Computational Linguistic Analysis of Submitted SEC Information (CLASSI)

FNU Shweta
Viterbi School of Engineering
University of Southern California
Los Angeles, California USA
s779682@usc.edu

Andrea Belz
Viterbi School of Engineering
University of Southern California
Los Angeles, California USA
abelz@usc.edu

***Abstract*—Text analysis is growing in research and practice of finance, management, and operations. Word associations offer deep insight at scale into dynamics, strategy, and tactics of industries, and thus automated text processing is of great interest. We report development of a new platform using a Latent Dirichlet Allocation (LDA) topic modeling process to analyze 10-K reports that publicly traded companies submit to the Securities and Exchange Commission (SEC). We describe evaluations of the system's intrinsic performance and an important external measure, the ability to sort documents into Standard Industrial Classifications (SICs), a widely used measure of industry categories. We discuss potential applications in operations, finance, and management.**

***Index Terms*—Latent Dirichlet Allocation, topic modeling, natural language processing, SIC**

I. Introduction

Text analysis lies at the heart of many central questions of business and management as strategies, plans, and performance are revealed. However, natural language processing (NLP) in business scholarship is still in its infancy, as it requires more extensive pre-processing and filtering than numeric financial data [1]. Term frequencies have been used to predict merger and acquisition outcomes [2] and stock volatility [3]. Pan et al. combined term frequencies with similarity constructs to explore firm strategies [4], whereas Basole and collaborators added network theory to these two dimensions to identify entrepreneurial systems [5].

Term frequencies and similarities can lead to significant insights, such as those of Hoberg and Phillips [6], [7]. Their methodology (denoted "HP" here) extracts words from 10-K filings for dynamic industry definitions. However, this approach does not address potential ambiguities for words used in multiple contexts. Furthermore, it does not leverage the complete information contained in inter-document word correlations evaluated on a continuous scale.

In this study, we address these concerns with a new approach based on Latent Dirichlet Allocation (LDA) [8], a popular topic modeling method extracting the statistical properties of documents, to study industries. The method estimates the topic-document and word-document distributions - in effect, modeling word frequencies with dice-like (generalized multinomial) probability distributions and identifying common co-occurrences. In essence, the LDA algorithm discovers the word-topic-document relationships to estimate the unobserved topic distribution. Among other applications, LDA has been used to study the role of financial analyst reports [9] and the firm innovation revealed therein [10]. Extensions have been used to map innovation spaces expressed in patents [11].

We use 10-K documents to create a LDA model of industries; LDA has been used previously on this data set to analyze risk disclosures [12] and has shown strong performance relative to other systems with a similar corpus [13]. We view the unobserved topic as the industry in which a firm operates, based on measuring how businesses in the same sector use similar words. To build a dictionary, we use 10-K filings, annual reports submitted in compliance with United States securities laws, a public data set used extensively in business scholarship. We implement a 80/20 train/test split architecture to evaluate the accuracy, categorizing reports by the nine relevant Standard Industry Classifications (SICs) used to sort firm industry. This paper describes the corpus construction, reports on the performance of this new topic modeling system, and suggests potential use cases.

II. Approach

Given a set of documents or reference corpus, a topic model is a construct that generates or replicates the statistical characteristics of the corpus. The model's objective is to reproduce the word frequency at the document level, supported by the insight that words do not appear strictly independently; instead, some words will occur frequently in combination ("assets" and "liabilities"), whereas others will rarely occur together ("assets" and "chocolate"). The concept linking word frequencies can be viewed as a latent or unobserved variable known as a "topic" in a so-called "bag of words" model - i.e., the order of the words in the document is irrelevant, as is the order of documents in the corpus. This approach offers the advantage that a word may belong to multiple topics, i.e., "our banks lend to businesses on river banks." In that particular case, one topic might be represented by "banks, lend" and another by "banks, river."

This research was funded in part by the National Science Foundation (NSF) I-Corps awards 1440080 and 1740721. Any opinions, findings, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the aforementioned organizations.

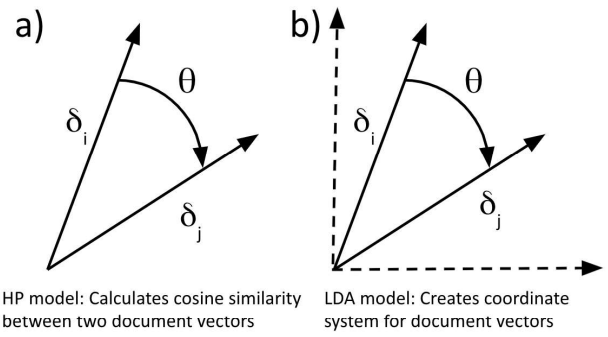

Fig. 1. Visualization of HP and LDA models.

Motivated by the HP methodology [7], we build a novel system to compare products of different companies by evaluating cosine similarities. After pre-processing, HP derive a unique vocabulary Ω from a set of documents δ. Each document δ_i is described by a vector of length Ω with π_i representing binary probabilities of appearing in a document, assigned a value of 1 if the word appears in the given document and 0 otherwise. The vector is normalized such that $\sum_\Omega \pi_\Omega = 1$. They calculate pairwise cosine similarities CS_{ij} between documents i and j to find closest matches. In effect, the HP method extracts a word's binary likelihood of appearing in a document to construct a vector of yes/no probabilities, then evaluates cosine similarities.

In a geometric view, this similarity can be viewed as a scalar product of two vectors oriented arbitrarily in a Ω-dimensional space (Fig. 1a). In our approach to LDA, we identify associations between the words and reduce the dimensionality of the problem. In this formulation, a topic effectively represents a basis vector in a space of fewer dimensions than the total vocabulary size Ω. For a system of K topics, the LDA model creates a $K-$dimensional system in which the document vectors lie (Fig. 1b). This approach offers a new interpretation for the exercise of optimizing the number of topics - namely, the objective is to estimate the space dimensionality. The topic vectors represent a set of basis vectors that span it the K-dimensional space. When the number of topics is too small, it is equivalent to insufficiently describing a three-dimensional space with only x and y. Conversely, assigning too high a value to K could be viewed as analogous to creating a four-dimensional coordinate system for three-dimensional space; the basis set has redundancies and the space is overspecified. The intuitive interpretation of the reduction of the space size is that the possible relationships between documents can be expressed by a smaller set of possibilities than that defined by each individual word; for instance, "gas" and "oil" are more likely to refer to the same general idea rather than two completely independent concepts.

Because of the aforementioned ambiguities, such as the "banks-lend"/"banks-river" case, the LDA topics will not be completely orthogonal; i.e., some words will appear with non-zero weights in multiple topics. However, for cases where the number of unique words is large relative to the number of topics (this ratio is approximately 400 in our case), overlaps in one or two dimensions still leads to topics that are relatively independent. Moreover, our formulation enables explicit testing of this assumption.

Our system offers a number of new capabilities. Importantly, we enable a finer measure of text co-occurrence than that of the HP system, which records binary probabilities if a word appears. This creates a sensitive analysis where "electric" can be linked to "cars" as well as to "utilities", two distinct industries. Second, we enable new capabilities in studying conglomerates because a document can contain multiple topics; for instance, Google's search business could be compared to Microsoft's, whereas its driverless car effort can be compared to Ford's. In addition, we retain the benefits of a customized dictionary, such as recognizing that "liability" is a financial term without the connotations of a more general use of the word [14]. Finally, the machine learning approach lends itself naturally to a train/test architecture and typical performance studies. Our dictionary can be used to evaluate different types of texts rather than just the reports. These benefits for operations and management research are significant and drive our study.

III. Developing a Topic Model

Topic modeling may be executed with various approaches, particularly Markov Chain Monte Carlo (MCMC) methods; fast Gibbs sampling has been used extensively in LDA applications adjacent to finance and accounting [12], [15], [16]. On the other hand, for large data sets, variational inference (VI) models are faster because they are optimization models rather than sampling schemes [17]. With a corpus of roughly 5,500 documents, we determined that the `gensim` VI model [18] was appropriate. Unless specifically noted, we use `gensim` data cleaning, processing, and topic modeling packages.

A. Formulation

Following related work [9], we seek to estimate the vocabulary representation in the reference corpus as multinomial distributions $\mathcal{M}$. This approach offers the advantage that in Bayesian statistical modeling, the conjugate prior of $\mathcal{M}$ is given by a Dirichlet distribution $\mathcal{D}(\delta)$ with hyperparameters δ. This multinomial-Dirichlet formulation allows for the use of standard numerical methods to improve the estimation of the hyperparameters δ until a convergence limit or other figure of merit is achieved.

Formally, the corpus $\mathcal{C}$ is defined by a vocabulary $\mathcal{V}$ consisting of unique words $\{v_1, v_2, \cdots, v_V\}$, and the set of documents containing $\mathcal{V}$ is given by $D = \{d_1, d_2, \cdots, d_D\}$. The latent topic set K links the words and documents through the probability distributions $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. $\theta_{d,k}$ describes the probability of topic k appearing in document d and thus the full $\boldsymbol{\theta}$ is represented as a D x K matrix; similarly, $\phi_{k,v}$ gives the probability that topic k is represented in vocabulary word v, so that the full $\boldsymbol{\phi}$ matrix is K x V in dimension. These probability distributions are estimated as Dirichlet functions $\mathcal{D}$

TABLE I
CORPUS CONSTRUCTION

Documents	Number
10-K documents	6,831
Documents lacking Part I	-676
Empty business text	-209
Duplicated CIKs (a newer version was filed later in the period)	-61
Duplicated+similar texts	-188
SICs not assigned	-95
Documents available	5,602
Training corpus $\mathcal{C}$	4,496
Testing data $\mathcal{T}$	1,106

TABLE II
SIC CATEGORIES AND CORPUS DISTRIBUTION

Label	Category	Total	Train	Test
A	Agriculture, Forestry and Fishing	30	25	5
B	Mining	343	275	68
C	Construction	58	48	10
D	Manufacturing	2,079	1,665	414
E	Transportation, Communications, Electric, Gas, And Sanitary Services	388	313	75
F	Wholesale Trade	145	118	27
G	Retail Trade	289	233	56
H	Finance, Insurance and Real Estate	1,306	1,046	260
I	Services	964	773	191
J	Public Administration	0	0	0
Total		5,602	4,496	1,106

so that $\boldsymbol{\theta} \sim \mathcal{D}(\alpha)$ and $\boldsymbol{\phi} \sim \mathcal{D}(\beta)$ for hyperparameters α and β. The terms α and β are varied until a convergence criterion is achieved. In other words, the analysis of the reference corpus proceeds as follows:

1) For a document $d \in \{1, 2, \cdots, D\}$ estimate a topic distribution $\theta_d \sim \mathcal{D}(\alpha)$;
2) For a topic $k \in \{1, 2, \cdots, K\}$, estimate a word distribution $\phi_k \sim \mathcal{D}(\beta)$;
3) Compare measured distributions with the model; and
4) Iterate α and β until figure of merit is reached.

B. Reference documents

Frequently, analysis of business texts requires specific dictionaries [19]–[22] to resolve context-specific discrepancies. To build a dictionary, we extracted 6,831 10-K reports filed between July 1, 2018, and June 30, 2019 from the Securities and Exchange Commission (SEC) web site (Tab. I).

After removing the images and tables, we created a customized parser to extract the business summary in Part I and labeled the report by the firm name, its unique Central Index Key (CIK), and the Standard Industrial Classification (SIC). A total of 676 documents lacking Part I were removed from the corpus, as were 209 documents with empty business text. Sixty-one reports were excluded because the company filed a later report in the period of our study. In 188 documents, we discovered exactly or nearly duplicated texts with different company names (typically in real estate); we retained only the last record. In 95 documents, the SIC code could not be extracted and thus they were excluded. This left 5,602 documents for analysis and potential sorting into either the corpus used to train the model ($\approx 80\%$) or those used for testing ($\approx 20\%$).

The candidate reports were sorted into the self-reported SIC categories defined by the SEC[1] (Tab. II) such that 4,496 comprised the main training corpus $\mathcal{C}$ for constructing the vocabulary, and the remaining 1,106 in each category were reserved as a test set $\mathcal{T}$. This train/test split was allocated on a proportional basis across SIC codes and submission quarter. One of the categories, Public Administration (J), did not contain documents and was excluded from later analysis with no impact.

[1]https://guides.loc.gov/industry-research/classification-sic

TABLE III
VOCABULARY EXTRACTION SUMMARY

Tokens	Number
Unigrams extracted from corpus	121,955
Bigrams added	8,296
Total unique token candidates	130,251
Stopwords, months removed	-3,784
Stemming and consolidation into types	-27,735
Frequent types (appear in <30%)	-841
Infrequent types (appear in less than 2 documents)	-54,509
Analysis vocabulary	43,382

C. Pre-processing

The `gensim` packgage was used to help prepare data for LDA analysis. We removed email addresses, new line characters, single quotes, non-standard (non-ASCII) characters, and proper nouns; and we changed hyphens to underscores. Bigrams were created for words commonly grouped together. The set of candidate words and bigrams formed just over 130,000 candidate unique types.

In addition to `ntlk` toolkit [23] stop words we manually removed the names of the months. We used the Snowball Stemmer for English package [24] for stemming; and manually some stemmed words, such as "optic*". For consideration in the vocabulary, the type was required to appear in more than two documents, but fewer than 30%. The final vocabulary consisted of 43,382 unique words (Tab. III), generally consistent with the 50-60,000 previously identified in the HP model [7].

IV. RESULTS

A. Parameter optimization

As an unsupervised learning process, LDA requires a so-called "burn-in" stage because the system is initialized in a random state, likely far from its optimum. System performance increases sharply with the number of iterations I at small values, then stabilizes. The objective is to generate satisfactory performance with a minimum number of iterations, as they impact processing time and after burn-in, other factors may limit system performance more dramatically than increased processing. In parallel, we sought to optimize the number of topics K based on the idea that a topic broadly represents an industry, which may comprise roughly 50 firms. Therefore, the number of topics in a set of 5,000 documents should be

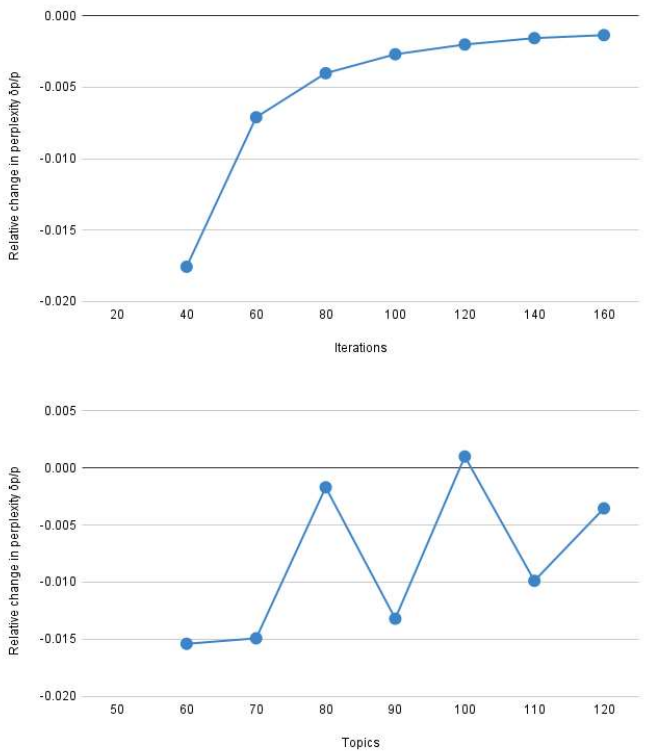

Fig. 2. Variation of the relative change in perplexity dp/p. Top: Variation with the number of iterations at fixed number of topics ($K = 90$). Bottom: Variation with the number of topics ($I = 100$)

≈ 100. As discussed in Sec. II, if K is too small, the topics are too broad; and if it is too high, the topics are redundant and thus imprecise. While some formulations determine the number of topics directly from the corpus [11]; we explicitly and systematically vary the number of topics.

As a system performance measure, we estimated the perplexity p (i.e., transformed entropy [25]), and in particular the relative change in perplexity $\delta p/p$ with respect to changes in either I or K. We optimized I by setting $K = 90$ and observed the expected asymptotic behavior (Fig. 2), electing to set I to 100 as $\delta p/p$ had levelled. This optimal value for I was independent of the value of K.

We conducted a similar exercise for the number of topics in increments of 10 and saw oscillating behavior (Fig. 2), but with a total amplitude $< 1.5\%$ for $\delta p/p$. Detailed tests indicated that when $K \approx 200$, the topics showed high levels of redundancy; for instance, in one topic, the types with the highest weights would be (in descending order) *patient, clinic, cancer* and another topic would be described by *clinic, cancer, patient*. This indicated that the system was effectively rearranging words to try to identify distinct models, and that the model was overspecified. We did not see this behavior in the vicinity of $K = 90$.

1) Average cosine similarities.: Another way to confirm that the model was not over-specified was to evaluate the average cosine similarity of the set of all topic pairs. A model with K topics has $K(K-1)/2$ unique pairs of topics. The cosine similarity γ can be calculated for each pair, and an average cosine similarity $\bar{\gamma}$ can be estimated for all pairs to characterize the model. An over-specified model will show a higher value for this quantity as the topics overlap further. We evaluated this for a series of models with constant I and varying K, and we confirmed relatively flat behavior.

Manual inspection. In each of the SIC groups $\mathcal{G}$, we examined the topic most commonly represented by the documents in that category (i.e., the topic appearing at the highest frequency with the highest weight). Manual examination suggests that the words are indeed related to each specific group; Tab. IV shows the three tokens with the highest frequency in the top topic.

TABLE IV
SIC DIVISIONS AND TOKENS WITH HIGHEST FREQUENCY IN THE TOP TOPIC

Label	Category	w1	w2	w3
A	Agriculture etc.	plant	agricultur	crop
B	Mining	gas	oil	drill
C	Construction	mine	land	expor
D	Manufacturing	clinic	patient	trial
E	Transportation, Electric, etc.	electr	gas	transmiss
F	Wholesale Trade	water	plant	residenti
G	Retail Trade	merchandis	shop	assort
H	Finance	loan	deposit	ratio
I	Services	app	marketplac	campaign

B. Train/test measures

We use the model developed with the training corpus $\mathcal{C}$ to assign labels to the 20% of the documents reserved in the test set $\mathcal{T}$ and measure the accuracy of the assignment. For each document i in $\mathcal{C}$, we identify the Most Significant Topic (MST) as the one with the highest value for θ_{ik} and denote this k'_i. Because a document is a linear combination of topics by definition, it is possible to create a synthetic document as a linear transformation of an arbitrary set of topics. For our purposes, we create a special synthetic document $\mathcal{S}_m$ to represent the mth SIC group. This synthetic document is the normalized sum of the MSTs identified for the j documents in that group ($\mathcal{S}_m = \frac{1}{j}\sum_j k'_{jm}$).

For instance, we extract each SIC "A" document (i.e., those defined as agriculture) from the training corpus $\mathcal{C}$ and identify the MST k', the topic with the highest weight, for each document. We then construct a synthetic document $\mathcal{S}_A$ to represent category A as the sum of each of these k'. We have $m = 9$ SIC groups because Category J, Public Administration, is not represented in the data. Therefore, we create nine synthetic documents $\mathcal{S}_m$, each of which is a vector of length k (the number of topics) and formulated as a normalized linear combination of the MST for all documents within that group.

A new document can be extracted from the test set $d_\mathcal{T}$ and decomposed in terms of the model's k topics. The new test document $d_\mathcal{T}$ is thus represented as a vector of length k. The goal is to assign the test document $d_\mathcal{T}$ to an SIC category. We do this by calculating cosine similarities with the synthetic representation of each of the m categories and finding the

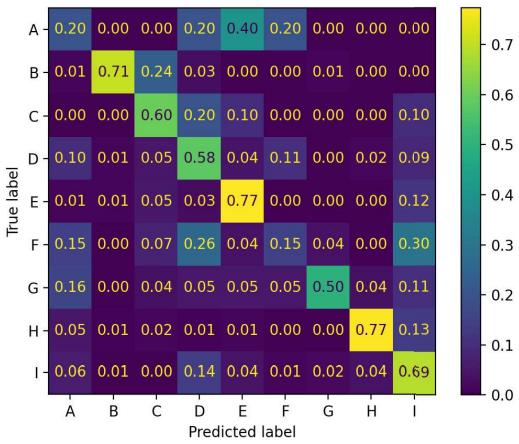

Fig. 3. Normalized confusion matrix for the predicted (horizontal) and true (vertical) SIC categorizations of Tab. II.

maximum value. In other words, we evaluate the m cosine similarities $\gamma_m = \sum_k d_{\mathcal{T}k}\mathcal{S}_{km}$ with each of the m synthetic documents, find the maximum value, and assign $d_{\mathcal{T}}$ to that group; i.e., the appropriate m of $d_{\mathcal{T}}$ is given by the value of m associated with max(γ_m).

Because the true SIC code of $d_{\mathcal{T}}$ is known, it is possible to compare the true and predicted values with a confusion matrix [26], shown in Fig. 3. The diagonal elements indicate that the lowest accuracy was seen in Category F (Wholesale Trade), whose documents represented approximately 2.5% of the total corpus - although notably, the least populated category, A (Agriculture), showed an accuracy of 20%, and suggesting that the confusion did not result from statistical limitations. The overall accuracy was approximately 65%. This is broadly consistent with the results of Hoberg and Phillips [7] that the overlap between 10-K text analysis and SIC codes is on the order of 45%, as well as observations that text-based similarity better describes industry momentum than SIC [27], [28]. Indeed, by their nature, SIC indicators lag and thus do not reflect the rapidly changing business landscape [29]. This is revealed in the present study; e.g., a significant fraction of the true Category F (Wholesale Trade) are misclassified as Category I (Services). However, the word "app", which could presumably apply to many industries, is the most highly weighted token in Category I (Tab. IV). Therefore, the measurement we report of 65% is reasonable, and we demonstrate that we go beyond the HP analysis to that of conglomerates and other types of entities.

C. Robustness checks

1) K-fold cross-validation: To ensure that these metrics were consistent across the set, we performed a cross-validation using $K = 2$ folds. We maintained the 80/20 train/test split, but extracted a different set of documents to represent the test set $\mathcal{T}$ and found an accuracy of 62%, comparable to that of the first split. We estimated the perplexity of this second split to be highly consistent to $<< 1\%$ with that of the first 80/20 split.

2) Changing train/test split: We repeated the two-fold cross-validation using a 90/10 split and a 70/30 split and found consistent accuracy values of 63% and 65%, respectively. The consistency of the 70/30 split, in particular, echoed the finding that the model was not limited by a paucity of training data.

3) Precision, sensitivity, and F-score: As a robustness check, we estimated another set of standard classification measures: precision (fraction of those that we classified that indeed belong to the category), sensitivity (also known as recall; fraction of those belonging to the category that we actually labeled), and F-score (a combined metric). These values were 77%, 65%, and 69%, respectively, in general agreement with the 65% accuracy measures.

V. Discussion

The topic models and methods described here can relate to a number of analytical exercises. They provide a way to understand groups of words that, when they appear jointly, convey significant meaning. Several applications of the specific system developed here are evident.

First, topic models can track industry evolution. For instance, one can imagine that the words "horse" and "buggy" were more commonly associated with transportation 100 years ago, versus "electric" and "car" today. Studies on the pace of innovation could benefit from simply tracking the rate at which words manifest within a topic. Another way to expand this research is to explore it on a finer scale than SICs, such as with the North American Industry Classification System (NAICS). This could be used to study the development of subfields, such as electric cars within the transportation field.

Similarly, the strategy of a single company could be examined, such as tracking General Electric through the years. The ability to assess conglomerate evolution through automated methods has not previously been reported. Yet another application could be to evaluate the alignment of other texts, such as analyst reports or web sites, with the financial reports to assess the way in which corporate communications align with regulated documents.

VI. Conclusion

Text analysis is an important element of finance and management research, and yet scholarship leveraging automated text processing is still embryonic. We use Latent Dirichlet Allocation as a method to identify companies that operate in the same industry. We report both qualitative and quantitative measures that our model is accurate and robust. This platform has a number of important applications in scholarship and business operations.

Acknowledgment

A.B. served previously as co-Principal Investigator (PI) and Research PI on the awards acknowledged herein. She now serves as Division Director of Industrial Innovation and

Partnerships at the National Science Foundation. To manage the potential conflicts of interest she resigned from all roles associated with the awards named here and is recused from NSF matters related to them. The authors thank Alexandra Graddy-Reed, Fernando Zapatero, Garima Adlakha, Chaitanya Ambegaonkar, Vaibhav Desai, Haoyuan Liu, Karan Nair, Anusha Ramakrishnan, Devansh Shah, Swetha Sivakumar, and the Management of INnovation, Entrepreneurial Research, and Venture Analysis (MINERVA) lab.

REFERENCES

[1] M. Gentzkow, B. Kelly, and M. Taddy, "Text as data." *Journal of Economic Literature*, vol. 57, no. 3, pp. 535 – 574, 2019.

[2] B. R. Routledge, S. Sacchetto, and N. A. Smith, "Predicting merger targets and acquirers from text," Working Paper, Carnegie Mellon University, Tech. Rep., 2017.

[3] S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith, "Predicting risk from financial reports with regression," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 272–280.

[4] Y. Pan, P. Huang, and A. Gopal, "Storm clouds on the horizon? new entry threats and r&d investments in the us it industry," *Information Systems Research*, vol. 30, no. 2, pp. 540–562, 2019.

[5] R. C. Basole, H. Park, and R. O. Chao, "Visual analysis of venture similarity in entrepreneurial ecosystems," *IEEE Transactions on Engineering Management*, vol. 66, no. 4, pp. 568–582, 2019.

[6] G. Hoberg and G. Phillips, "Product market synergies and competition in mergers and acquisitions: A text-based analysis," *Review of Financial Studies*, vol. 23, no. 10, pp. 3773–3811, 2010.

[7] G. Hoberg and G. M. Phillips, "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, vol. 124, no. 5, pp. 1423–1465, 2016.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] A. H. Huang, R. Lehavy, A. Y. Zang, and R. Zheng, "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science*, vol. 64, no. 6, pp. 2833–2855, 2018.

[10] G. Bellstam, S. Bhagat, and J. A. Cookson, "A text-based analysis of corporate innovation," *Management Science*, vol. 67, no. 7, pp. 4004–4031, 2021. [Online]. Available: https://doi.org/10.1287/mnsc.2020.3682

[11] F. Teodoridis, J. Lu, and J. L. Furman, "Measuring the direction of innovation: Frontier tools in unassisted machine learning," *SSRN*, 2021. [Online]. Available: https://ssrn.com/abstract=3596233

[12] Y. Bao and A. Datta, "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science*, vol. 60, no. 6, pp. 1371–1391, 2014. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2014.1930

[13] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, no. July, 2012, pp. 952–961.

[14] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, vol. 46, no. 1, pp. 35–65, 2011.

[15] H. Yuan, R. Y. K. Lau, and W. Xu, "The determinants of crowdfunding success : A semantic text analytics approach," *Decision Support Systems*, vol. 91, pp. 67–76, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2016.08.001

[16] A. Saif, M. J. Ab Aziz, and N. Omar, "Reducing explicit semantic representation vectors using Latent Dirichlet Allocation," *Knowledge-Based Systems*, vol. 100, pp. 145–159, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2016.03.002

[17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017. [Online]. Available: https://doi.org/10.1080/01621459.2017.1285773

[18] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45—-50.

[19] T. Loughran and B. McDonald, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, vol. 54, no. 4, pp. 1187–1230, 2016.

[20] P. J. Stone, "Thematic text analysis: new agendas for analyzing text content," in *Text Analysis for the Social Sciences*, C. Roberts, Ed. Lawrence Erlbaum Associates, 1997, ch. 2. [Online]. Available: https://trove.nla.gov.au/version/19209764

[21] M. Caylor, M. Cecchini, and J. Winchel, "Analysts' Qualitative Statements and the Profitability of Favorable Investment Recommendations," *Accounting, Organizations and Society*, vol. 57, pp. 33–51, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.aos.2017.03.005

[22] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making Words Work: Using Financial Text as a Predictor of Financial Events," *Decision Support Systems*, vol. 50, pp. 164–175, 2010.

[23] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*, 2009.

[24] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130–137, 1980.

[25] C. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[26] K. M. Ting, *Confusion Matrix. In: Encyclopedia of Machine Learning and Data Mining*, S. C. and W. G.I, Eds. Boston, MA: Springer US, 2017.

[27] G. Hoberg and G. M. Phillips, "Text-based industry momentum," *Journal of Financial and Quantitative Analysis*, vol. 53, no. 6, pp. 2355–2388, 2018.

[28] J. Fan, K. Cohen, L. M. Shekhtman, S. Liu, J. Meng, Y. Louzoun, and S. Havlin, "Topology of products similarity network for market forecasting," *Applied Network Science*, vol. 4, no. 1, pp. 1–15, 2019.

[29] C. Bean, "Independent Review of UK Economic Statistics – Professor Sir Charles Bean," Tech. Rep. March, 2016. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/507081/29