An Unsupervised Algorithm for Qualitative Coding of Text Data: Artifact Design, Application, and Evaluation

Alysson De Oliveira Silveira and Anol Bhattacherjee(✉)

University of South Florida, Tampa, FL 33620, USA
{alysson,abhatt}@usf.edu

Abstract. This study presents an artifact for qualitative coding of large volumes of text data using automated text mining techniques. Coding is a critical component of qualitative research, where the "gold standard" involves human coders manually assigning codes to text fragments based on their subjective judgment. However, human coding is not scalable to large corpora of text with millions of large documents. Our proposed method extends the latest advancements in semantic text similarity using sentence transformers to automate qualitative coding of text for predefined constructs with known operationalizations using cosine similarity scores between individual sentences in the text documents and construct operationalizations. We illustrate our approach by coding corporate 10-K reports from US SEC filings for two organizational innovation processes: exploration and exploitation.

Keywords: Text mining · Text analytics · Natural language processing · Qualitative research · Coding · Organizational innovation · Semantic text similarity

1 Introduction

Thematic coding of text documents, such as interview transcripts, is a core process in qualitative research [1]. This process involves searching for key concepts or themes (called "codes") in blocks of texts. The coding process may be concept-driven or data-driven [1]. In concept-driven coding, we search the text for patterns of words that match a predefined set of concepts. In data-driven coding, we approach the text without any predefined conceptualization, letting the text speak for itself and allowing concepts to emerge from the text. Data-driven coding, also called open coding in grounded theory research, is used to inductively build theories, while concept-driven coding is typically used to deductively test hypotheses [1]. This study provides an algorithm for concept-driven coding of text data using automated text mining techniques.

The "gold standard" of qualitative text coding is using human coders, who must manually comb through text documents to search for relevant themes or concepts, and assigned codes to text fragments based on their subjective interpretation of the text. This

© Springer Nature Switzerland AG 2021
L. Chandra Kruse et al. (Eds.): DESRIST 2021, LNCS 12807, pp. 272–284, 2021.
https://doi.org/10.1007/978-3-030-82405-1_27

process has two limitations. First, human coding is laborious, resource-intensive, and not scalable to thousands or millions of text documents. Consequently, qualitative research tends to employ small samples, which limits the generalizability of such research. Second, while human coding works for research projects, where the data collection process is purposively designed to extract core themes relevant to the phenomenon of interest, it does not work very well for secondary text data, such as corporate reports, published news articles, or social media posts. Because secondary data are not designed for research data collection, they tend to be extremely "noisy" (with low signal content), rendering them inefficient for human coding.

Nonetheless, secondary text corpora offer interesting possibilities for information systems [2] and organizational research [3, 4]. First, large volumes of such data already exist in the form of electronic mails, social media posts, online reviews, text messages, and such. If properly mined, such huge datasets may allow for the detection of small effects, the investigation of complex relationships, comparison of sub-samples, and the study of rare phenomena [2]. Second, because secondary data are not created for research purposes, they are unbiased by researchers and the research process (e.g., Hawthorne effect), which often exists in primary data, for example, when a researcher's questions may lead an interviewee to respond in a "socially desirable" manner.

Researchers are increasingly turning to automated tools like Linguistic Inquiry and Word Count (LIWC) for coding text into predefined constructs. These are "bag of words" approaches, because they require specifying a bag of similar words (e.g., synonyms) representing a construct of interest, and the algorithm will count the number of words in that bag in a text corpus. However, words are often ambiguous and mean different things (e.g., "Apple" could refer to a technology company or a fruit), different words (e.g., automobile, car, "wheels," or "Toyota") may refer to the same object, an dthe same idea can be represented using different combinations of words (e.g., "AT&T merges with Time-Warner" and "Time-Warner is bought by AT&T"). Further, the semantic meaning of a word depends on its context in use (e.g., "battery" means very different in "My cell phone battery is low" and "He was charged of battery"). Hence, word-based automated coding like LIWC tend to do a poor job at deciphering the semantic meaning and context of text. In addition, word-based approaches suffer from high dimensionality and high sparsity, requiring substantial computational resources to process, while producing inferior results due to low signal-to-noise content.

In this paper, we present an alternative approach, where the unit of textual analysis is sentences, rather than words. As evident from the examples above, sentences provide a much better representation of semantic meaning and context than words, and help us make more sense of ambiguous natural languages than word-based approaches. Recent developments in deep learning involving sentence transformers offer new possibilities in our ability to compare sentences in large corpora of secondary text data with minimal human intervention, that were not feasible even a couple of years ago.

Although industry leaders like Google and Facebook have made tremendous progress in the development of pretrained language models, these models are largely unknown in academic research. In this paper, we demonstrate how a sample of annual 10-K reports filed by publicly traded corporations in the USA can be mined for concept-based coding of research constructs, such as exploration and exploitation – two popular types of organizational innovation. We evaluate our semantic text similarity (STS) based approach by comparing the automated coding with a manually coded subsample of the same data. The resulting codes represent whether or not the companies in our sample are engaging in exploration and/or exploitation activities, which can be used as dummy variables in statistical models to test research hypotheses involving these constructs. Our research lies at the interface of design science and qualitative research, and the automated coding approach that we developed can be extended to coding of other research constructs with minimal human intervention.

2 Related Literature

Text mining techniques, such as sentiment analysis, topic modeling, dimensionality reduction, classification, and clustering, are still novel in academic research. In one of the earliest studies in this area, Abbasi et al. [5] analyzed 300,000 web pages (and 30,000 images) to extract fraud cues from word phrases using similarity metrics, which were then used to identify fake web pages. Among more recent studies, Muller et al. [2] used topic modeling and classification to identify which aspects of product reviews users find "helpful." Others have employed similar techniques to predict stock price movements based on published news articles [6], forecast tourism demand by analyzing online travel forum data [7], identify smoking status based on online forum posts [8], identify helpful content from online knowledge community posts [9], assess user sentiments toward products or services [10, 11], identify fake online reviews [12], categorize users of online communities [13], categorize products competing for the same market [14], and analyze social interaction among top management members [15] or between guests and hosts [16]. A more detailed literature review is not presented here to conserve space but is available from the authors upon request.

The dominant approach of feature extraction in the above studies is the bag of words approach (e.g., [6–8, 10, 12, 14, 17, 18]). In this approach, the text corpus is tokened into words, followed by stopword removal, punctuation removal, text normalization (lowercasing and/or stemming or lemmatization). The remaining words are then converted into a term frequency-inverse document frequency (TF-IDF) matrix based on word co-occurrence frequency within and across documents. This approach helps identify influential words but does not preserve the ordering or semantic meaning of words. A variation of this approach uses n-grams (word sequences) instead of individual words for TF-IDF vectorization (e.g., [9, 11, 19]), which may provide slightly better performance than word-based TF-IDF in large text corpora.

Two drawbacks of the TF-IDF-based bag of words approach are (1) high-dimensional sparse vectors (with many zeroes) that are computationally inefficient to process, and (2) their inability to store the semantic content of language. To address these problems, in the mid-2010s, researchers (e.g., [16, 20]) developed pre-trained word vector models by using dense vectors that are computationally more efficient and can compare words using metrics such as Euclidean distance or cosine similarity. However, words often have multiple meanings (e.g., Apple: a fruit or a technology company?), which cannot be interpreted if a word is divorced from its context of use. As linguist Ludwig Wittgenstein once said, the meaning of a word lies in its use. Hence, sentences are a more appropriate unit of linguistic interpretation than words. However, we did not find any prior instance of sentence-based vectorization in the literature, and understandably so, given that sentence models are just about two years old and are still being developed. Our study is one of the earliest to apply this technique to information systems research.

Secondly, much of prior research has been inductive, searching for latent topics or features from text and using them to classify or cluster text documents. There is little instance of using text mining for deductive research, for example, to score specific theoretical constructs of interest for hypotheses testing. Our goal in this research is not to discover new patterns but to use text mining as a tool to support classical theory-driven, hypotheses testing research. We propose a semantic text similarity (STS) method that uses sentence as the unit of analysis, computes cosine similarity between sentence vectors in a text corpus and those in predefined operationalizations of theoretical constructs to code sentences, to mimic human coding.

3 Problem Context

We demonstrate our STS approach in the context of organizational innovation research. Two key innovation processes described in the organizational literature are exploration and exploitation. *Exploration* refers to discovering new products, new resources, new knowledge, and new opportunities that may lead to new product or service offerings or new markets. In contrast, *exploitation* refers to better utilization of existing products, existing resources, existing knowledge, and existing competencies to reduce production costs or improve efficiency [21]. The two approaches require diametrically opposite organizational structures, processes, capabilities, and cultures. Exploration requires risk-taking, experimentation, improvisations, radical changes, and chaos, while exploitation requires risk avoidance, incremental refinement, focus on efficiency, stability, and order. Exploration is generally associated with organic structures, loosely coupled systems, autonomy and chaos, and breaking new grounds, while exploitation is associated with mechanistic structures, tightly coupled systems, control and bureaucracy, and stable markets and technologies [22].

Consequently, organizations that excel in exploitation tend to struggle with exploration and vice versa. However, both approaches are important for organizations because exploitation generates current revenues, which is needed to fuel exploration for future revenues. Hence, a key theme in strategic management research is that organizations that are “ambidextrous” or can concurrently manage exploration and exploitation processes, outperform those that excel only at exploration or exploitation [23].

Empirical research have measured exploration and exploitation using multi-item, Likert-scaled instruments. A representative instrument for measuring exploration and exploitation, adapted from He and Wong [24], is shown in Table 1.

Two typical problems in survey research are common method bias and social desirability bias. Common method bias stems from the use of a common instrument (the same survey form) for measuring independent and dependent variables at the same time, while social desirability bias is a tendency among respondents to portray a positive view of themselves and their organizations, irrespective of the ground reality. Moreover, cross-sectional surveys provide a snapshot of contemporaneous levels of exploration and exploitation in organizations, but cannot provide any information on historical trajectories of such constructs in their organizations, or the extent to which organizations have built or lost exploration and exploitation capabilities over time. Mining historic 10-K reports filed by public corporations to the United States Securities and Exchange Commission (SEC) can help us reconstruct a historical fossil record of organizations and analyze innovation patterns within and across industry sectors, while avoiding common method bias and social desirability bias.

Table 1. Operational measures of exploration and exploitation

Exploration measures	Exploitation measures
Introduce new generation of products	Improve existing product quality
Extend product range	Improve production flexibility
Open up new markets	Reduce production cost
Enter new technology fields	Improve yield or reduce material consumption

*Adapted from He and Wong [24]

However, exploration and exploitation metrics are not readily available on financial statements. Although some researchers have considered research & development (R&D) expense as a measure of innovation, it is unclear whether R&D refers to exploration, exploitation, or both. Moreover, R&D is not comparable across industry sectors and many organizations, such as banks, do not have R&D expense, but still innovate in the form of online or mobile banking or new financial products. However, it may be possible to “infer” exploration and exploitation from “business” and “management discussions & analysis” (MD&A) sections of 10-K reports, where corporate management may discuss new product or market developments, product extensions, and/or internal process improvement initiatives for the benefit of their shareholders. If we can use text mining techniques to efficiently mine these text sections, we may be able to create corporate profiles of innovation across time and industry sectors.

4 Method

4.1 Data Sourcing

Data for our analysis was sourced from SEC’s Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website, which makes available all 10-K reports filed with the agency in HTML and text files. Starting with companies that filed 10-K reports with SEC during 2020, we had 5,573 reports from 5,505 unique companies. Using Standard Industrial Classification (SIC) codes, we dropped finance, insurance, and real estate, public administration, non-classifiable companies, and those with unassigned SICs. From the remaining pool of 4,147 companies, we randomly chose a sample of 201 companies. We then searched the EDGAR website for 10-K reports of these companies for each fiscal year between 2016 to 2020 to create a five-year longitudinal panel for each company’s innovation activities. Companies that did not have 10-K reports for the entire five-year period were dropped, leading to a final sample of 134 companies and 670 10-K documents. The "business" section of these documents were parsed to remove XML tags and extract the text for our text mining.

4.2 Artifact Design

Our deductive coding pipeline is shown in Fig. 1. We adopted He and Wong's [24] eight exploration and exploitation measures (Table 1) as the ontology for our automated coding process.

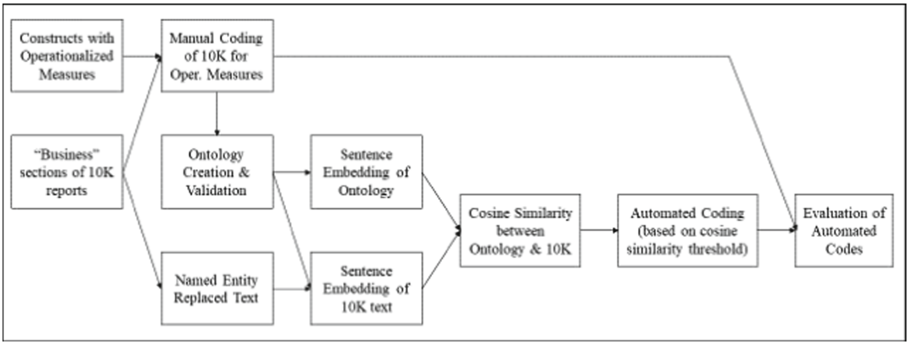

Constructs with Operationalized Measures
Manual Coding of 10K for Oper. Measures
"Business" sections of 10K reports
Ontology Creation & Validation
Sentence Embedding of Ontology
Named Entity Replaced Text
Sentence Embedding of 10K text
Cosine Similarity between Ontology & 10K
Automated Coding (based on cosine similarity threshold)
Evaluation of Automated Codes

Fig. 1. Automated deductive coding of text

The extracted text was tokenized into sentences for coding. Unlike the generic wording in our exploration and exploitation ontologies, we found that products (or services) were usually addressed in 10-K reports by their specific names, such as "iPhone 12". Because the STS model did not show good match between generic "product" and specific product names like "iPhone 12" to the generic "product", we employed a named entity recognition (NER) model to replace all product names with a generic "product".

Our next decision choice was selecting a sentence transformer model for encoding each sentence in the 10-K reports and our ontologies. Transformers are pretrained deep learning models for transforming an input sequence (e.g., text) into a different output sequence (e.g., vector) using an attention mechanism that learns contextual relations in the input sequence [25]. Among different classes of sentence transformers, Bidirectional Encoder Representations from Transformers (BERT), developed by Google AI Lab in 2019, are particularly suited for STS tasks [26]. Initially intended for language translation, the BERT model uses multi-layer bidirectional architecture and shows excellent performance across various tasks, including STS, text summarization, and autocompleting search queries [27]. Robustly optimized BERT approach (RoBERTa), retrained with a much larger text corpus, more compute power, and improved training methodology, demostrate significantly improved performance over the original BERT model [28]. SRoBERTa is further retuning of the RoBERTa model to enhance its computational efficiency. With a 10,000 sentences collection, SRoBERTa reduced the time to find the most similar pair of sentences from 65 h to about 5 s and computed cosine-similarities in approximately 0.01 s [26]. Because of the above reasons, SRoBERTa was chosen as the desired sentence embedding model in this study.

We used RoBERTa to generate sentence embeddings for each sentence in the business section of 10-K reports and the eight ontology categories for exploration and exploitation, and computed cosine similarities between text and ontology sentence vectors, with and without NER replacements (for comparison). Cosine similarity is a measure of semantic similarities between two encoded sentences obtained using Eq. 1, where $\vec{a}$ and $\vec{b}$ vectors representing sentences that we are comparing:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \left|\left|\vec{b}\right|\right|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2}\sqrt{\sum_1^n b_i^2}} \tag{1}$$

The outputs from our processing stages were a m × 8 matrix for each 10-K document, where m is the number of sentences in the business section of the 10-K, and the eight columns correspond to eight categories in our ontology. Table 2 shows that NER resulted in significant improvement in cosine similarity scores. To aggregate our analysis from the sentence-level to the organizational level, we calculated the maximum similarity between each category and all sentences in the 10-K as a measure of category code for that document as follows, where $v_{i,j}^x$ is the vector of cosine similarities for category x and all sentences in the 10-K report of company i in year j:

$$m_{i,j}^x = \text{arg max}\left(\vec{v}_{i,j}^x\right) \tag{2}$$

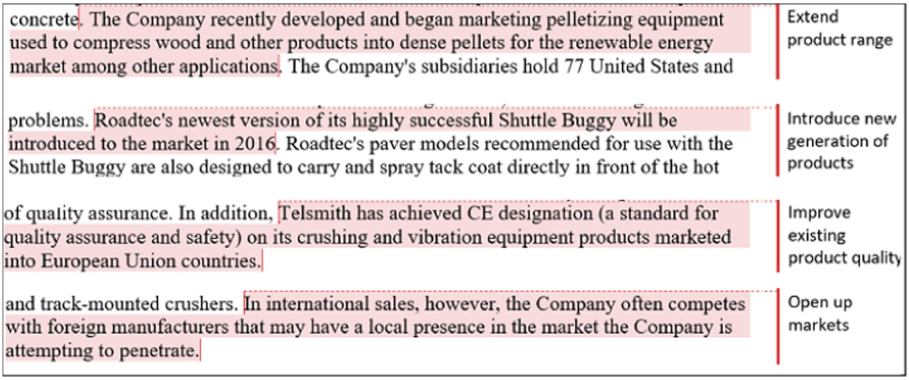
Table 2. Effect of NER on Similarity Score

Sentence without NER [Sentence with NER]	Cosine similarity	
	Without NER	With NER
During 2020, the Company released an updated iPad Pro [During date, the Company released an updated product.]	0.4285	0.8215
During 2020, the Company released AirPods Pro® [During date, the Company released product.]	0.4181	0.7210
(…) released Apple Watch Series 6 and a new Apple (…) [released product and a new product]	0.3510	0.8080

We assigned company i to belong to category x in year j if its maximum similarity $m_{i,j}^x$ with category x equaled or exceeded a certain threshold.

$$c_{i,j}^x \begin{cases} 1 \; if \; m_{i,j}^x \geq threshold \\ 0 \; if \; m_{i,j}^x < threshold \end{cases} \tag{3}$$

Determining an appropriate threshold in Eq. (3) was an important part of our methodology. To identify what threshold value would yield the best classification performance, we manually coded a randomly chosen subsample of twenty 10-K documents (3,742 sentences) in our sample into the eight categories of interest (see Fig. 2 for our sample manual coding). We used these manual codes as benchmark to assess the performance of our automated coding at different cosine similarity thresholds from 0.0 to 1.0, with intervals of 0.1. Confusion matrices for each threshold level were used to compute recall, precision, and F1-score as metrics of our classification performance. These performance metrics for exploration and exploitation are shown in Fig. 3. These plots suggested an optimum similarity threshold of 0.50 for best automated classification performance.



concrete. The Company recently developed and began marketing pelletizing equipment used to compress wood and other products into dense pellets for the renewable energy market among other applications. The Company's subsidiaries hold 77 United States and

Extend product range

problems. Roadtec's newest version of its highly successful Shuttle Buggy will be introduced to the market in 2016. Roadtec's paver models recommended for use with the Shuttle Buggy are also designed to carry and spray tack coat directly in front of the hot

Introduce new generation of products

of quality assurance. In addition, Telsmith has achieved CE designation (a standard for quality assurance and safety) on its crushing and vibration equipment products marketed into European Union countries.

Improve existing product quality

and track-mounted crushers. In international sales, however, the Company often competes with foreign manufacturers that may have a local presence in the market the Company is attempting to penetrate.

Open up markets

Fig. 2. Manual coding of text segments of 10-K Report

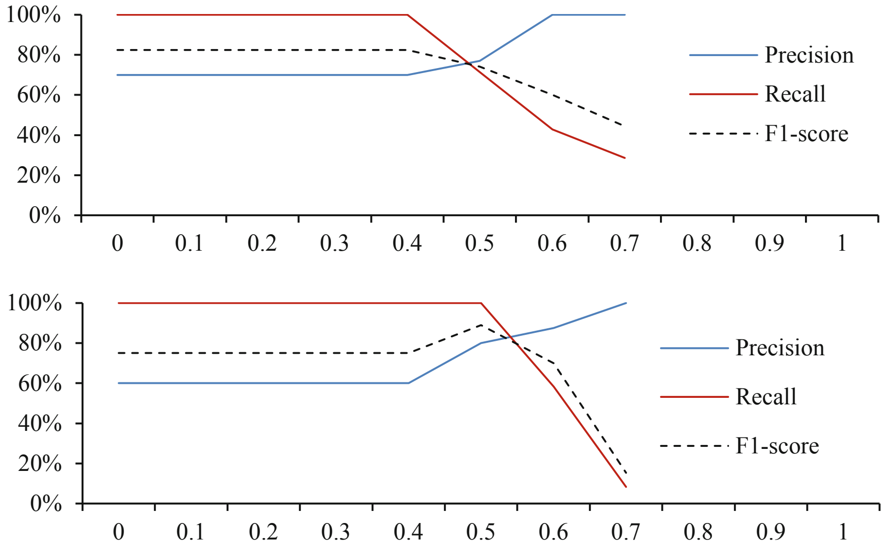


Fig. 3. Exploration (top) and exploitation (bottom) performance by cosine similarity threshold

Lastly, we classified company *i* as "explorative" in year *j*, if it belonged to at least one of the four categories: (1) introduce new generation of products, (2) extend product range, (3) open up new markets, or (4) enter new technology fields, and "exploitative" if it belongs to at least one of the categories: (5) improve existing product quality, (6) improve production flexibility, (7) reduce production cost, or (8) improve yield or reduce material consumption. This classification can be used for inductive theory building or deductive theory testing.

$$exploration_{i,j} = \arg\max(c^1_{i,\ j},\ c^2_{i,\ j},\ c^3_{i,\ j},\ c^4_{i,\ j}) \tag{4}$$

$$exploitation_{i,j} = \arg\max(c^5_{i,\ j},\ c^6_{i,\ j},\ c^7_{i,\ j},\ c^8_{i,\ j}) \tag{5}$$

Table 3. Classification comparison and performance

Company	Year	Exploration		Exploitation	
		Manual	Automated	Manual	Automated
Teradata	2018	1	1	1	1
Currency works	2018	1	1	0	0
Soligenix Inc.	2018	1	0	1	1
Beazer homes USA	2020	1	1	1	1
Cedar fair	2020	0	0	0	1
Citrix systems	2020	1	1	1	1
Defense technologies	2019	1	0	0	0
Concierge technologies	2017	0	0	0	0
Moneygram international	2016	0	1	1	1
Pc Tel	2018	1	1	0	0
Strattec security	2018	1	1	1	1
Zynga	2017	1	1	1	1
Union Pacific	2020	0	0	0	1
Telos	2019	0	1	1	1
Pattern wwnergy group	2017	1	0	1	1
Applied energetics	2020	1	0	0	0
Astec industries	2016	1	1	1	1
American superconductor	2016	1	1	1	1
Houston American energy	2019	0	1	0	1
Superior Drilling products	2017	1	1	1	1
Overall recall		0.714		1.000	
Overall precision		0.769		0.800	
Overall F1-score		0.741		0.889	

4.3 Design Evaluation

Using the cosine similarity threshold of 0.50, our manual and automated exploration and exploitation coding for our subsample of 20 companies, along with the overall recall, precision, and F1-score, are shown in Table 3. Overall F1-score for exploration was 0.74, and that for exploitation was 0.89 using our designed artifact.

A closer examination of classification performance by category reveals substantial variations in performance across categories. Among exploration measures, Category C1 (introduce new generation of products) had the highest F1-score of 0.74, while Category 4 (enter new technology fields) had the lowest F1-score of 0.33. Exploitation categories showed slightly better results, with F1-scores ranging between 0.957 for Category 5 (improve existing product quality) and 0.43 for Category 8 (improve yield or reduce material consumption). Terms like "yield" and "material consumption" did not appear to be referenced in most corporate 10-K reports.

Table 4. Performance measures by ontology category

Categories	Precision	Recall	F1-score
C1	63.64%	87.50%	73.68%
C2	71.43%	45.45%	55.56%
C3	33.33%	80.00%	47.06%
C4	28.57%	40.00%	33.33%
C5	91.67%	100.00%	95.65%
C6	33.33%	75.00%	46.15%
C7	85.71%	75.00%	80.00%
C8	27.27%	100.00%	42.86%

Note: C1 = introduce new generation of products; C2 = extend product range; C3 = open up new markets; C4 = enter new technology fields; C5 = improve existing product quality; C6 = improve production flexibility; C7 = reduce production cost; and C8 = improve yield or reduce material consumption

Given these variations, it may be prudent to use a different threshold to separately classify each of the eight categories, and then aggregate those binary classifications into ordinal measures of exploration and exploitation, rather than classify at the aggregate level of exploration and exploitation for each company. Given conference submission deadlines, we were unable to complete this analysis, but plan to present it at the conference, if accepted. We also plan to expand our manual analysis from 20 to 40 companies and employ two coders to assure intersubjectivity in our manual coding.

5 Implications and Conclusions

The goal of this study was to create an artifact that could be used to automatically code text documents for theoretical constructs with predefined operationalizations. We described the process of creating such an artifact for coding corporate 10-K filings for two types of organizational innovation: exploration and exploitation. We evaluated our artifact using a subsample of 20 manually coded organizations.

Our study proposed an unique method, employing pretrained sentence transformers for automatic coding of text documents into theoretical constructs. We view this as an important methodological contribution as it extends the labor-intensive manual coding process to large text corpora. Unlike prior word count based automated approaches like LIWC, our approach is based on sentences that captures semantic meaning and context better than words can, and is not sensitive to improper specification of bags of words. This is a very promising technique, given our large and growing corpora of user-generated text content such as SEC filings, online reviews, social media posts, and so forth.

Our proposed method (artifact) leverages known operationalizations of research constructs (e.g., exploration and exploitation) an as ontology to look for sentences in text corpora that are semantically close to sentences in the ontology. This sentence-based approach is also novel to academic research.

Of course, coding is just one step in the research process. The final goal of research is to understand phenomena of interest. Qualitative codes generated from our method can be linked to other constructs to support inductive theory building or may be used as dummy variables in statistical models for deductive theory testing.

Though this paper was a proof-of-concept for our proposed artifact, it raised several questions on its application to text coding that we want to explore next. For example, we do not know the sensitivity of our approach is to the size of input text documents. We also do not know how well this approach works if we increase the number of categories. Lastly, though we found NER to significantly improve matching of text entences to ontology, it may be argued that sentence transformers should ideally be able to match the word product with names of products and that NER may be unnecessary. Our current sentence transformer models are not yet able to match product with product names; however as we develop better pretrained models, NER may become unnecessary. We plan to explore these issues, with a bigger sample of manually and automatically coded 10-K documents in our subsequent research.

References

1. Gibbs, G.R.: Analyzing Qualitative Data, 6th edn. Sage, Thousand Oaks (2007)
2. Müller, O., Junglas, I., Brocke, J.V., Debortoli, S.: Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur. J. Inf. Syst. **25**, 289–302 (2016). https://doi.org/10.1057/ejis.2016.2
3. Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G., Den Hartog, D.N.: Text Mining in Organizational Research. Org. Res. Methods **21**, 733–765 (2018). https://doi.org/10.1177/1094428117722619
4. Janasik, N., Honkela, T., Bruun, H.: Text Mining in Qualitative Research. Organ. Res. Methods. **12**, 436–460 (2009). https://doi.org/10.1177/1094428108317202

5. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., Nunamaker, J.F.: Detecting fake websites: the contribution of statistical learning theory. Manag. Inf. Syst. Quart. **34**, 435–461 (2010). https://doi.org/10.2307/25750686
6. Nam, K.H., Seong, N.Y.: Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. Decis. Support Syst. **117**, 100–112 (2019). https://doi.org/10.1016/j.dss.2018.11.004
7. Colladon, A.F., Guardabascio, B., Innarella, R.: Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. Decis. Support Syst. **123**, 113075 (2019). https://doi.org/10.1016/j.dss.2019.113075
8. Wang, X., et al.: Mining user-generated content in an online smoking cessation community to identify smoking status: a machine learning approach. Decis. Support Syst. **116**, 26–34 (2019). https://doi.org/10.1016/j.dss.2018.10.005
9. Liu, X., Alan Wang, G., Fan, W., Zhang, Z.: Finding useful solutions in online knowledge communities: a theory-driven design and multilevel analysis. Inf. Syst. Res. **31**, 731–752 (2020). https://doi.org/10.1287/ISRE.2019.0911
10. Chatterjee, S.: Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents. Decis. Support Syst. **119**, 14–22 (2019). https://doi.org/10.1016/j.dss.2019.02.008
11. Xu, X.: What are customers commenting on, and how is their satisfaction affected? examining online reviews in the on-demand food service context. Decis. Support Syst. **142**, 113467 (2021). https://doi.org/10.1016/j.dss.2020.113467
12. Hu, N., Bose, I., Koh, N.S., Liu, L.: Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis. Support Syst. **52**, 674–684 (2012). https://doi.org/10.1016/j.dss.2011.11.002
13. Hwang, E.H., Singh, P.V., Argote, L.: Jack of all, master of some: information network and innovation in crowdsourcing communities. Inf. Syst. Res. **30**, 389–410 (2019). https://doi.org/10.1287/isre.2018.0804
14. Pan, Y., Huang, P., Gopal, A.: Storm clouds on the horizon? New entry threats and R & D investments in the U.S. IT industry. Inf. Syst. Res. 30, 540–562 (2019). https://doi.org/10.1287/isre.2018.0816.
15. Zhang, T., Liu, F.C., Gao, B., Yen, D.: Top management team social interaction and conservative reporting decision: a language style matching approach. Decis. Support Syst. **142**, 113469 (2021). https://doi.org/10.1016/j.dss.2020.113469
16. Wu, J., Cai, J., Luo, X.R., Benitez, J.: How to increase customer repeated bookings in the short-term room rental market? a large-scale granular data investigation. Decis. Support Syst. **143**, 113495 (2021). https://doi.org/10.1016/j.dss.2021.113495.
17. Cao, Q., Duan, W., Gan, Q.: Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach. Decis. Support Syst. **50**, 511–521 (2011). https://doi.org/10.1016/j.dss.2010.11.009
18. Goes, P., Lin, M., Yeung, Ching-man Au.: "Popularity Effect" in user-generated content: evidence from online product reviews. Inf. Syst. Res. **25**(2), 222–238 (2014). https://doi.org/10.1287/isre.2013.0512
19. Singh, P.V., Sahoo, N., Mukhopadhyay, T.: How to attract and retain readers in enterprise blogging? Inf. Syst. Res. **25**, 35–52 (2014). https://doi.org/10.1287/isre.2013.0509
20. Zhang, L., Yan, Q., Zhang, L.: A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behavior on Airbnb. Decis. Support Syst. **133**, 113288 (2020). https://doi.org/10.1016/j.dss.2020.113288
21. March, J.G.: Exploration and Exploitation in Organizational Learning. Organ. Sci. **2**, 71–87 (1991)
22. Lewin, A.Y., Long, C.P., Carroll, T.N.: The coevolution of new organizational forms. Organ. Sci. **10**, 535–550 (1999). https://doi.org/10.1287/orsc.10.5.535

23. Tushman, M.L., O'Reilly, C.A.: Ambidextrous organizations: managing evolutionary and revolutionary change. Calif. Manage. Rev. **38**, 8–29 (1996). https://doi.org/10.2307/41165852
24. He, Z-L., Wong, P-K.: Exploration vs. exploitation: an empirical test of the ambidexterity hypothesis. Organ Sci. **15**(4), 481–494 (2004). https://doi.org/10.1287/orsc.1040.0078
25. Vaswani, A., et al.: Attention is all you need. In: 31st Conference Neural Information Processing System (2017). https://doi.org/10.1109/2943.974352.
26. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. In: 2019 Conference Empirical Methods Natural Language Processing 9th International Jt. Conference Natural Language Processing Proceedings Conference, pp. 3982–3992 (2020). https://doi.org/10.18653/v1/d19-1410.
27. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference North American Chapter Association Computer Linguistics Human Language Technology - Proceedings Conference, vol. 1, pp. 4171–4186 (2019)
28. Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv (2019)