldwin: 2010, 'Automatic evaluation of topic coherence'. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 100–108.

Pennington, J., R. Socher, and C. Manning: 2014, 'Glove: Global vectors for word representation'. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pp. 1532–1543.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al.: 2016, 'SemEval-2016 task 5: Aspect based sentiment analysis'. In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. pp. 19–30.

Puranam, D., V. Narayan, and V. Kadiyali: 2017, 'The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors'. *Marketing Science*.

Ras, G., M. van Gerven, and P. Haselager: 2018, 'Explanation methods in deep learning: Users, values, concerns and challenges'. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, pp. 19–36.

Ribeiro, M. T., S. Singh, and C. Guestrin: 2016, '" Why should i trust you?" Explaining the predictions of any classifier'. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1135–1144.

Rizkallah, J.: 2017, 'The Big (Unstructured) Data Problem'.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand: 2014, 'Structural topic models for open-ended survey responses'. *American Journal of Political Science* **58**(4), 1064–1082.

Rubenstein, H. and J. B. Goodenough: 1965, 'Contextual correlates of synonymy'. *Communications of the ACM* **8**(10), 627–633.

Rudin, C.: 2019, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. *Nature Machine Intelligence* **1**(5), 206.

Sahlgren, M.: 2008, ‘The distributional hypothesis’. *Italian Journal of Disability Studies* **20**, 33–53.

Shi, B., W. Lam, S. Jameel, S. Schockaert, and K. P. Lai: 2017, ‘Jointly Learning Word Embeddings and Latent Topics’. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 375–384.

Timoshenko, A. and J. R. Hauser: 2018, ‘Identifying customer needs from user-generated content’. *Forthcoming at Marketing Science*.

Tirunillai, S. and G. J. Tellis: 2014, ‘Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation’. *Journal of Marketing Research* **51**(4), 463–479.

Toubia, O., G. Iyengar, R. Bunnell, and A. Lemaire: 2018, ‘Extracting Features of Entertainment Products: A Guided LDA Approach Informed by the Psychology of Media Consumption’. *Journal of Marketing Research*.

Wernicke, S.: 2015, ‘How to use data to make a hit tv show’.

Wexler, R.: 2017, ‘When a computer program keeps you in jail: How computers are harming criminal justice’. *New York Times*.

Xun, G., Y. Li, J. Gao, and A. Zhang: 2017, ‘Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts’. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 535–543.

Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann: 2018, ‘Confounding variables can degrade generalization performance of radiological deep learning models’. *arXiv preprint arXiv:1807.00431*.

Zhang, Q., Y. N. Wu, and S.-C. Zhu: 2017, ‘Interpretable convolutional neural networks’. *arXiv preprint arXiv:1710.00935* **2**(3), 5.

Zhu, J., A. Ahmed, and E. P. Xing: 2012, ‘MedLDA: maximum margin supervised topic models’. *Journal of Machine Learning Research* **13**(Aug), 2237–2278.

Appendix A - Product Categories in Data

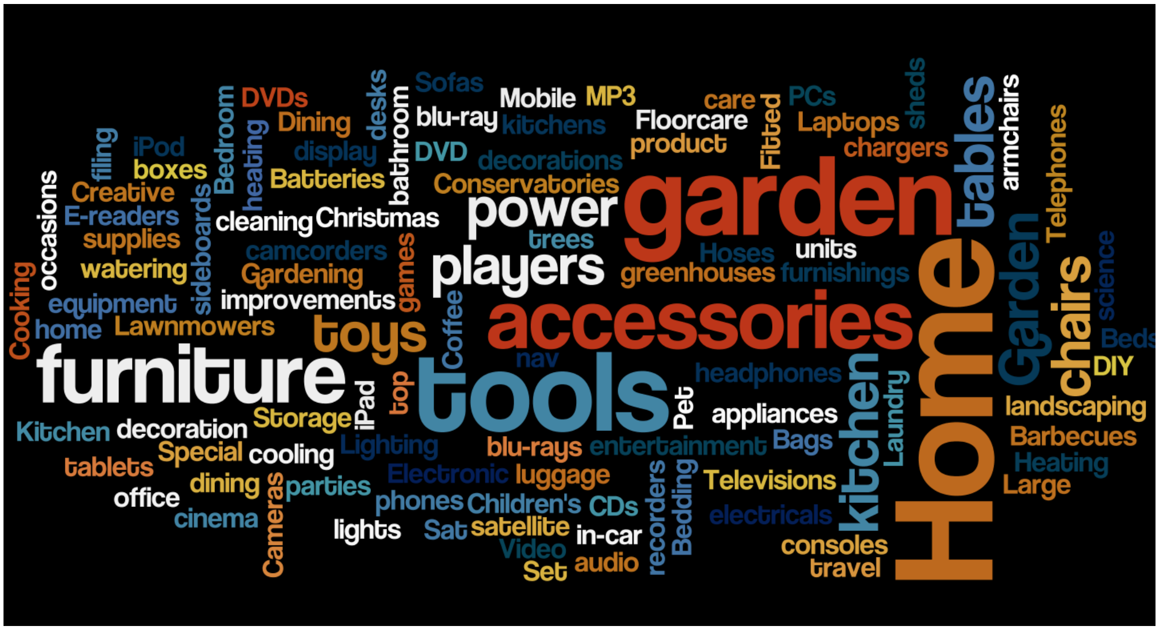


Figure 10: **Word Cloud of Product Categories**

Appendix B - Survey Instrument Used to Measure Interpretability

Survey 1

Given a set of words, please identify the number of distinct high-level concepts or topics described by these set of words.
For example, if the set of words are as follows,
Example Set: Game, Team, Year, Play, Good, Player, Win, Season, Fan, Hockey, Baseball
This example set of words describe one concept or topic "Sports" and # of concept is 1

Here are few more examples

- ai algorithm machine automation robot self-driving [the topic is about artificial intelligence - # of concept =1]
- canon dslr mirrorless 48pixel fullframe fuji lens film tripod [the topic is about DSLR camera - # of concept =1]
- gate earbuds fence speaker pasta tomato keyboard [# of concepts = 3, "Gate Fence", "Earbuds Speaker Keyboard", "pasta tomato"]
- good iron printer buy bin mattress great price board fridge [# of concepts=5, "board iron", "good great buy price", "printer bin", "mattress", "fridge"]
- globe photos frank mass bear mountain cell group area york [# of concepts = 7, "Bear mountain", "cell group area", all other words go by themselves]

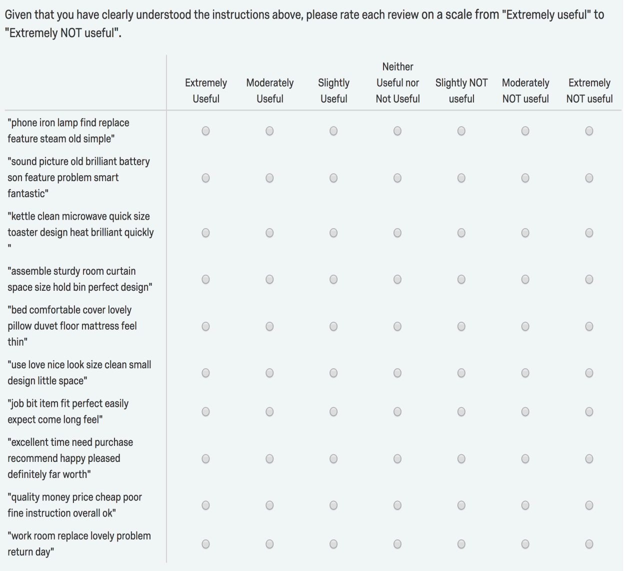
==

Given that you have understood the examples above, please take a look at the following set of words and identify the number of distinct high-level concepts in each set.
Topic 1: phone iron lamp find replace feature steam old simple
Topic 2: kettle clean microwave quick size toaster design heat brilliant quickly
Topic 3: use love nice look size clean small design little space
Topic 4: quality money price cheap poor fine instruction overall ok
Topic 5: assemble sturdy room curtain space size hold bin perfect design
Topic 6: job bit item fit perfect easily expect come long feel
Topic 7: excellent time need purchase recommend happy pleased definitely far worth
Topic 8: sound picture old brilliant battery son feature problem smart fantastic
Topic 9: bed comfortable cover lovely pillow duvet floor mattress feel thin
Topic 10: work problem return replace lovely room day

Survey 2

We are interested in studying how useful certain product reviews are for making a purchase decision on an e-commerce site. Imagine that you are shopping for a particular product on Amazon.com. You are already decided on a product to purchase and are comparing several different options. To make a better decision, you decide to read customer generated reviews for more information regarding all different aspects of products.

For hypothetical 10 reviews, we provide few topic keywords about the review. Please first look at all 10 topic description of product reviews. Please rate them on a scale from **"Extremely useful"** to "**Extremely NOT useful**" on whether you would choose to read the reviews to make a purchase decision.



Given that you have clearly understood the instructions above, please rate each review on a scale from "Extremely useful" to "Extremely NOT useful".

	Extremely Useful	Moderately Useful	Slightly Useful	Neither Useful nor Not Useful	Slightly NOT useful	Moderately NOT useful	Extremely NOT useful
"phone iron lamp find replace feature steam old simple"	○	○	○	○	○	○	○
"sound picture old brilliant battery son feature problem smart fantastic"	○	○	○	○	○	○	○
"kettle clean microwave quick size toaster design heat brilliant quickly "	○	○	○	○	○	○	○
"assemble sturdy room curtain space size hold bin perfect design"	○	○	○	○	○	○	○
"bed comfortable cover lovely pillow duvet floor mattress feel thin"	○	○	○	○	○	○	○
"use love nice look size clean small design little space"	○	○	○	○	○	○	○
"job bit item fit perfect easily expect come long feel"	○	○	○	○	○	○	○
"excellent time need purchase recommend happy pleased definitely far worth"	○	○	○	○	○	○	○
"quality money price cheap poor fine instruction overall ok"	○	○	○	○	○	○	○
"work room replace lovely problem return day"	○	○	○	○	○	○	○

Appendix C-Review Examples for Case Study and Visualization

We present review examples for case study and visualization. The last sigmoid layer ($\sigma(.)$) of FCM enables us to measure the impact of documents on the business outcome; in other words, how the predicted conversion rate changes after the consumer reads a particular document (series of reviews). To do this, we first calculate the predicted conversion rate $Conv\hat{e}rsion$ using the true document-concept distribution $DocConceptD$. Then, holding explanatory variables the same, we compare the predicted conversion rate using the average document-concept distribution $\overline{DocConceptD}$, which is calculated by averaging the document-concept across all the available documents in the training set. The difference in these two predicted conversions can be interpreted as the impact of reviews on the consumer's decision to purchase, as represented in Equation 18.

$$\Delta Conversion = \sigma(\theta_0 + \sum_i \theta_i DocConceptD_i + \sum_j \theta_j ExpVar_j) - \sigma(\theta_0 + \sum_i \theta_i \overline{DocConceptD}_i + \sum_j \theta_j ExpVar_j) \quad (18)$$

The differences in the two predicted conversion rates are calculated across all documents in the dataset.

Table 10 shows six reviews along with the product information, predicted conversion difference, and concept distribution assessed by FCM. For diversity of examples, we randomly choose two reviews with the most positive predicted conversion difference (i.e., the most likely to have increased conversion), the two most negative, and the two most neutral. For each phrase that matches a dimension of concept, we manually apply a tag that indicates the corresponding concept right after it. Comparing the tagged reviews and the estimated concept distribution shows that the concept distributions assessed by FCM are generally well aligned with the actual semantics of the reviews.

The predicted conversion difference captures the correlation between the textual data and the conversion beyond other explanatory variables. For example, without utilizing the textual data, the FCM would have predicted that the "Single Door Display Cabinet" would have a low conversion probability. However, with text, FCM predicts a high conversion rate. In fact, it predicts a 99.34% higher probability compared to the review-blind model. The increase in conversion probability seems to stem from reviews that discuss features at length.

However, the textual value becomes less obvious when a review embodies a compound of several high-volume concepts. From a review of a Hoover vacuum cleaner, FCM reports that reading reviews might have no impact on conversion. Although the review discusses the product's features in detail, (which, by our model, is likely to stimulate the conversion), it also mentions the concepts for serviceability, such as "return" or "refund", which offsets the stimulus of feature-related concepts.

Product	Title	Review with Manually Tagged Concepts	Predicted Conversion Difference	Assessed Concept Distribution
Single Glass Door Display Cabinet - Black	Useful storage	I needed a replacement for an old corner unit which used to house small momentoes we collect from our travels, both in the UK and abroad.This little unit is perfect. It has7 glass shelves and of course the floor of the cabinet to display items <FEATURE>. I needed something with more shelving rather than a unit with 3 larger spaced shelves.Fits nicely <CONFORMANCE> in the same corner and has a light in the top.My husband put it together fairly easily <FEATURE>, with my son helping when putting glass in place <CONFORMANCE>. Atlhough good value, it is lightweight and the glass is fairly thin<AESTHETICS>. Comes with fixing strap to hold against the wall <FEATURE> if required.Quick delivery.	99.34%	Aesthetics: 10.27% Conformance: 34.01% Features: 35.18% Value: 10.27% Serviceability: 10.27%
Hoover Turbo Power Bagless Upright Vacuum Cleaner	replacement hoover	Easy to assemble <FEATURE> and light weight<AESTHETIC>. The extension <FEATURE> hose for the stairs is a great. That's the good points. The suction <FEATURE> is not that great, especially if you have pet hair to remove. Difficult to see <AESTHETIC> through the cylinder to see if it needs emptying. On the 3rd time I used it the belt snapped. I returned it to Argos and got a full refund <SERVICEABILITY>.	˜ 0	Aesthetics: 10.18% Conformance: 6.13% Features: 43.76% Value: 18.88% Serviceability: 21.04%
Challenge Xtreme 70 Piece Titanium Drill Driver Set	good variety but breaks quickly	We bought the drill driver set to utilize <FEATURE> some of the parts for building <FEATURE> flat pack furniture as well as outdoor decking <FEATURE> . The variety and amount of bits is great <CONFORMANCE> but unfortunately the pieces break <SERVICEABILITY> very quickly and easily. The screwdriver heads wear out <AESTHETICS> rapidly and the drill bits break even when drilling into soft woods.	˜ 0	Aesthetics: 12.95% Conformance:12.95% Features: 48.19% Value: 12.95% Serviceability: 12.95%
2m Heavy-weight PVC Curtain Track - White	Has a drawback	I bought this to hang curtains with 2 sets of linings fitted (thermal and blackout) when these linings are fitted <CONFORMANCE> you have to use the top row of the tape on the curtains to hang them. This makes the curtain sit low <AESTHETICS> on the rail and causes <SERVICEABILITY> a gap between the curtain and rail which allows light in. I got round this by using a moulded skirting board fitted to the rail a bit like a pelmet, it works for me. They rail itself is really easy <FEATURE> to adjust to the correct size and to fit .	-79.48%	Aesthetics: 20.92% Conformance: 28.17% Features: 24.46% Value: 10.92% Serviceability: 15.53%
Eve 3 Light Ceiling Fitting - Clear	lovely light but...	I bought 2 eve lights for my narrow hall and was pleased <CONFORMANCE> with them so much I bought another 2 for my living room. However, I am so disappointed <SERVICEABILITY> that although the sun ray effects on the ceiling is lovely <AESTHETIC> -the rest of the ceiling is very dark(room size 12ftx15ft) They also cast <FEATURES> massive gloomy shadows on the walls which are driving me mad and I am going to replace <SERVICEABILITY> them. In themselves - the lights are lovely and a bargain <VALUE> but they are only good enough for narrow spaces like landings and halls.	-75.25%	Aesthetics: 12.21% Conformance:29.50% Features: 17.99% Value: 17.67% Serviceability: 22.63%

Table 10: **Example Reviews and Concept Tagged by FCM.**

Appendix D - Concept-Positional Importance

We further investigate if certain concepts matter more if presented on different review pages. That is, does the order in which certain review concepts are shown to customers correlate with more or less conversion? We enhanced our prediction model by breaking a journey-level concept distribution into two position-level concept distributions. Here, the position of a review is defined as the page in which the review is presented. Recall that the reviews are presented in groups of five and only the final (latest) 10 reviews are considered. Thus, there are two positions for the reviews: the first group of five reviews on the first page and the second group of five reviews on the second page.

We further estimate the coefficients of five concept dimensions at two different positions. As shown in Figure 11, the coefficients give the relative importance of concepts at each position. The results show that the importance of all concepts decreases in absolute value. This means that the reviews the consumer read earlier (or on the first page) are more influential than those read later (or on the second page). Among the five concepts, the coefficient drops in features and value are especially significant. From the perspective of business owners, it might be a good strategy to display more reviews concentrated on features and value earlier to extract as much conversion as possible, rather than displaying reviews concentrated in conformance and serviceability, which may be so critical that it does not matter where they are presented.

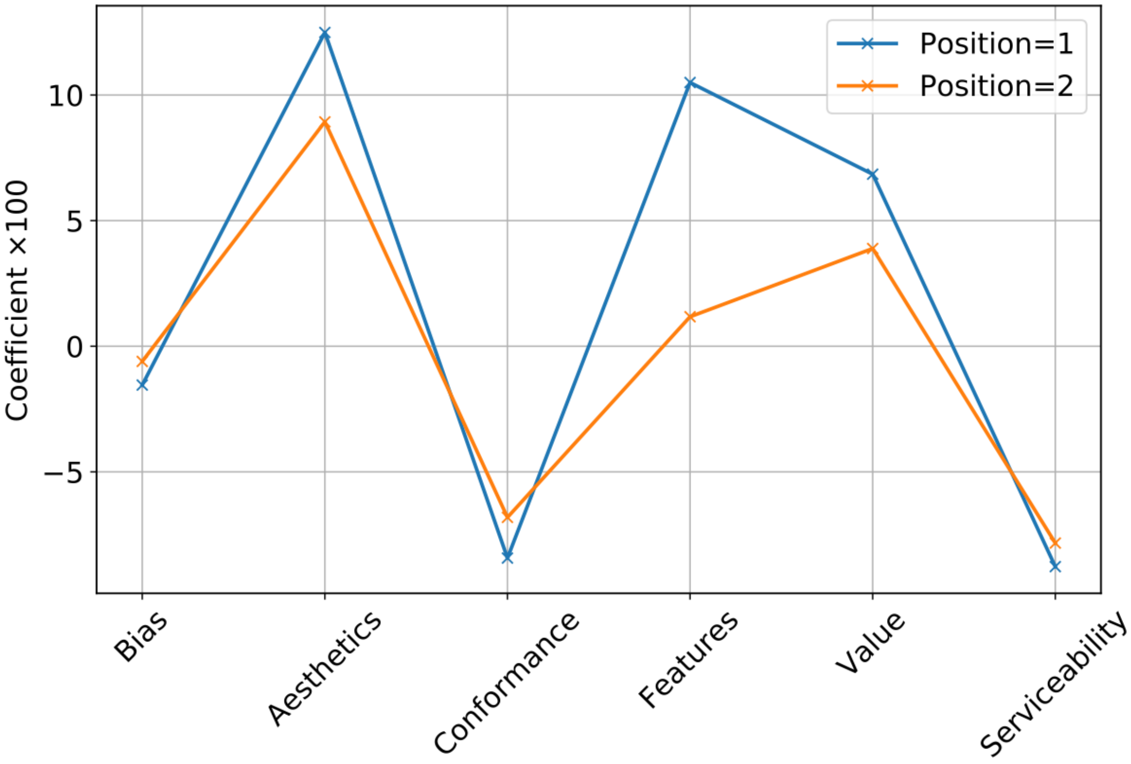


Figure 11: **Positional Importance of Concepts.**

Appendix E - 20-Newsgroup Data Performance

To demonstrate the predictive performance of FCM on a publicly available dataset, we train and evaluate FCM on the 20newsgroups dataset. The 20newsgroups dataset consists of 20 collections of documents, each of which contains 1,000 emails from a single newsgroup. Each newsgroup is associated with some topic (such as science, politics, computer graphics, etc.), which is also used as the label for all the documents within the newsgroup. The newsgroups may be broadly categorized as in Table 11. We evaluate FCM on the binary classification task of distinguishing between emails from a pair of different newsgroups. Instead of evaluating on every pair of newsgroups (which is a total of 190 pairs), we select a single newsgroup from each of the 5 broad categories in Table 11 (selected newsgroups emphasized in italics), and evaluate on all 10 pairs derived from the selected newsgroups. For comparision, we use the XGBoost classifier. We tuned the XGBoost hyperparameters to perform the best on the test data, setting the maximum depth to 1 and eta to 0.1 for 1000 training iterations. For both FCM and the XGBoost, we report area-under-the-curve classification metrics (ROC-AUC and average-precision) as well as thresholded classification metrics (precision, recall, F1-score and accuracy) in Tables 12 and 13. We find that FCM performs on-par with or better than XGBoost for a majority of the newsgroup pairs on all metrics.

Subject	Newsgroups
Computers (Comp)	*comp.graphics*, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware comp.sys.mac.hardware, comp.windows.x
Recreation (Rec)	*rec.autos*, rec.motorcycles, rec.sport.baseball, rec.sport.hockey
Science (Sci)	*sci.med*, sci.crypt, sci.electronics, sci.space
Politics (Pol)	*talk.politics.mideast*, talk.politics.guns, talk.politics.misc
Religion (Rel)	*talk.religion.misc*, alt.atheism, soc.religion.christian

Table 11: Newsgroups in the 20newsgroups dataset grouped by category. Selected representative dataset is in italics.

	ROC-AUC		F1-Score		Average Precision	
Dataset	FCM	XGB	FCM	XGB	FCM	XGB
Comp + Rec	**0.966**	0.915	**0.903**	0.810	**0.952**	0.922
Comp + Sci	**0.965**	0.937	**0.891**	0.852	**0.961**	0.937
Comp + Pol	**0.991**	0.948	**0.956**	0.838	**0.989**	0.946
Comp + Rel	**0.972**	0.891	**0.904**	0.780	**0.959**	0.899
Rec + Sci	**0.944**	0.866	**0.878**	0.796	**0.945**	0.853
Rec + Pol	**0.983**	0.944	**0.932**	0.868	**0.982**	0.933
Rec + Rel	**0.952**	0.908	**0.841**	0.809	**0.941**	0.913
Sci + Pol	**0.978**	0.932	**0.925**	0.819	**0.975**	0.934
Sci + Rel	0.841	**0.958**	0.684	**0.878**	0.809	**0.961**
Pol + Rel	0.942	**0.970**	0.833	**0.892**	0.920	**0.973**

Table 12: (Classification Metrics) Area under the ROC curve (AUC), average precision (AP) and F1-score (F1) for each dataset and method. 1.000 is the best score for all metrics. Best method for each dataset is in bold.

	Accuracy		Precision		Recall	
Dataset	FCM	XGB	FCM	XGB	FCM	XGB
Comp + Rec	**0.903**	0.826	0.885	**0.886**	**0.922**	0.746
Comp + Sci	**0.892**	0.856	**0.892**	0.885	**0.890**	0.822
Comp + Pol	**0.957**	0.849	**0.956**	0.892	**0.956**	0.790
Comp + Rel	**0.927**	0.800	**0.926**	0.865	**0.883**	0.711
Rec + Sci	**0.876**	0.787	**0.879**	0.770	**0.877**	0.825
Rec + Pol	**0.932**	0.865	**0.936**	0.839	**0.928**	0.899
Rec + Rel	**0.878**	0.820	**0.881**	0.869	**0.804**	0.756
Sci + Pol	**0.927**	0.832	**0.933**	0.864	**0.997**	0.779
Sci + Rel	0.800	**0.884**	0.893	**0.916**	**0.918**	0.844
Pol + Rel	0.869	**0.895**	0.845	**0.936**	0.821	**0.851**

Table 13: (Classification Metrics) Accuracy, precision and recall for each dataset and method with the prediction threshold fixed at 0.5. The threshold is not tuned for any metric. 1.000 is the best score for all metrics. Best method for each dataset and metric is in bold.