Journal of Big Data: Theory and Practice

Volume 1 **June 2022** **Number 1**

Jim Samuel, Ratnakar Palle, Eduardo Correa Soares

Textual Data Distributions: Kullback–Leibler Textual Distributions Contrasts on GPT-2-Generated Texts, with Supervised, Unsupervised Learning on Vaccine and Market Topics and Sentiment

Amit Mokashi, J.D. Jayaraman, Priyanka Mahakul, Rutu Patel, Anthony Picciano

Global Perception of the Belt and Road Initiative: A Natural Language Processing Approach

Miaojie Zhou, Satish Mahadevan Srinivasan, Abhishek Tripathi

Four-Class Emotion Classification Problem Using Deep Learning Classifiers

Jim Samuel, Margaret Brennan-Tonetta, Yana Samuel, Pradeep Subedi, Jack Smith

Strategies for Democratization of Supercomputing: Availability, Accessibility and Usability of High Performance Computing for Education and Practice of Big Data Analytics

Ethné Swartz, Rashmi Jain, Margaret Brennan-Tonneta, Marina Johnson, Stanislav Mamonov, Matthew Hale, J.D. Jayaraman

In Search of Pedagogical Approaches to Teaching Business Ethics in the Era of Digital Transformation

Journal of Big Data: Theory and Practice

Journal of Big Data:
Theory and Practice

Volume 1 **June 2022** **Number 1**

WWW.JBDTP.ORG

ISSN: 2692-797

JBDTP Professional Vol. 1, No. 1, 2022 DOI: 10.54116/jbdtp.v1i1.20

TEXTUAL DATA DISTRIBUTIONS: KULLBACK–LEIBLER TEXTUAL DISTRIBUTIONS CONTRASTS ON GPT-2-GENERATED TEXTS, WITH SUPERVISED, UNSUPERVISED LEARNING ON VACCINE AND MARKET TOPICS AND SENTIMENT

Jim Samuel
Rutgers University
jim@aiknowledgecenter.com

Ratnakar Palle
Apple Inc
rrpalle@gmail.com

Eduardo Correa Soares
Cerence B.V.
ecsoares@yahoo.com

ABSTRACT

Efficient textual data distributions (TDD) alignment and generation are open research problems in textual analytics and natural language processing (NLP). It is presently difficult to parsimoniously and methodologically confirm that two or more natural language datasets belong to similar distributions and to identify the extent to which textual data possess alignment. This study focuses on addressing a segment of the broader problem described above by applying multiple supervised and unsupervised machine learning (ML) methods to explore the behavior of TDD by (i) topical alignment and (ii) by sentiment alignment. Furthermore we use multiple text generation methods including fine-tuned GPT-2, to generate text by topic and by sentiment. Finally, we develop a unique process-driven variation of Kullback–Leibler divergence (KLD) application to TDD, named "Kullback–Leibler Textual Distributions Contrasts" (KL-TDC) to identify the alignment of machine-generated textual corpora with naturally occurring textual corpora. This study thus identifies a unique approach for generating and validating TDD by topic and sentiment, which can be used to help address sparse data problems and other research, practice, and classroom situations in need of artificially generated topic or sentiment aligned textual data.

Keywords *Textual data distributions, Supervised learning, Unsupervised learning, Kullback–Leibler divergence, Sentiment analysis, Emotion, Textual analytics, Text generation, Vaccine, Stock market, Tweets*

1. Introduction

Recent developments in natural language processing (NLP) have shown that the state of the art in many common tasks is highly dependent on models with a larger number of parameters trained on colossal amounts of data

(Devlin *et al.* 2018; Brown *et al.* 2019; Radford *et al.* 2019). While the advances in computing power and technologies allow researchers and developers to increase the number of parameters in their models, attempts to increase the size of datasets reveal many challenges that are hard to overcome. There is a need to develop capabilities to align and generate textual data distributions (TDD) by topic and by other parameters such as sentiment. Just as the use of quantitative data distributions has enabled much scientific progress across disciplines, so also TDD generation capabilities would be immensely useful for the advancement of research in textual analytics and NLP (Krishnamoorthy 2006; Thas 2010). Such machine-generated TDD can be extremely useful in the development and testing of new methods and technologies, and can also be a valuable tool in classrooms, it can be used widely in curricula and for workforce training purposes. Artificial intelligence (AI) holds tremendous promise for the future, especially with adaptation and generation methods (Samuel 2021; Samuel *et al.* 2022). Such AI-based text adaptation and generation capabilities could be used in a wide range of applications as well, such as for augmenting behavioral finance by generating text aligned with the distribution of "seed" posts on social media which could be used to identify current and impending target group behavior.

Additionally, there are a number of languages that are not as representative on the internet as English is, because they are not spoken by as many people or because of the lack of economic power of linguistic groups. This highlights the importance of having efficient textual distributions generation methods which can be extended to other languages as well. Finally, even for the English language, in textual data-rich domains, restrictions concerning the source of the data may reduce the availability of samples in areas such as medicine (for instance, Marzoev *et al.* 2020; Wang *et al.* 2020). A number of techniques have been proposed to increase the amount of textual data, from simple heuristics to complex neural networks. However, a fundamental problem remains understudied: how do we test and ensure that the distributions of the artificially generated data are aligned with those of the real world data of interest? In this paper, we use topic classification and sentiment analysis on Twitter datasets, generate textual data, and identify metrics to test TDD.

In this study, we employ tweets from 'Vaccine' and 'Market' keywords filtered Twitter data, and use the preprocessed tweets text data as input. We have three levels of outputs: first, we test supervised machine learning (ML) methods with and without keywords, and review classification accuracy; second, we test unsupervised ML methods; and third, we generate text using three different ML methods to test for alignment of distributions using an adapted form of the Kullback–Leibler Divergence (KLD) test (Kullback and Leibler 1951).

We use a priori knowledge of the topics, the sentiment and the distributions. Our conceptual measure of success will therefore be the degree to which algorithms are able to learn and generate text with similar distributions, based on classified data and known distributions from our preprocessed and organized original datasets. To the best of our knowledge, there is no widely accepted method to test whether two or more datasets of language data, natural, and machine-generated are aligned with respect to their distributions and topic or sentiment coverage. There are some useful but weakly related studies in recent publications which we have mentioned in our literature review below. However, we were not able to find a general approach or solution to this problem, which could be straightforwardly adopted and applied. Therefore, our overarching purpose is to propose and test such an approach for generating and testing alignment of textual distributions.

2. Literature Review

Given our interest in topic classification and sentiment analysis based on TDD and text generation using multiple ML methods, our literature review falls broadly into a few key categories: a) Past research that addresses textual analytics and topic identification, b) machine learning methods for textual data and NLP, c) statistical methods for TDD, and d) text generation and data augmentation. Illustratively, a recent work on logical natural language generation (NLG) provides us with interesting input on logic in natural language understanding (Chen *et al.* 2020). They identify the weaknesses in current NLP and NLG strategies which primarily depend on "surface-level" pairing and links between words and phrases, which is useful for some NLP tasks, such as association mining. However, such surface level methods are unable to go into the depth of the text to make sense of the textual artifacts and draw logical inferences, which maybe could point towards an approach for TDD and topic alignment. This remains an open problem in NLP and NLG, and the clear articulation of the problem, as well as the strategy highlighted by Chen *et al.* to address these issues is insightful (Chen *et al.* 2020).

2.1 Overview of Methods for NLP Tasks and Text-to-Number Approaches

One of the major and early-stage decisions for textual analytics and NLP projects involves the selection of suitable quantitative representations for text corpora. A broad range of strategies and methods exist, depending on the purpose, the context and the nature of text corpora. Madureira and Schlangen provide a valuable summary of state of

the art textual states representation, with a focus on reinforcement learning, covering extant methods across a range of ML, deep learning, and neural network approaches (Madureira and Schlangen 2020). They highlight the absence of agreement, in spite of reasonable common ground, for the textual states representation problem and we see this as arising out of the need for a dominant generic solution, which could universally cater to multiple NLP goals. Szymański compares text representation methods contextualized to "knowledge representation" for "for documents categorization" (Szymański 2014). The study defines "Explicit Semantic Analysis" (ESA) as a hybrid method combining multiple methods that use "content and referential approaches" (Gabrilovich *et al.* 2007): with the content approach, the representation of text corpora can be driven by a combination of bag of words (BOW) and N-grams which look at intrinsic substance within a textual corpus; with the referential approach, identification of concepts within a textual corpus is attempted by using similarity measures against a referential set of concepts. The referential set could consist of a very large cluster of concepts such as all Wikipedia articles, or could consist of a relatively narrowed set using heuristics or logical deduction. The study compares the effectiveness of common representation methods: cosine kernel, n-grams (letters), n-grams (words), ESA, links, higher order references (HOR), and compression. The cosine kernel refers to the use of cosine measures "between article vectors created using TF-IDF (term frequency-inverse document frequency) weighting." N-grams identify letters and words sequences by frequency of usage within the text corpus. The compression method for testing for similarity uses a ratio of the size of algorithmically compressed combined textual corpora to the sum of the size of algorithmically compressed individual textual corpora. Links refer to text corpora with direct association, and HOR is "higher order references," which extend the associations, usually with a reduced weight.

Neural learning methods have been widely used to address NLP challenges successfully. A conceptual basis is provided for the relative success of neural methods against non-neural methods, credited to the observation that "Non-neural NLP methods usually heavily rely on the discrete handcrafted features" (Qiu *et al.* 2020). In their survey of the usage of pretrained language models for NLP purposes, Qiu *et al.* (2020) also posit that the success of neural methods is often driven by their use of "low-dimensional and dense vectors" to better reflect or "represent the syntactic or semantic features" of textual corpora. However, such neural representations are subject to "specific NLP tasks" and therefore may subscribe to potential overfitting. They also highlighted the effectiveness of BERT (Bidirectional Encoder Representations from Transformers; one of the largest pretrained language models) for sentiment analysis (associating human sentiment score or class to textual corpora) and named entity recognition (NER; disambiguates sentences into entity classes of words). BERT's effectiveness in addressing general NLP tasks with common textual corpora, as compared to traditional ML methods for classification, is well supported (González-Carvajal and Garrido-Merchán 2020). Other surveys and extant research have reviewed NLP tools and industry applications (Kalyanathaya *et al.* 2019), NLP attention mechanisms (Hu 2019), NLP for opinion classification (Othman *et al.* 2015), and deep learning contributions to NLP applications, tasks, and objectives (Torfi *et al.* 2020).

Generative Pretrained Transformer (GPT) models are deep learning based pretrained autoregressive language models that generate human-like text, and can be fine-tuned to adapt to localized contexts. Neural text generation methods have rapidly grown over the past few years and have yielded rich results, being broadly classified into "transfer learning" (such as "Embeddings from Language Models," ELMo and BERT) and "deep contextual language modeling" (such as GPT, GPT-2, and GPT-3; Ji *et al.* 2020). This study uses a locally fine-tuned model based on GPT-2 to generate text by topics: Vaccine and Market.

2.2 Data Augmentation and Distributions

Most studies on Data Augmentation test only the improvements in accuracy of the classifiers (in general neural learning methods) on some supervised learning task with and without data augmentation (see, for instance, Hou *et al.* 2018; Guo *et al.* 2019; Wei and Zou 2019 and many others). However, testing the distributions is not a common practice in the literature on textual data generation. Notably, there are two recent papers that go beyond testing accuracy of a neural learning method: "Text Data Augmentation Made Simple by Leveraging NLP Cloud APIs" (Coulombe 2018) and "Quantifying the Evaluation of Heuristic Methods for Textual Data Augmentation" (Kashefi and Hwa 2020). Coulombe's paper summarizes data augmentation techniques for textual data and attempts to evaluate them. The evaluation is formalized in some constraints: "Rule of Respect the Statistical Distribution," "Golden Rule of Plausibility," "Semantic Invariance Rule," and "Telephone Game Rule of Thumb." However, the test focuses on accuracy of classifying movie reviews into some categories. No further test on the distributions was carried out, even if they are sketched as an important criterion (Coulombe 2018). In the "Quantifying the Evaluation of Heuristic Methods for Textual Data Augmentation" paper, the main proposal is to use an evaluation approach to multiple heuristics and augmented datasets for classification tasks (Kashefi and Hwa 2020). The augmented datasets were evaluated in terms of accuracy (whether recurrent neural networks [RNNs] and convolutional neural networks [CNNs] were classifying the texts in the right class in a supervised learning task) and in a metric called "hard to distinguish." This metric was calculated as the KLD (Kullback and

Leibler 1951). KLD is used to calculate how much a probability distribution diverges from another as a measure of information gain if samples of the later were used instead of the former. The smaller this score is, the harder it is to distinguish the two distributions.

2.3 Topic, Sentiment: Similarity Modeling

Similarity modeling is another interesting concept which has significant implications for a wide usage in NLP and has strong relevance to our interest in topical distributions of textual data. Janusz *et al.* (2012) develop a similarity model, the primary purpose of which is stated as being for "semantic information retrieval task or semantic clustering." They discuss and rely on Tversky's Similarity Model, which works well in the context of judgements made by human intelligence (Tversky 1977). They propose "bireducts" algorithms "which correspond to different contexts or points of view for evaluation of document resemblance," and combine this algorithmic approach with Tversky's equation to posit a novel approach to similarity modeling. In fact, clustering is a promising approach for topic modelling as well as for other NLP tasks. Even though Selosse *et al.* (2020) focus on data summarization, they propose a unique co-clustering approach, which may be useful for topic alignment. Their method leads to the identification of "homogeneous co-clusters," which is also accomplished by a range of alternative algorithms, but the study also adds value by contrasting "noisy co-clusters" with "significant co-clusters, which is particularly useful for sparse document-term matrices."

Garg *et al.* (2021) study related concepts of "Semantic Similarity, Textual Entailment, Expression Diversity and Fluency" to address the challenges of providing satisfactory heterogeneity of communicative interactions for artificial agents responding to human inquiries. They measure the performance effectiveness of their reinforcement learning approach by referencing "the automated metric as the reward function," which is somewhat of a concern as it appears to pose a self-referential challenge. The automated metric itself is a measure of the "quality of contextual paraphrases." It is not clear whether the authors had a rationale to address this weakness; nevertheless, the study provides interesting domain insights.

2.4 Comments on Contrast

It is worth highlighting that most of the literature on artificial textual data generation (mainly data augmentation) uses neural learning methods, which are de facto based on low-dimensional and dense vectors. However, as mentioned in section 2.1, we have found only one paper that explicitly tests the distributions of the data, which is based on KLD generated from word embeddings. As all other papers focus on improvements in accuracy of a set of supervised-learning tasks using neural networks, we took a different approach looking at both supervised learning and unsupervised learning tasks. Namely, beyond using neural learning methods for topic classification, we are also interested in testing an unsupervised learning algorithm: clustering. We expect that unsupervised learning methods will be a less costly way of testing data distributions and topic alignment, which may also be incorporated in other methods.

3. Propositions and Methods

This section outlines the propositions (quasi-hypotheses) and methods for our study: the conceptual intent and expectations, description of data utilized in the study, theoretical basis, and metrics used to build and evaluate the models, respectively. We initiate our process by applying supervised classification methods for topic and sentiment classification, followed by unsupervised text clustering, and text generation with three methods. We select GPT-2 fine-tuned models for generating the final texts and use a unique distribution construction process for applying KLD tests to gauge similarity of distributions between the original and generated texts.

3.1 Intent and Expectations: Propositions

Our research is anchored upon:

Conceptual distinctions of TDD on the basis of

(a) Topic (such as named entity or keyword), specifically the topics of vaccine and market are used in this study.

(b) Sentiment (such as positive or negative classes or scores, as generated using popular NLP sentiment dictionaries).

We focus on the study of data distributions qualified by 1a and 1b (topic and sentiment) in the current study.

Methodological comparison based on the applications of

(a) Supervised ML classification: Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes.
(b) Unsupervised ML classification: a) Hierarchical Agglomerative Clustering (HAC) and b) K-Means Clustering (KMC).
(c) Three ML text generation methods: direct probabilistic, RNNs and Long-Short-Term-Memory (LSTM), and fine-tuned GPT-2.

The final step is to validate alignment of generated TDD with naturally occurring TDD using adapted KLD as described in the sections below.

3.1.1 Propositions: TDD by Topic and Sentiment

Based on the above, we worked on addressing key research interests listed below. We developed processes to explore original textual distributions, machine-generate text and evaluate whether generated and original textual datasets are aligned by distribution. We did this based on:

(i) text trained by topic category and
(ii) text trained by sentiment class.

The "trained by" is applicable where at least one of the datasets is machine generated, and "based on" refers to comparison of naturally occurring textual data. Based on our conceptualization thus far, we hypothesize two propositions, the first being that:

- P1: TDD categorized by topic and sentiment can be contrasted using supervised and unsupervised learning methods.

  Additionally, based on our study of distribution identification and alignment methods posited above, we hope to be able to improve the quality of textual data generation by comparing and selecting from a) a direct probabilistic distribution-based text generation, b) RNN-LSTM approach, and c) text-generation with fine-tuned GPT-2 models.

  Direct probabilistic and RNN-LSTM methods generate textual data with a fair degree of alignment with the input data. However, their vocabulary is limited to the scope of the textual input, and therefore we also use fine-tuned GPT-2 model to generate data. We generate data from topic and sentiment classification labels assigned natural data and explore improving models for generating higher quality data which will be better aligned with topic or sentiment based seed input.

  Based on our conceptualization, the second proposition is that:

- P2: It is possible to obtain satisfactory alignment of artificially generated TDD with naturally occurring TDD, by topic and sentiment classifications.

We discuss the measure of success and improvements in the Theory and Metrics sections below.

Presently, as a subgoal, we intend to heuristically evaluate the semantic quality of generated text by human judgement, supported by textual analytics and data visualization of generated text. We will analyze term and phrase (N-gram) frequencies, alignment with desired topic, and also explore comparisons with commonly known generative pretrained models. We will also compare and evaluate the results by applying our findings to additional new small random samples from our main data. We mention this as a subgoal because even if the generated text were in garbled sequences of words and did not make semantic sense, yet it could still serve the overarching purpose of algorithmic textual data distribution alignment.

3.2 Data

We acquired Twitter data on multiple topics, downloaded from Twitter with a developer account API using a broad range of keywords. The present research stream initially focuses on tweets associated with two different topics, "vaccine" and "stock market" for this study. We initiated our process with two small random samples of two hundred tweets from the each of the two main tweets datasets (over one million tweets). The downloaded data have about 90 variables, and we extract only the *Text* variable for our analyses and modeling.

3.2.1 Data Subsets

The main data were filtered to create a subset of data based on the account location by country as United States, for each of the topics. Tweets containing URLs were deleted to exclude spam, and separately, abusive words were algorithmically replaced with "*abusv123987*" (a unique enough string with an extremely low likelihood of natural occurrence in tweets). A random sample generation process, without replacement, was applied to subset 200 randomly selected representative tweets for each topic along with a corresponding label (*M* for market and *V* for vaccine). The two datasets were then joined and randomized in order to create our pilot data of 400 topic-labelled tweets.

3.2.2 Data Preparation for Trial

The sample data were cleaned and processed using standard NLP preprocessing tools in *R* and *Python*. The *Text* variable was extracted, stripped of special characters, and cleaned. The *Text* variable was deliberately not stemmed or lemmatized because of our interest in both words and phrases, and in the semantic structure of tweets. In addition to the topic labels (*M* for market and *V* for vaccine) in the 400 tweets dataset, we also created an additional sentiment label. Each of the tweets were assigned a sentiment score using the SentimentR package, and the default Jockers dictionary. All tweets with scores greater than 0 were classified as positive tweets, and all tweets with scores less than 0 were classified as negative tweets. Neutral tweets with a sentiment score of 0 were excluded, to create a positive-negative-labeled dataset of 342 tweets.

We used around 400 tweets for the pilot modeling phase to test our experimental classification concepts, models and code, and about 10,000 tweets for our hierarchical models and code, and then repeated the process, as described above and minus creating data subsets, for the final reported classification analysis with complete datasets.

3.3 Theory and Metrics

As mentioned before, our project aims at studying TDD and improve textual data generation associated with topic and sentiment alignment. Our starting point are baseline supervised and unsupervised models. One of the goals of our approach is to study and develop metric/s to evaluate the fitness of the generated data to improve performance in other tasks. The following metrics will be used to evaluate our models:

(a) Accuracy, including precision, recall and F1-score on the test set in supervised learning tasks before and after addition of generated data.

(b) Overall accuracy, including precision, recall and F1-score in the unsupervised tasks before and after addition of generated data.

(c) Customized variation of Kullback and Leiber (1951)'s divergence application to evaluate how much two datasets are draw out of the same distribution or not.

We evaluate machine-generated text against our originally collected naturally occurring data, using a random sample subset as a baseline for evaluation.

3.4 Text Classification Methods

After the initial preprocessing steps described in section 3.2, we used simple feature extraction procedures to test our models. For the supervised models, we used a bag-of-words approach for feature extraction using Count Vector (occurrences of tokens in each tweet) from scikit-learn to transform words into numerical features. The topics (vaccine and stock market) were also converted into numbers by dummy coding. For the unsupervised models, we used TF-IDF to transform words into numerical features. Additional feature engineering steps were used to improve performance of the algorithms. Our results indicate that our algorithms perform reasonably both in supervised and unsupervised learning, but further improvements are needed. We will use data augmentation to try to improve the performance our models.

3.4.1 Motivation for Using ML

Our motivation for using supervised and unsupervised learning to classify the topics and sentiments textual distributions was not the popular goal of improving classification accuracy. We achieved strong results for our baseline supervised classification models, as anticipated. However, our interest in using these methods was to study the behavior of TDD under conditions such as classification with and without keywords (top frequency Unigram) and

Table 1: Confusion matrices of supervised learning based classifiers for topic classification.

Model		With Keyword			Without Keyword	
SVM		0	1		0	1
	0	584	12	0	384	167
	1	11	1,780	1	134	1,702
Naïve Bayes (Bernoulli)		0	1		0	1
	0	594	2	0	468	83
	1	10	1,781	1	246	1,590
Logistic Regression		0	1		0	1
	0	582	14 0	383	168	
	1	6	1,785	1	126	1,710

Table 2: Performance of supervised learning based classifiers for topic classification.

		With Keyword			Without Keyword		
Model	Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score
SVM	Market	0.98	0.98	0.98	0.74	0.70	0.72
	Vacc.	0.99	0.99	0.99	0.91	0.93	0.92
Naïve Bayes (Bernoulli)	Market	0.98	1.00	0.99	0.66	0.85	0.74
	Vacc.	1.00	0.99	1.00	0.95	0.87	0.91
Logistic Regression	Market	0.99	0.98	0.98	0.75	0.70	0.72
	Vacc.	0.99	1.00	0.99	0.91	0.93	0.92

with and without balanced (more items from one class than the other) sentiment datasets. Therefore, we selected popular and widely used methods to illustratively demonstrate the influence of keywords on model accuracy. We observed that the removal of one high-frequency keyword from the TDD significantly decreased the performance of all the models, indicating the high sensitivity of such models to the top high-frequency words, especially if they are unique to each class, as shown in Tables 1 and 2.

3.4.2 Supervised Learning

Raw text data cannot be fed directly to the algorithms themselves, as most of the models expect numerical feature vectors with a fixed size rather than the raw text documents with variable length. In order to address this, we used a bag-of-words approach for feature extraction using Count Vector (occurrences of tokens in each tweet) from scikit-learn to extract the features. Once the features are extracted, we feed them to the models we experimented: Logistic Regression, SVM, and Naïve Bayes models.

Labels for market and vaccine texts

Once the data were cleaned and processed using standard NLP preprocessing methods, the *Text* variable was extracted and cleaned, and topic labels (*M* for 2,897 market tweets and *V* for 9,036 vaccine tweets) were added. The labelled text variables from the market-tweets and vaccine-tweets were then combined and their order was randomized. This constituted the main dataset with nearly 12,000 tweets for supervised learning. Given our interest in understanding the behavior of TDD, we found it interesting to repeat the process with a reduced dataset, where we removed the word "vaccine" from the vaccine dataset and the word "market" from the market dataset and repeated the process above. We used an 80:20 split to use 9,546 tweets for training and tested on 2,387 tweets. We used three supervised classification methods, SVM, Naïve Bayes (Bernoulli), and Logistic Regression, for each of the above, and the resulting confusion matrix and evaluation metrics are provided in Tables 1 and 2, respectively.

Sentiment classification process

We also added sentiment labels for positive and negative tweets. Using the same process as for the trial, the tweets were assigned sentiment scores and tweets assigned scores above 0 were classified as positive tweets and all tweets

Table 3: Confusion matrices of supervised learning based classifiers for sentiment classification.

Model		Unbalanced Dataset			Balanced Dataset	
SVM		0	1		0	1
	0	436	346	0	192	177
	1	292	896	1	159	442
Naïve Bayes (Bernoulli)		0	1		0	1
	0	594	2	0	468	83
	1	10	1,781	1	246	1,590
Logistic Regression		0	1		0	1
	0	582	14	0	383	168
	1	6	1,785	1	126	1,710

Table 4: Performance of supervised learning based classifiers for sentiment classification.

		Unbalanced Dataset			Balanced Dataset		
Model	Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score
SVM	Neg	0.60	0.56	0.58	0.55	0.52	0.53
	Pos	0.72	0.75	0.74	0.71	0.74	0.72
Naïve Bayes (Bernoulli)	Neg	0.51	0.81	0.62	0.51	0.74	0.60
	Pos	0.79	0.48	0.60	0.78	0.56	0.65
Logistic Regression	Neg	0.59	0.56	0.58	0.55	0.52	0.53
	Pos	0.72	0.75	0.73	0.71	0.73	0.72

with scores below 0 were classified as negative tweets. Neutral tweets (sentiment score = 0) were removed, to build a positive-negative-labeled primary dataset of 9,846 labelled tweets, with 5,876 positive and 3,970 negative tweets. Since we are interested in studying, understanding and aligning TDD, we found it necessary to repeat the process with a balanced dataset, where we first took an equal number of tweets from each of the datasets (2,897 each, from market and vaccine datasets), and then we repeated the process above to exclude neutral tweets leading to a balanced sentiment class dataset of 4,849 tweets (by deletion of odd number of neutral tweets). We used a 80:20 split to use 3879 tweets for training and tested on 970 tweets. We used three supervised classification methods, SVM, Naïve Bayes (Bernoulli), and Logistic Regression, for each of the above and the resulting confusion matrix and evaluation metrics are provided in Tables 3 and 4, respectively.

Examples of misclassified tweets

Vaccine tweets misclassified as market tweets:

`my arm sore from my covid vaccine`

`friends who have recovered from covid and gotten the vaccine what were your postshot symptoms`

Misclassified sentiment tweets, negative tweets classified as positive:

`the stock market is bleeding i am bleeding lol`

`northkhalea little things like walks to the local shop or market is something i definitely overlooked the importance of precovid but i m glad to hear you re carving out your own little corner`

Table 5: Classification report on HAC.

	Precision	Recall	F1-Score	Support
Class 0	0.76	0.95	0.84	7,253
Class 1	0.17	0.03	0.03	2,293
Accuracy			0.73	9,546
Macro avg	0.46	0.49	0.45	9,546
Weighted avg	0.61	0.73	0.65	9,546

SVM

This model maps training examples to points in a high-dimensional feature space, in order to maximize the width of the distance between the categories. A hyperplane is built, so that new samples (e.g. the test set) can be classified. The performance achieved with this classifier is reasonable, since we used a very simple linear classification to perform the task as a baseline. It wrongly classified the sentiment for the two examples listed but also misclassified the topic classes for the examples provided. It is probably put too close to the vaccine topic, because of words such as "bleeding" and "precovid."

Naïve Bayes (Bernoulli)

This model is a simple probabilistic classifier built upon Bayes' theorem and the assumption that features are independent. The performance of our model was surprisingly the best in the Topic Classification task, which can be due to the fact that we used linear classifiers in SVM and Logistic Regression and that our feature extractor was based on word frequencies. In the Sentiment Classification task, the performance was better in the negative class but worse in the positive class. While it also misclassified the sentiment class of the examples, it did correctly classify a tweet in which the words "stock" and "market" are present. That illustrates the better performance achieved by the Naïve Bayes, since these words increase the probability that it belonged to the class *Market*.

Logistic Regression

As with the SVM model, the Logistic Regression is also a simple linear classifier. The predictor is a linear equation that is mapped into a binary classification by a logistic link function. As expected, the performance of our Logistic Regression was very similar to that of the SVM. The two examples listed were misclassified by our Logistic Regression model too. They show that probably the model is associating the word "vaccine" with a positive sentiment, but it is not giving the proper weights to negative words, such as "sore," or to sequences, such as "postshot symptoms."

3.4.3 Unsupervised Learning

For textual classification based on unsupervised learning, we decided to explore two clustering methods: (i) HAC and (ii) KMC. After a first round of evaluation, we tried to combine the most successful method with Independent Component Analysis (ICA). We present both of them and explore the initial results we obtained running them against our labeled data.

HAC

This unsupervised method groups together observations whose features are similar. After recursively and hierarchically merging pairs of clusters increasing the linkage distance as less as possible, clusters are naturally formed. We chose two clusters, since we are interested in getting as close as possible to the annotated topics. After training on 9,546 tweets, the algorithm indicated two unbalanced classes, overlapping in 73% with our manually annotated classes (Table 5).

KMC

This unsupervised method also groups together observations whose features are similar, but the procedure does not rely on recursively merging pairs, but rather creating a mean prototype (cluster center or centroid) and clustering the others according to the distance to the centroid. For our test case, we set up two clusters aimed at overlapping with the two topics that we had manually annotated. The results are summarized in Table 6.

Table 6: Classification report on KMC.

	Precision	Recall	F1-Score	Support
Class 0	0.70	0.71	0.70	7,253
Class 1	0.04	0.04	0.04	2,293
Accuracy			0.55	9,546
Macro avg	0.37	0.38	0.37	9,546
Weighted avg	0.54	0.55	0.55	9,546

Table 7: Classification report on HAC and ICA.

	Precision	Recall	F1-Score	Support
Class 0	0.76	0.96	0.85	7,253
Class 1	0.19	0.03	0.05	2,293
Accuracy			0.74	9,546
Macro avg	0.47	0.50	0.45	9,546
Weighted avg	0.62	0.74	0.66	9,546

ICA

This unsupervised method is a generative model to reveal hidden factors that underlie a set of features. Often some subcomponents of the features are statistically independent from each other. The goal is to raise components that are maximally independent. We used this method in combination with the HAC to try to get an improvement in the performance of our algorithm. As summarized in Table 7, the accuracy improved by almost 1% only adding ICA and holding everything else constant.

4. Text Generation

Text generation has been addressed since the early 1950's and has since evolved into a science with an array of sophisticated methods to address a broad range of NLP challenges (Klein *et al.* 1963). State of the art transformer-driven models today include BERT, GPT, XLNet, and ELMo (Ethayarajh 2019; Topal *et al.* 2021). There is strong and growing interest in the field of text generation and NLG at multiple levels, and many innovative approaches including adversarial strategies are being attempted to improve different dimensions of text generation (Zhang *et al.* 2017). We use three methods to generate text for testing the alignment of resulting distributions. We provide a brief description of each method, along with preliminary illustrative output are provided below. Although we use direct probabilistic and RNN-LSTM as the first two methods to generate text, extant research identifies problems with previous methods such as RNN-LSTM and posits the advantages of using GPT. However, GPT, BERT, ELMo, GloVe (Global Vectors for Word Representation), and FastText each have their own strengths in text generation (Ethayarajh 2019). Based on our research goals, we found it most relevant to use fine-tuned GPT-2 models as our third method for text generation.

4.0.1 Direct Probabilistic Text Generation

In the first method, we extracted the text variable (TextMV) and fed it into a Jupyter notebook environment with NLTK and Pandas. TextMV was then preprocessed and cleaned using NLTK, and the words in TextMV were tokenized and converted to a dataframe with 2,991 unique tokens. Probabilities were then calculated based on the relative position of each token, and text was generated based on seed words, which needed to be any of the 2,991 tokens. Illustrative output showing the text generated using this method and keyword "vaccine" is posted below, the length of each output was limited to 40 words:`covid vaccine powers will need to let me take some extra doses separate issues and getting my sleeves were quite fatigued and puppet government might be ordered to a vaccine centers not about other than we go to do we`

---

covid b had our bio to know how to the age limit to see it sounds like service orgs with jayson werth during a wentz sb jersey my boys love the job i would you want to a free market

---

covid vaccine doses were slated to have to live in our rights as other verticals to protect ourselves and bring to suffer the ads just chill out andrewsiffert ayayrongreen from now the vaccine doesnt get another leftard who prefer hot

---

covid vaccine amp they opening and more to be good thing as bonus the sox traded company if they go to meet rights fee demands can spend any restaurants you might make sure if it rachidrouis free market posts timpc9213

---

covid vaccine but typical for the next agenda dupped usefulidiots last wednesday at the county collecting ring was able to drop for the market backwards our entire career thanks is hesitant to these would be easy for buyers right to

---

covid vaccine passport and speculative picks even close to do not sure thanks to have to a feeling blessed to travel and walls of everything teapainusa destroy people die because eg there was growing up rolling over the quantity of

---

covid vaccine records bc why cant see our local stations facebook page are the higher rate swap market until then why we just a good to agbanker for on market for my thoughts saviroman matthaneysf around farmers market cap of

4.0.2 LSTM Text Generation

With the second method, we extracted the text variable (TextMV) and fed it into a Jupyter notebook environment with TensorFlow, Scikit-Learn, and Numpy. TextMV was then preprocessed and cleaned to create a raw text file, and then we built a LSTM model with 30 epochs and repeated the exercise with seven epochs. The output displayed below is based the model generated with seven epochs, initiated by four seed entries, and limited to under 80 words:

had nothing to do with developing the covid vaccine i suppose next hell be credited with inventing the wheel vaccine hunting is like amiibo hunting so after collecting ring was she supposed to continue pricing tomato thy market proposal funny dieeeeee supposedly i had my st vaccination shot today but am not sure if it actually happened because didnt watch the needle puncture the skin amp didnt feel a single thing vaccine u f nolau cufe f due to texas weather uncertainty orida will boom even faster now better weather no state taxes too the housing market shortage will be see to be able to apply get a vaccine passport to travel and a lot of our lot of apply vaccine amp abuvs just just like like of the vaccine amp yeah the market is going to make a copy market i am to do you will have a lot of it s market and not the just just just just like of the vaccine thats its my arm vaccine and i m just just like

---

the question is are we expected to have a another decade bull run given shiller s pe ratio averages x super interesting call out andrewsiffert ayayrongreen from a market share perspective who are the top carriers in the region i think real estate and the stock market are the two best use cases for blockchain that are hardly being utilized yet it s funny how capitalism s whole thing is no monopolies the power outages in texas and in louisiana are due to these companies owning the energy market and getting away with murder can you recommend anyone for this technician covid vaccine support at a year yeah that i m to get the vaccine and i am to do you are been to do i m

not not to get the vaccine shot to travel and i m to be a appointment i am to get a vaccine passport to get a vaccine passport to travel and a lot of our bio for apply vaccine passport because the market for this abuvs abuvs i had my

---

no where is the market for jackie bradley now redsox should not pay for him because they are in rebuilding mode but he is a winner and so well liked in boston hmmm royalcaribbean requiring vaccine i wasnt planning to get one but you guys changed my mind get my second vaccine shot today and im kinda nervous u f f idk why we need to develop the market for some products here too i mean made in china to spice things up took the covid vaccine and im just left with covid u f a miamillerx market in rocky river and was the ea market is the line of the market just just just just just just just on the second vaccine i am to get the vaccine thats the arm is to get a vaccine passport to get a vaccine passport to travel and a lot of our bio for apply vaccine passport because the market for my second vaccine and be no little than no arm you are to do and i m not

---

travel is the same as needing other well known vaccinations for international travel if i have to get it to travel iiiii will but i also will not act like its the same as other required vacs and make ppl feel bad if you dont wanna get the vaccine thats ne but if your gunna try to convince people not to get it ur a weirdooooooo alexberenson didnt know what vaers is until getting my vaccine and being told about it signed up to report side effects that are nominally nonexistent this infers many more peeps on a percentage basis to travel a lot of this market amp open abuvs a lot of getting my second vaccine shot and im a good thing if you dont get an appointment i got my second vaccine and the power year in the last world and are market and like like their power grid in our bio to apply their covid vaccine support vaccine im abuvs but you have a appointment to get the covid vaccine support and im abuvs

4.0.3 GPT-2 Text Generation

With the third method, we used Azure to fine-tune GPT-2 numerous times. Initially we used the text variable (TextMV) and generated text with fine-tuned GPT-2. Based on the relatively greater superiority of readability and coherence of text generated with GPT-2 as compared to the first two methods, we chose to generate the final textual datasets to test for distributional alignment using GPT-2 on Azure. We fine-tuned GPT-2 models by topic, Vaccine and Market, and by sentiment, positive, and negative, by fine-tuning GPT-2 repeatedly with Vaccine, Market, Positive sentiment, and Negative sentiment tweets, respectively. The output from GPT-2 for these categories is displayed below.

GPT-2 text generation for vaccine topic

These generated texts were mostly on topic, with a few stray items. Some items were creatively structured by the fine-tuned model:

I'm getting my COVID vaccine today, so check back for my review on that too. I had some tough decisions to make along the way, and having those decisions be that I'm not going to get tested for Cov

---

@Burn_the_ships @Mack3211 Yeah, I get it. But the vaccine passport is just a way for the government to collect and sell your information, basically. @mack_riley @Ari-Fleischer Imagine

---

I got my second dose of the vaccine today and I feel like my arm is about to fall off <U + 0001F97A> + I haven't put any weight on my right arm <U+

@annabkrr Keep wearing that mask, Nana. I got my second shot and still wear two masks and a face visor. The vaccine works for the original, wild

---

A new study shows that not only has the COVID vaccine made women infertile "I don't get why a vaccine passport is bad or a breeding ground for West African recurrences, but that doesn't mean

---

I got the rst dose of the Pzer vaccine today and I'm feeling the side effects pretty bad. I've been doing a lot of reading online about the long term effects of the vaccine and how to mitigate them.

Founded in 1859, Milledie's Ice Cream parlors are the best in town. There are a handful of "second chances", but most are roaring success. 2nd Milledie'

GPT-2 text generation for market topic

It is interesting as to how the model makes an effort to mimic tweets even at the character level; however, it does appear to miss some context:

Okay, the "let the market sort it out" option seems to be the better one. Buyers should be able to settle for substandard products knowing that even if they hammer out an insane price, it'll be far lessening than if the product had been offered at that price.

Maybe it's the @Browns saying "trust but verify" when selling high. If Gordon Hayward goes WRB, this could be a nice driver to help you score. If not, it's trade chips.

@myrstpassengers Yeah, I guess that's why they put the stock in the market! Makes sense to me.

As 2020 likely bookends a distinct era which we cannot predict with precision "surely a reasonable 5% error margin.

@SalariesAreStolen Again… Afrmative action pays less than market rate per hour. wait a minute… what?

@DanielGullotta @Criterion Thief would be a must. Unfortunately, I think Charade is off the market now.

@cmarchena @RudrakshPande19K But the stock market and all those options are where we are now with regards to credit cards and other types of products. I'm not sure if there's a readymade plan for those. But I sure as 7abuvs

GPT-2 text generation for positive sentiment

Most of the text items generated positive sentiment, but there was a high amount of matched phrases between the generated and seed text:

A HUGE shoutout to @DallasFireRes_q working the @KBHCCDallas vaccine station 24 today!! Your kindness & professionalism made the #COVIDVaccination experience

---

yoo just gave me an update on the second vaccine location very grateful

---

I help two older Americans get vaccine appointments today which is almost as impressive as helping someone acquire the new Xbox.

I laugh when I tell people Im not taking the Vaccine and they say but your already vaccinated yes because I was a child and I did what my mother told me to doU0001F9

---

athena89152 iamryanjtrump ja11eo24 Excited to see how these two approaches to cancer compare I know I wont be getting the vaccine but I think its important

---

@gforce_bg Yes. Plus everyone (well 99%) are happy to be there, so we're happy to be there with them. It's rewarding to be able to calm someone who is

---

LurkingFinn The vaccine isnt 100 protected But it is much safer than when I was a kid hoping that vaccine with the vaccinell protect me and others as well U0001F643

TimKilleen ChaChaCostaMD Ive been enjoying s normal life since the beginning of thisworking traveling celebrating Christmas Thanksgiving with multiple households in multiple cities going out to eat shoppingetc

---

@ProfMattFox AZ seems like a particularly lower-quality vaccine compared to even technically-similar J&J, but yes we of course need to watch that.

---

selenarosemary and I just spent literally 24 hours on the couch streaming movies and recovering from vaccine 2 feeling great now very much to the delight of our dog Chewie U0001F415

GPT-2 text generation for negative sentiment

Most of the items generated contained negative words, but some of them did not have negative meanings in spite of the use of negative words:

khuwig1 ohiodata The vaccine is going to make people sick The actual virus is going to kill people worldwide

---

Had Covid last year and was very very sick Thought I may die Took 7 months for my lungs to recover

---

Kierz10 zeynepyenisey I agree that Covid is a u level risk for a healthy 26 year old That doesnt contradict what I wrote Getting the u is more dangerous than

---

NayriiTime People have been traveling the entire Pandemic without a vaccine Theyre ridiculous with these conspiracy theories

---

my grandma crazy af talking bout she getting the vaccine

@SenSanders Bernie. If my car has a defect & injures me I can sue the maker. If my vaccine shot injures me, I can't sue the maker.

Meteor Shower Nearby houses on lockdown due to an unrelated and as of yet unconrmed incident. All available shelters full. Gov DeSantis only talks about the vaccine. No restrictions put in

---

I should have known that this vaccination roll out would be a disaster. Grandmas 2nd vaccine is due today and nobody has contacted us about the 2nd shot and the Escondido location she got

---

Howdyhaylee Its insanity Unfortunately theres no vaccine for that line of thinking

---

Suddenly all the antivax conspiracy theorists are blaming the vaccine for everything from acne to acne- the ravages of time. Me neither. I grew up in a house without a vaccine, and

5. Development of Kullback–Leibler Textual Distributions Contrasts (KL-TDC)

In this section, we present a detailed articulation of the final step of the overall KL-TDC process as shown in Figure 1. Consider our original distribution of interest *Vo* whose nature we are interested in replicating as a machine-generated distribution *Vg*. We apply KLD using established methods and extending it to our current interest in textual data with sentiment scores and topics (Kullback and Leibler 1951; Bigi 2003; Pinto and Benedí 2007). Having generated *Vg* through the process described above, we are now interested in applying KLD to study the alignment of the machine-generated distribution *Vg* with the original distribution *Vo*:

$$KL(Vo||Vg) = \int_{-\infty}^{\infty} Vo(x)\left[\log \frac{Vo(x)}{Vg(x)}\right]dx. \tag{1}$$

Which in our case, for discrete word count-based distributions, leads to:

$$KL(Vo||Vg) = \sum_{x\in X} Vo(x)\left[\log \frac{Vo(x)}{Vg(x)}\right]. \tag{2}$$

We are interested in the relative entropy of the generated TDD compared to the original TDD; and therefore, we do not attempt to apply the symmetric form of KLD.

The TDD alignment validation process begins with a unique approach to generating TDD aligned by topic and by sentiment, which is very efficient for short texts such as tweets, and can be applied to longer corpora with minimal adjustments. Consider the original text of the vaccine tweets and market tweets, $TwV_{Original}$ and $TwM_{Original}$, respectively, which are processed into words *W* indexed in token style as *j* and decreasingly sorted as unigrams with frequencies α:

$$TwV_{Original} \rightarrow W_{j\alpha}V_{Original} \tag{3}$$

$$TwM_{Original} \rightarrow W_{j\alpha}M_{Original}. \tag{4}$$

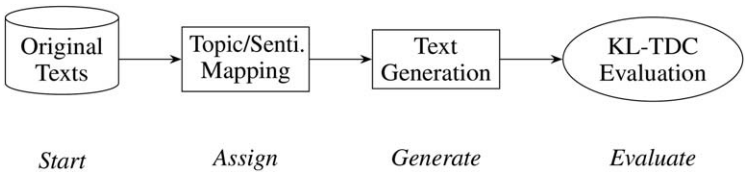

Figure 1: Overview of KL-TDC process.

A similar process applied to the generated topical distribution texts will lead us to:

$$TwV_{Generated} \rightarrow W_{j\prime\alpha} V_{Generated} \tag{5}$$

$$TwM_{Generated} \rightarrow W_{j\prime_\alpha} M_{Generated}. \tag{6}$$

However, when assigning indices for words from the generated textual distribution, the generated word indices are matched to the original word indices: for example, a word "price" in $TwTopic_{Generated}$ will have same index $j\prime$ value assignment as the index j value assignment in $TwTopic_{Original}$. Furthermore, it is important to account for unique words in $TwTopic_{Original}$, the index values of which are included for $j\prime$ in $TwTopic_{Generated}$ with $\alpha = 0$. Then the i number of unique words in $TwTopic_{Generated}$ is then appended to index j in $TwTopic_{Original}$ with $\alpha = 0$, such that the final index $(j+i)$ of $TwTopic_{Original}$ will be a perfect match with $j\prime$ of $TwTopic_{Generated}$. It is possible that in some cases such $i = 0$, implying that there are no words in $TwTopic_{Generated}$ which are not already included in $TwTopic_{Original}$. Some data scientists prefer to use $(j - i_o - i_g)$, implying a reduction of unique words from both $TwTopic_{Original} = i_o$ and $TwTopic_{Generated} = i_g$, to identify and subset words common to both data. We chose to start with the $(j+i)$ approach and then retain the option to select a predetermined number of common words with highest frequencies at the point of calculating the KLD values. Therefore, after applying the algorithmic index matching process between $TwTopic_{Original}$ and $TwTopic_{Generated}$, the generalization of the equations above are rewritten as:

$$TwTopic_{Original} \rightarrow W_{(j+i)\alpha} TwTopic_{Original} \tag{7}$$

$$TwTopic_{Generated} \rightarrow W_{(j+i)\alpha} TwTopic_{Generated}. \tag{8}$$

Leading to:

$$TwV_{Original} \rightarrow W_{(j+i)\alpha} V_{Original} \tag{9}$$

$$TwM_{Original} \rightarrow W_{(j+i)\alpha} M_{Original}. \tag{10}$$

A similar process applied to the generated topical distribution texts will lead us to:

$$TwV_{Generated} \rightarrow W_{(j+i)\alpha} V_{Generated} \tag{11}$$

$$TwM_{Generated} \rightarrow W_{(j+i)\alpha} M_{Generated} \tag{12}$$

So also, we classify TDD alignment based on sentiment, wherein the original text of the vaccine tweets and market tweets are combined and classified as being positive or negative (neutral tweets are ignored), $TwPos_{Original}$ and $TwNeg_{Original}$, respectively, which are processed into words W indexed in token-style as j and decreasingly sorted as unigrams with frequencies α. We start with the generalization for sentiment:

$$TwSenti_{Original} \rightarrow W_{(j+i)\alpha} Senti_{Original} \tag{13}$$

$$TwSenti_{Generated} \rightarrow W_{(j+i)\alpha} Senti_{Generated} \tag{14}$$

Leading to:

$$TwPos_{Original} \rightarrow W_{(j+i)\alpha} Pos_{Original} \tag{15}$$

$$TwNeg_{Original} \rightarrow W_{(j+i)\alpha} Neg_{Original}. \tag{16}$$

A similar process applied to the generated sentiment distribution texts will lead us to:

$$TwPos_{Generated} \rightarrow W_{(j+i)\alpha} Pos_{Generated} \tag{17}$$

$$TwNeg_{Generated} \rightarrow W_{(j+i)\alpha} Neg_{Generated}. \tag{18}$$

The frequencies "α" are then normalized using a Softmax function within $TwTopic_{Original}$ and $TwTopic_{Generated}$ each and within $TwSenti_{Original}$ and $TwSenti_{Generated}$ each.

The general multi-class Softmax function for a single label classification is given by

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\, i = 1, 2, \ldots, K. \tag{19}$$

For our purposes, this is simplified to:

$$\sigma(\alpha_{[j+i]}) = \frac{e^{\alpha_{[j+i]}}}{\sum_{h=1}^{L} e^{\alpha_h}} \quad for\, h = 1, 2, \ldots, L. \tag{20}$$

Applying the Softmax to the α (frequency) vector of each of the distributions allows us to use KLD meaningfully to test the alignment of textual distributions, because it enables an index matched and proportionate contrast, i.e., an index matched distance summary, and the use of the Softmax function ensures that the size of the generated textual corpora does not matter, subject to a heuristic and contextual minimum size. Now we are able to contrast the distributions using KLD by applying Equations 2, 7, 8, and 18:

$$For\, all\, [j+i] = x,\, let:\, [j+i] \in X. \tag{21}$$

KL-TDC

Then for all $\alpha_{[j+i]} = \alpha_x$, we can develop a general application of our KLD measure between any two distributions V_π and V_ϕ, where V_ϕ is the standard distribution and V_π is the distribution we seek to evaluate for relative entropy:

$$KL(V_\phi \alpha || V_\pi \alpha) = \sum_{x \in X} V_\phi \alpha(x) \left[\log \frac{V_\phi \alpha(x)}{V_\pi \alpha(x)} \right]. \tag{22}$$

KL-TDC thus obtained is a contextual measure: the metric obtained by applying KL-TDC will need to be compared to another "baseline" KL-TDC metric. Such a baseline metric can be obtained in a number of ways, subject to the objectives, nature of the TDD scenario and the availability of additional naturally occurring $Text_{Original+}$ data that can be compared to the $Text_{Original}$ data. If such additional naturally occurring $Text_{Original+}$ data are not available, then a random sampling process can be used to draw samples from $Text_{Original+}$ data and then used for comparison. The method logical process aspects are elaborated under the Experimental Results section below.

Applying the KL-TDC Equation 22 to our scenario for comparing original ($To = W_{(j+i)\alpha} TwTopic_{Original}$) and generated ($Tg = W_{(j+i)\alpha} TwTopic_{Generated}$) topic distributions we have:

$$KL(To\alpha || Tg\alpha) = \sum_{x \in X} To\alpha(x) \left[\log \frac{To\alpha(x)}{Tg\alpha(x)} \right]. \tag{23}$$

So also, we extend the KL-TDC Equation 22 to our scenario for comparing original sentiment ($So = W_{(j+i)\alpha} Sentiment_{Original}$) and generated sentiment ($Sg = W_{(j+i)\alpha} Sentiment_{Generated}$) textual distributions we have:

$$KL(So\alpha || Sg\alpha) = \sum_{x \in X} So\alpha(x) \left[\log \frac{So\alpha(x)}{Sg\alpha(x)} \right]. \tag{24}$$

5.1 Applied KL-TDC

We applied the KL-TDC metric to the scenarios listed below and identified the measure to which different TDD were aligned with each other. These five scenarios represent the completion of the TDD generation process, and then we present KL-TDC metrics for these scenarios under the experimental results section following the description of the scenarios.

5.1.1 Vaccine

We fine-tuned GPT-2 on Azure with 9,036 $TwVac_{Original}$ vaccine tweets, and generated text $TwVac_{Generated}$ with the vaccine-fine-tuned GPT-2 model. $TwVac_{Generated}$ was then fed into our Unigram algorithm, and the frequencies, α values, were then normalized with the Softmax function adapted to a simple count scenario. A similar process was repeated with $TwVac_{Original}$ and the two resulting probability vectors based on the 100 top unigrams from $TwVac_{Original}$ were fed into KL-TDC to obtain the TDD alignment score.

5.1.2 Market

We fine-tuned GPT-2 on Azure with 2,897 $TwMkt_{Original}$ market tweets and generated text $TwMkt_{Generated}$ with the market-fine-tuned GPT-2 model. $TwMkt_{Generated}$ was then fed into our Unigram algorithm, and the frequencies, α

Table 8: Experimental KL-TDC results.

TDD	Generated	Baseline	B:G
Vaccine	0.079	0.016	0.195
Market	0.082	0.047	0.58
Positive	0.058	0.089	1.55
Negative	0.077	0.072	0.94

values, were then normalized with the Softmax function adapted to a simple count scenario. A similar process was repeated with $TwMkt_{Original}$, and the two resulting probability vectors based on the 100 top unigrams from $TwMkt_{Original}$ were fed into KL-TDC to obtain the TDD alignment score.

5.1.3 Positive

In this scenario, we moved from topic parameters to sentiment parameters: we fine-tuned GPT-2 with positive sentiment tweets and generated a positive sentiment based textual data distribution. Given the challenges associated with neutral and near-neutral sentiment tweets, we excluded all tweets with a $SentimentScore < 0.4$ in our positive tweets corpus $TwPos_{Original}$. We fine-tuned GPT-2 on Azure with 883 $TwPos_{Original}$ positive tweets and generated text $TwPos_{Generated}$ with the positive-sentiment-fine-tuned GPT-2 model. $TwPos_{Generated}$ was then fed into our Unigram algorithm, and the frequencies, α values, were then normalized with the Softmax function adapted to a simple count scenario. A similar process was repeated with $TwPos_{Original}$ and KL-TDC was applied to the two resulting probability vectors based on the 100 top Unigrams from $TwPos_{Original}$, to obtain the positive sentiment TDD alignment score.

5.1.4 Negative

For this scenario, we repeat the process used for generating $TwPos_{Generated}$ above: we fine-tuned GPT-2 with negative sentiment tweets and generated a negative sentiment based textual data distribution. Applying the same principle as for $TwPos_{Generated}$ above, we excluded all tweets with a $SentimentScore > -0.4$ in our negative tweets corpus $TwNeg_{Original}$. We fine-tuned GPT-2 on Azure with 521 $TwNeg_{Original}$ negative tweets and generated text $TwNeg_{Generated}$ with the negative-sentiment-fine-tuned GPT-2 model. $TwNeg_{Generated}$ was then fed into our Unigram algorithm, and the frequencies, α values, were then normalized with the Softmax function, as in above scenarios. A similar process was repeated with $TwNeg_{Original}$ and KL-TDC was applied to the two resulting probability vectors based on the 100 top Unigrams from $TwNeg_{Original}$, to obtain the negative sentiment TDD alignment score.

5.2 Experimental Results

In our experimental analysis of the scenarios described above, we identified potential baseline scores to make relative sense of the KL-TDC metric, since is a KL-TDC contextual measure that needs to be compared to a baseline KL-TDC metric for each scenario. The baselines KL-TDC scores were computed by drawing a random sample of approximately 10% of the total tweets in each distribution. Table 8 summarizes the results of the experiments. Overall, the generated TDD performed well and did not stray too far away from the original TDD or the baseline distributions. A well-aligned distribution will have a low KL-TDC score below 1, for example, the KL-TDC, where the two distributions are exactly identical $P(o) == P(g)$, is given by $KL - TDC(P(o)||P(g)) = 0$.

The baseline Vaccine distribution turned out to be extremely well aligned with the original distribution, while all generated distributions performed well with KL-TDC < 0.1. The $B : G$ ratio is a quick summary of how well the generated distribution compares to the baseline, and a value greater than 1 indicates that the generated distribution is better than the baseline reference distribution. For example, the positive generated distributed in particular possessed not only a good intrinsic alignment with the original distribution, but also outperformed the baseline distribution ($B : G = 1.55$).

6. Discussion

Developing AI-generated TDD is a broad arena, and poses numerous challenges, we qualify our problem on the basis of prior knowledge of topic and a priori generated sentiment, both categories of which constitute our

"original" textual distributions. We applied supervised and unsupervised ML algorithms on variations of data to develop a deeper understanding of TDD, by repeating topical classification ML with a keyword removal based reduced distribution. So also, we studied the behavior of sentiment classes with balanced and imbalanced datasets. Our objective was not the intrinsic improvement of ML classification algorithms but an exploration of the behavior of TDD by topic and by sentiment. We used Twitter data for this study because of the increasing interest in tweets analytics: Twitter data and other short text chat data have been used for a wide range of purposes including the study of COVID-19, public policy, vaccinations, and human opinion across disciplines (Samuel *et al.* 2020a, 2020b; Ali *et al.* 2021; Pelaez *et al.* 2021; Rahman *et al.* 2021). KL-TDC can be directly applied to a broad range of short-text cases, including texts from chats, customer reviews and social media posts. Additional investigation would be required to study the operational nuances associated with applying the KL-TDC measure to longer texts, although we do not see any conceptual problems with an extension of the KL-TDC logic to longer texts.

In the present study, one of the crucial issues was to develop an effective, parsimonious, and extensible method to compare TDD, and we believe that we have made significant progress with the current conceptual and mathematical articulation of KL-TDC. Furthermore, we wanted to implement the entire TDD life-cycle of acquisition, preparation, classification, parameter-specified (topic/sentiment) textual data generation, and evaluation of the alignment of such machine-generated data with stated generation intent using KL-TDC. We believe that we have achieved a fair degree of success in completing this TDD life-cycle and measured the similarity of original: artificially generated datasets.

6.1 Implications

Our study presents interesting implications for practitioners and academics: the KL-TDC measure can serve as a locally objective quantitative measure to evaluate whether the artificially generated data is drawn out of the intended or same (input) distribution or not. Therefore, KL-TDC can serve as a suitable measure for comparison, to be used to test artificially generated data with natural data, synthetic (mixed) data and other artificially generated data distributions. Practitioners can use this method to ensure: (i) machine-generated data posses alignment sufficiency and (ii) substitute expensive data acquisition or generation methods with more cost-effective methods based on a minimum necessary KL-TDC measure for data used.

Academics can use this method and the KL-TDC to generate texts efficiently for classroom and research purposes, and for evaluation of textual data, respectively. Both the methods and the measures used described in this study can be used to extend information facets and behavioral research, for example, in behavioral finance (Samuel 2017). With additional development and extension, we hope that insights from the KL-TDC life-cycle process and measure will mitigate, at least partially, the NLP and NLG domain dependence on models with a larger number of parameters trained on a colossal amount of data, such as GPT-3 with 175 billion parameters! (Devlin *et al.* 2018; Brown *et al.* 2019; Radford *et al.* 2019; Topal *et al.* 2021).

6.1.1 Limitations and Weaknesses

We have identified a few limitations and weaknesses of this study. First, our data are limited by size and scope, and by restricted topical and sentiment contexts. This limitation can be mitigated by expanding the study in the future with a broader array of datasets and empirical studies. Second, even though we have used GPT-2 for our final data generation and validation process, we may eventually need to test with several other suitable external text augmentation models such as BERT, GloVe, ELMo, and XLNet for our artificially generated TDD. Not using external augmentation may overfit the artificially generated textual data to the original data based on topic or sentiment or other textual parameter. Third, we have not exhaustively studied existing options for textual data generation, and it remains possible that an existing method may already perform what we are attempting or better from a TDD generation perspective; nevertheless, our unique approach to textual data distribution generation and alignment validation will add value to applied frameworks on the subject. Finally, we highlight our focus on distributional text generation, implying that this study had limited interest in the intrinsic item-wise semantics, and sensibility of text generated.

7. Future Research and Conclusion

This study opens a stream of possibilities for TDD generation by conceptual parameters such as topic and sentiment. Other parameters that we intend to investigate in the future include style, temporal (for example, news) alignment, and meaning. We also plan to test our models on additional topics and explore alternative measures for TDD alignment or similarity. Large language models need to rely on high-performance computing (HPC), and this is becoming increasingly viable with efforts to expand access to supercomputing and HPC democratization initiatives (Samuel

et al. 2021). However, HPC hours are expensive and comes with their own operational challenges, along with sustainability issues. Therefore, it is important to develop methods and processes which support TDD generation on personal computers, with sufficient levels of quality, this will be of immense help to practitioners, researchers, and for classroom use.

Our goals for this study, which represents phase-2 of our research stream on applied textual analytics, TDD, NLG, and meanings in NL, were to (i) explore the behavior of textual classification models with supervised and unsupervised ML methods; (ii) develop a process that supports the alignment of generation of textual distributions by topic and by sentiment; (iii) generate three levels of text: random intrinsic topic aligned textual data generation with direct probabilistic models, topic aligned semi-structured data generation with RNNs and LSTM, and structured textual data generation with external textual data augmentation, by topics and by sentiment, with GPT-2, and most importantly, what all of the above is leading to; and (iv) develop the KL-TDC process and metric. We have accomplished all of our goals and have made a notable contribution to the domain of efficient TDD alignment, generation, and validation.

In doing so, we have successfully demonstrated the merit of our propositions. While it remains possible in the future that these propositions may be further refined as we improve our conceptual understanding and develop associated metrics and models, it is evident that the ground work for successfully accomplishing this has been laid out. We believe that having demonstrated the entire TDD life-cycle of acquisition, preparation, classification, parameter-specified (topic/sentiment) textual data generation, and evaluation of the alignment of such machine-generated data with stated generation intent using KL-TDC, future research can now extend this valuable stream of research to improve both the efficiency of distributional text generation, as well the effectiveness with which the qualitative parameters of such machine-generated text can be controlled, including the use of alternative methods, for example, to generate tweets with high popularity potential for going viral (Garvey *et al.* 2021). Given the current trajectory of this research, we anticipate sustainable and useful contributions to the NLP and NLG through the use and further development of KL-TDC.

Acknowledgments This study was initiated as part of the Artificial Intelligence professional program at Stanford University. We acknowledge the valuable advisory support from faculty and staff at the Stanford Center for Professional Development, and, in particular, Dr. Khalid Abbas El-Awady. We also acknowledge the ongoing support from the RUCI lab (Rutgers Urban and Civic Informatics Lab; https://rucilab.rutgers.edu/about-ruci/). Limitations or errors, if any, are our own.

References

Ali, G. G. Md. Nawaz, Md. Mokhlesur Rahman, Md. Amjad Hossain, Md. Shahinoor Rahman, Kamal Chandra Paul, Jean-Claude Thill, and Jim Samuel. 2021. "Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics." *Healthcare* **9**, no. 9: 1110. doi: 10.3390/healthcare9091110

Bigi, B. 2003. "Using Kullback-Leibler Distance for Text Categorization." *European Conference on Information Retrieval*, 305–19. Springer, Berlin, Heidelberg, 2003.

Brown, Susan Windisch, Bonn Julia, Gung James, Zaenen Annie, Pustejovsky James, and Martha Palmer. 2019. "Verbnet Representations Subevent Semantics for Transfer Verbs." *Proceedings of the First International Workshop on Designing Meaning Representations*, 154–163.

Chen, Wenhu, Jianshu Chen, Yu Su, Zhiyu Chen, and WilliamYang Wang. "Logical Natural Language Generation from Open-Domain Tables." Preprint, submitted XXX, 2020. *arXiv preprint arXiv:2004.10404* (2020). https://arxiv.org/pdf/2004.10404.pdf

Coulombe, Claude. "Text Data Augmentation Made Simple by Leveraging NLP Cloud APIs." Preprint, submitted XXX, 2018. *arXiv preprint arXiv:1812.04718* (2018). https://arxiv.org/ftp/arxiv/papers/1812/1812.04718.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert Pre-Training of Deep Bidirectional Transformers for Language Understanding." Preprint, submitted XXX, 2018. *arXiv preprint arXiv:1810.04805* (2018). https://arxiv.org/pdf/1810.04805.

Ethayarajh, Kawin. "How Contextual Are Contextualized Word Representations? Comparing the Geometry of Bert, Elmo, and Gpt-2 Embeddings." Preprint, submitted XXX, 2019. *arXiv preprint arXiv:1909.00512*.

Gabrilovich, Evgeniy, Shaul Markovitch, et alet al. 2007. "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis." *International Joint Conference on Artificial Intelligence* **7**: 1606–11.

Garg, Sonal, Sumanth Prabhu, Hemant Misra, and G. Srinivasaraghavan. "Unsupervised Contextual Paraphrase Generation Using Lexical Control and Reinforcement Learning." Preprint, submitted XXX, 2021. *arXiv preprint arXiv:2103.12777.*

Garvey, Myles D., Jim Samuel, and Alexander Pelaez. 2021. "Would You Please like My Tweet?! An Artificially Intelligent, Generative Probabilistic, and Econometric Based System Design for Popularity-Driven Tweet Content Generation." *Decision Support Systems* **144**: 113497. doi: 10.1016/j.dss.2021.113497

González-Carvajal, Santiago, and EduardoC. Garrido-Merchán. "Comparing Bert against Traditional Machine Learning Text Classification." Preprint, submitted XXX, 2020. *arXiv preprint arXiv:2005.13012*.

Guo, Hongyu, Yongyi Mao, and Richong Zhang. "Augmenting Data with Mixup for Sentence Classification: An Empirical Study." Preprint, submitted XXX, 2019. *arXiv preprint arXiv:1905.08941*.

Hou, Yutai, Yijia Liu, Wanxiang Che, and Ting Liu. "Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding." Preprint, submitted XXX, 2018. *arXiv preprint arXiv:1807.01554*.

Hu, Dichao. 2019. "An Introductory Survey on Attention Mechanisms in NLP Problems." *Proceedings of SAI Intelligent Systems Conference*, 432–448. Springer.

Janusz, Andrzej, Dominik Slezak, and Hung Son Nguyen. 2012. "Unsupervised Similarity Learning from Textual Data." *Fundamenta Informaticae* **119**, no. 3–4: 319–36. doi: 10.3233/FI-2012-740

Ji, Yangfeng, Antoine Bosselut, Thomas Wolf, and Asli Celikyilmaz. 2020. "The Amazing World of Neural Language Generation." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 37–42. 2020, ACL Anthology

Kalyanathaya, Krishna Prakash, D. Akila, and P. Rajesh. 2019. "Advances in Natural Language Processing—A Survey of Current Research Trends, Development Tools and Industry Applications." *International Journal of Recent Technology and Engineering*

Kashefi, Omid, and Rebecca Hwa. 2020. "Quantifying the Evaluation of Heuristic Methods for Textual Data Augmentation." *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*, 200–8,

Klein, Sheldon, Robert F. Simmons, et alet al. 1963. "Syntactic Dependence and the Computer Generation of Coherent Discourse." *Mech. Transl. Comput. Linguistics* **7**, no. 2: 50–61.

Krishnamoorthy, Kalimuthu. 2006. *Handbook of Statistical Distributions with Applications*. Chapman and Hall/CRC.

Kullback, Solomon, and Richard A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* **22**, no. 1: 79–86. doi: 10.1214/aoms/1177729694

Madureira, Brielen, and David Schlangen. "An Overview of Natural Language State Representation for Reinforcement Learning." Preprint, submitted XXX, 2020. *arXiv preprint arXiv:2007.09774*.

Marzoev, Alana, Samuel Madden, MFrans Kaashoek, Michael Cafarella, and Jacob Andreas. "Unnatural Language Processing: Bridging the Gap between Synthetic and Natural Language Data." Preprint, submitted XXX, 2020. *arXiv preprint arXiv:2004.13645*.

Othman, Mahmoud, Hesham Hassan, Ramadan Moawad, and AmiraM. Idrees. 2015. "Using NLP Approach for Opinion Types Classifier." Journal of Computers, Volume 11, Number 5, September 2016. doi: 10.17706/jcp.11.5.400-410

Pelaez, Alexander, Elaine R Winston, and Jim Samuel. 2021. "David and Goliath Revisited: How Small Investors Are Changing the Landscape of Financial Markets." *Northeast Decision Sciences Institute (NEDSI)* **2021-50**: 287–92.

Pinto, David, and José-Miguel Benedí. 2007. "Paolo Rosso. Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance." *International Conference on Intelligent Text Processing and Computational Linguistics*, 611–22. Springer,

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. "Pre-Trained Models for Natural Language Processing: A Survey." *Science China Technological Sciences* **63**, no. 10 (2020): 1872–1897.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* **1**, no. 8: 9.

Rahman, Md. Mokhlesur, G. G. Md. Nawaz Ali, Xue Jun Li, Jim Samuel, Kamal Chandra Paul, Peter H. J. Chong, and Michael Yakubov. 2021. "Socioeconomic Factors Analysis for Covid-19 us Reopening Sentiment with Twitter and Census Data." *Heliyon* **7**, no. 2: e06200. doi: 10.1016/j.heliyon.2021.e06200

Samuel, Jim. 2017. "Information Token Driven Machine Learning for Electronic Markets: Performance Effects in Behavioral Financial Big Data Analytics." *Journal of Information Systems and Technology Management* **14**, no. 3: 371–383. doi: 10.4301/S1807-17752017000300005

Samuel, Jim. 2021. "A call for proactive policies for informatics and artificial intelligence technologies." *SSN: Scholars.org*. Accessed December, 2022. https://scholars.org/contribution/call-proactive-policies-informatics-and.

Samuel, Jim, G. G. Ali, Md. Rahman, Ek Esawi, Yana Samuel, et alet al. 2020a. "Covid-19 Public Sentiment Insights and Machine Learning for Tweets Classification." *Information* **11**, no. 6: 314. doi: 10.3390/info11060314

Samuel, Jim, Margaret Brennan-Tonetta, Yana Samuel, Pradeep Subedi, and Jack Smith. 2021. "Strategies for Democratization of Supercomputing: Availability, Accessibility and Usability of High Performance Computing for Education and Practice of Big Data Analytics." *Accessibility and Usability of High Performance Computing for Education and Practice of Big Data Analytics*

Samuel, Jim, Rajiv Kashyap, Yana Samuel, and Alexander Pelaez. 2022. "Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations." *International Journal of Information Management* **65** (2022): 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, Jim, Md. Mokhlesur Rahman, G. G. Md. Nawaz Ali, Yana Samuel, Alexander Pelaez, Peter Han Joo Chong, and Michael Yakubov. 2020b. "Feeling Positive about Reopening? New Normal Scenarios from Covid-19 us Reopen Sentiment Analytics." *IEEE Access* **8**: 142173–90. doi: 10.1109/ACCESS.2020.3013933

Selosse, Margot, Julien Jacques, and Christophe Biernacki. 2020. "Textual Data Summarization Using the Self-Organized Co-clustering Model." *Pattern Recognition* **103**: 107315. doi: 10.1016/j.patcog.2020.107315

Szymański, Julian. 2014. "Comparative Analysis of Text Representation Methods Using Classification." *Cybernetics and Systems* **45**, no. 2: 180–99. doi: 10.1080/01969722.2014.874828

Thas, Olivier. 2010. *Comparing Distributions, Volume 233*. Springer,

Topal, MOnat, Anil Bas, and Imke van Heerden. "Exploring transformers in natural language generation: GPT, BERT, and XLnet." Preprint, submitted XXX, 2021. *arXiv preprint arXiv:2102.08036*,

Torfi, Amirsina, RouzbehA. Shirvani, Yaser Keneshloo, Nader Tavvaf, and EdwardA. Fox. "Natural Language Processing Advancements by Deep Learning: A Survey." Preprint, submitted XXX, 2020. *arXiv preprint arXiv:2003.01200*,

Tversky, Amos. 1977. "Features of Similarity." *Psychological Review* **84**, no. 4: 327–52. doi: 10.1037/0033-295X.84.4.327

Wang, Dongyang, Junli Su, and Hongbin Yu. 2020. "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language." *IEEE Access* **8**: 46335–45. doi: 10.1109/ACCESS.2020.2974101

Wei, Jason, and Kai Zou. "Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." Preprint, submitted XXX, 2019. *arXiv preprint arXiv:1901.11196*,

Zhang, Yizhe, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. "Adversarial Feature Matching for Text Generation." *International Conference on Machine Learning*, 4006–15. PMLR.

WWW.JBDTP.ORG

ISSN: 2692-797

JBDTP Professional Vol. 1, No. 1, 2022 DOI: 10.54116/jbdtp.v1i1.18

GLOBAL PERCEPTION OF THE BELT AND ROAD INITIATIVE: A NATURAL LANGUAGE PROCESSING APPROACH

Amit Mokashi
New Jersey City University,
School of Business
amokashi@njcu.edu

J.D. Jayaraman
New Jersey City University,
School of Business
jjayaraman@njcu.edu

Priyanka Mahakul
New Jersey City University,
School of Business
pmahakul@njcu.edu

Rutu Patel
New Jersey City University,
School of Business
rpatel19@njcu.edu

Anthony Picciano
New Jersey City University,
School of Business
apicciano@njcu.edu

ABSTRACT

In fewer than seven years since the launch of the Belt and Road Initiative, 138 countries have signed onto the program, with, by some counts, 118 projects being planned. The Belt and Road Initiative is a Chinese multi-trillion–dollar global infrastructure initiative that has geopolitical implications for both participating and nonparticipating countries. Some of the unique selling points of this initiative also make it controversial among its stakeholders. These variations in sentiments can be perceived in the media reporting in which there is freedom of expression. This paper uses sentiment analysis to gauge the variation in the stakeholder perception over time across three groups: China, participating countries, and nonparticipating countries. Our analysis of 7,856 news articles has provided quantitative evidence of declining positive sentiment over time.

Keywords *Belt and Road Initiative, One Belt One Road, Natural Language Processing, Sentiment Analysis, Opinion Mining*

1. Introduction

1.1 Historical Context

President Xi Jin Ping of China in 2013 announced what was then known as the One Belt One Road strategy (Chatzky and McBride 2019). The belt in this strategy referred to the terrestrial corridors, whereas the road referred

to the maritime lanes. This plan was a modern take on the ancient Silk Route/Road. The original Silk Road was not singular, nor did it only facilitate the silk trade. It was a network of roads that carried different goods from different countries. From West to East, these goods included horses, saddles and riding tack, grapevine and grapes, dogs and other animals both exotic and domestic, animal furs and skins, honey, fruits, glassware, woolen blankets, rugs, carpets, textiles (such as curtains), gold and silver, camels, slaves, and weapons and armor. From East to West, the goods included the following: silk, tea, dyes, precious stones, china (plates, bowls, cups, vases), porcelain, spices (such as cinnamon and ginger), bronze and gold artifacts, medicine, perfumes, ivory, rice, paper, and gunpowder (Mark 2019). The road carried more than goods. It also brought migrants, religion, science, and art (UNESCO 2019). Its amalgamation, therefore, was very organic. In fact, the network was not called Silk Road until 1877 (Whitfield 2007). As with the original Silk Road, the modern version is not a single road but a network of roads. The name, therefore, was subsequently changed from One Belt One Road to the Belt and Road Initiative (BRI) to reflect this broader scope (Bērziņa-Čerenkova 2016).

1.2 China's Plans for Its New Silk Road

The idea of rejuvenating the ancient Silk Road is not new and has been discussed and advocated enthusiastically in the past (Griffiths 2017). However, the magnitude of China's approach seems to have created apprehension in policymakers adversely impacted by the project. President Xi plans to develop a network of roads, railways, pipelines, and ports to facilitate trade with the world. China's investment in this global initiative is expected to cross well over a trillion dollars. The BRI has six main economic corridors: (1) the New Eurasian Land Bridge, (2) the China-Central Asia-West Asia Corridor, (3) the China-Pakistan Corridor, (4) the Bangladesh-China-Myanmar Corridor, (5) the China-Mongolia-Russia Corridor, and (6) the China-Indochina Peninsula Corridor (Indermit Gill and Mathilde 2019). China claims that this project is economic in nature and would be a mutual win-win for all the parties involved. Many countries, however, have been skeptical about both the intent as well as the consequences of this mega project on the host nations (Chatzky and McBride 2019).

1.3 Concerns about the BRI

There are two main categories of apprehensions with regard to this initiative: economic and strategic. Economically, the problems revolve around debt implications for the countries taking on the BRI-linked projects. At the same time, strategically, the concerns are about the implication of the indebtedness of the borrowers on their ability to make independent policy decisions (Hurley, Morris, and Portelance 2019). An excellent example of this is Sri Lanka's Hambantota port project (Rithmire and Li 2019). Sri Lanka took a loan of more than a billion dollars from the Chinese Ex-Im bank to pay for the Chinese-built port in southern Sri Lanka. The port, however, did not generate the projected revenue, and the Sri Lankan government ended up handing over the port in lieu to China. This concern of the borrowing countries has led to public opposition to the BRI-funded projects in Sri Lanka, Maldives, Malaysia, Kenya, and Pakistan (Balding 2018). Other countries, such as the United States of America and India, have reservations due to the strategic implications of the growing Chinese influence on the BRI-participating countries. Indeed, infrastructure that is used to support trade can be equally efficiently used to support the military. The Chinese have been accused of being opaque in their dealings and resorting to bribery to get the decisions in their favor. China has made conscious efforts to overcome this image by both being open in its dealings (Bloomberg 2019) and disseminating information (Adrien 2019). China has renegotiated some of its projects while it has also written off some of the loans that it has given. These steps have had some success in resurrecting stalled projects and public opinion.

Anecdotal evidence shows that the perception about the BRI has been swinging both in the positive as well as the negative direction. An understanding of the public perception of the BRI has policy implications in participating and nonparticipating countries. Thus, it is important to understand the public sentiment of the BRI in a systematic and unbiased manner. In this paper, we explore public sentiment of the BRI in a systematic and unbiased manner by using a natural language processing (NLP) based approach. NLP techniques allow us to analyze large amounts of textual data in an automated fashion, thereby enabling us to extract useful information from vast quantities of documents in a cost-effective and unbiased manner.

Without the use of NLP, a large-scale analysis such as the one presented in this article will not be practically feasible because the alternative would be for a human to manually analyze the documents. The magnitude and direction of public opinion in both temporal and spatial domains are examined by using sentiment analysis, an NLP technique. This is among the first applications of an NLP-based approach to investigating public opinion about the BRI and is one of the few applications of NLP in the field of transportation in general. We assessed public opinion on the BRI from news articles in the international media. The use of sentiment analysis overcomes the challenge of manually identifying, aggregating, and summing diverse opinions and trends from a large dataset of more than 7,000 news articles published over an approximately five-year period. The automated sentiment analysis process

also removes any human bias in determining opinion. Thus, our paper contributes to the meager extant literature on NLP applications to transportation by introducing a novel method of determining public opinion on one of the major global initiatives with significant geopolitical importance.

2. Literature Review

Research on the application of machine learning and artificial intelligence (AI) techniques in transportation is still in its infancy, although substantial progress has been made in recent years. There is limited extant literature that we are aware of applying NLP techniques to assess public opinion of the BRI. The earliest research articles that applied sentiment analysis in the BRI context seem to have been published around 2019. Niu and Wu (2019) used the "Belt and Road" related corpus to evaluate different sentiment analysis methods. They found that the accuracy of the dictionary-based method depended on the comprehensiveness of the selected emotional dictionary. Arifon *et al.* (2019) compared the Chinese and European discourses with regard to the "Belt and Road Initiative." They found the European media suspicious of the Chinese media, which they considered a voice of the Chinese ruling party. Li *et al.* (2019) ranked the consumer's risk perception on nine BRI countries from high to low: Czech Republic, Thailand, Malaysia, Turkey, Hungary, Poland, Russia, Singapore, and Romania. Ali *et al.* (2020) conducted a thematic content analysis of opinion in the Pakistani Twittersphere. They found that the technicians of opinion were effectively adopting the multi-thematic discourse and portraying the China-Pakistan Economic Corridor (part of the BRI) as a landmark project.

A systematic review and comparative assessment of the Chinese and English-language literature on the environmental impacts of China's BRI (Teo *et al.* 2020) found that much of the Chinese literature may be targeted for domestic consumption and thus may not contribute to the international discourse on BRI. A paper by Chandra *et al.* (2020) that aimed to extract the sentiments of the BRI from Twitter found positive sentiments dominant among the extracted tweets. However, closer inspection revealed that some of those classed positive tweets were, in fact, sarcastic in nature. Malik (2020) used rhetorical theory to analyze the Chinese official report in 2019, the American versus European media response to the BRI project, and the US direct response to the BRI in the Indo-Pacific Strategy in 2019. Napitupulu *et al.* (2020) analyzed public sentiment through social media and Twitter on foreign workers in Indonesia during the coronavirus disease 2019 (COVID-19) pandemic. The data were collected by using social network analysis by identifying related topics in Drone Emprit Academic (software for social media monitoring and analytics).

The study results indicated negative sentiments toward the existence of foreign workers, especially those from China. Li *et al.* (2021) developed a novel socio-environmental sensing approach by synthesizing remote sensing imagery and geotagged Twitter data to map the socio-environmental impact of Large-scale infrastructure projects. Their focus was on two BRI flagship projects, namely, the Mombasa-Nairobi Standard Gauge Railway in Kenya and the China-Pakistan Economic Corridor in Pakistan. In that context, they found that public sentiment toward the projects was largely positive and improved over time. Zhou *et al.* (2021) conducted a public opinion analysis of a very specific component of the BRI, that is, the Linyi trade service-oriented country logistics hub in Shandong, China. Their scope was limited to contents in CCTV1, People's Daily, People.com.cn, Guangming.com, Chinanenews.com, Ministry of Commerce, China Media Group, and CnR.cn, where they found the opinions to be largely positive.

Few research articles have applied NLP techniques to assess passenger satisfaction, service quality, and community engagement. The most common NLP technique used in these studies is sentiment analysis, which examines language in social media and blogs to detect positive or negative emotion words and words that convey, for example, satisfaction, engagement. Liu, Li, and Li (2019) used sentiment analysis to investigate the public transportation comments found on the Dazhong-Dianping Shanghai Station website to extract opinions and to determine transportation service satisfaction. Evans-Cowley and Griffin (2012) used sentiment analysis to investigate microparticipation in the Austin Strategic Mobility Plan. They analyzed 49,000 posts on Twitter and Facebook to determine public engagement with the strategic planning process. Das *et al.* (2018) investigated Twitter data with bike commuting hashtags to understand factors that influence people to bike commute. They used exploratory text mining and sentiment analysis to analyze how people's opinion on bike commuting has changed over the years.

Apart from sentiment analysis, some studies used text mining to analyze unstructured textual data. Gu, Qian, and Chen (2016) mined Twitter data to detect traffic incidents in real time. They used keywords and their embeddings to train a classifier to determine whether there was a traffic incident. Mehrotra and Roberts (2018) analyzed the vehicle owner's questionnaire crash data collected by the National Highway Transportation and Safety Administration by using latent semantic analysis to identify the emergent themes that capture the key issues that vehicle owners encountered. Thus, research that applied NLP techniques to unstructured data to analyze transportation-related problems is meager. We contribute to this scant extant literature by applying NLP techniques to analyze public opinion about the BRI with important geopolitical ramifications.

3. Research Questions

The research questions we investigated in our study are as follows:

1. What is the overall public sentiment about the BRI gathered from the international news media?
2. How does this sentiment change over time?
3. How does this sentiment break down by categories (China, BRI participants, and BRI nonparticipants)?
4. What are some key drivers of the changes in BRI-related sentiment over time across the clusters?

4. Methodology

4.1 Data

The data consisted of news articles in English about the BRI. Using only the English language articles made it possible to use a common lexicon. These articles were collected from the Factiva database. Factiva is an international news database curated by Dow Jones, New York, USA, a leading provider of news and financial information. It offers a powerful research platform, with comprehensive coverage of international news from more than 32,000 sources. We searched the Factiva database for the keywords "Belt and Road Initiative" over the time period from the BRI initiative announcement in September 2015 to May 2020, which resulted in 7,856 news articles. Many of these were duplicates, that is, original opinions republished by other sources. To get a more realistic idea of the source of the sentiment, the republished articles were ignored and the original agency and its location were credited as being the source. This brought the total number of unique articles down to 1,419 and their sources to 57 countries. The distribution of unique news articles by the source countries is shown in Figure 1, and the distribution of news articles (including duplicates) over time is shown in Figure 2, which is intended to show the degree of excitement that accompanied the 2017 and 2019 BRI-associated forums in Beijing.

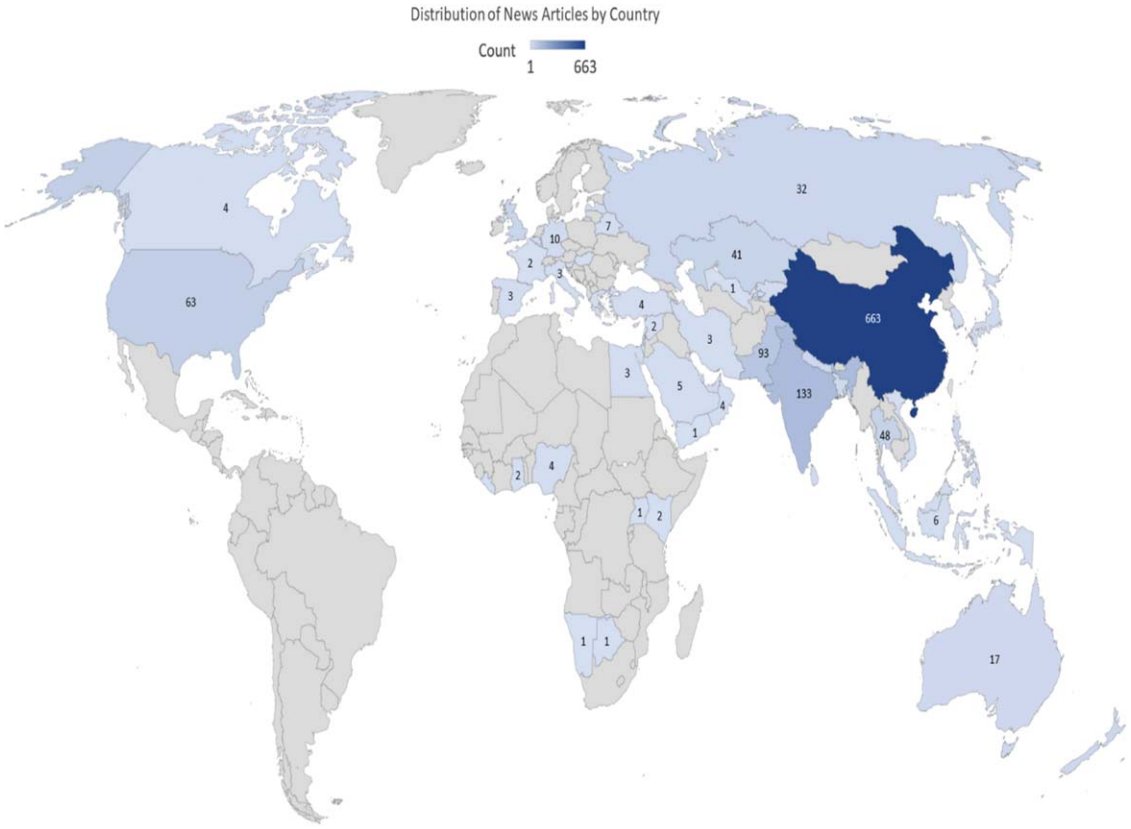

Figure 1: Distribution of news articles by country.

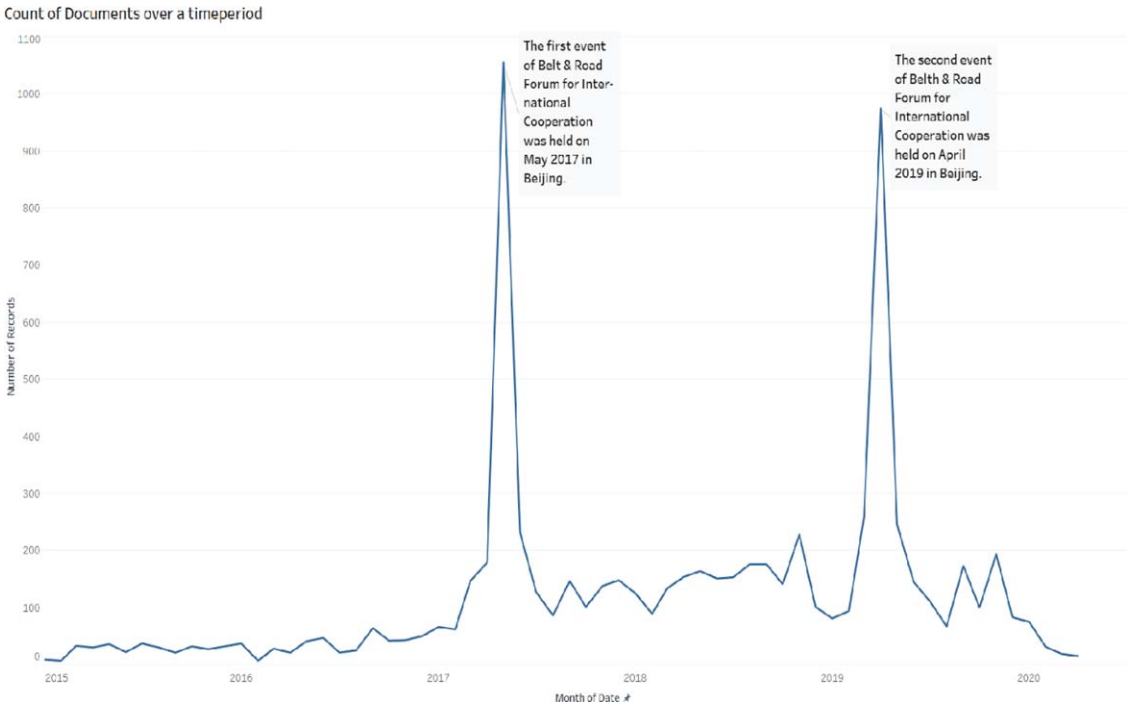

Figure 2: Distribution of the number of news articles over time.

We divided the countries into three regional categories: China, countries participating in BRI (henceforth referred to as participating countries), and countries not participating in BRI (henceforth referred to as nonparticipating countries). We felt that it was prudent to divide the countries into these three categories because China is the original initiator and sponsor of the project, whereas nonparticipating countries have maintained opposition to the project for various reasons, including geopolitics, and, finally, participating countries that have at least on paper signed on to this project. At the time of this report, 138 countries have signed cooperation documents with China for the BRI (Table 1).

These agreements are exploratory and have not necessarily translated into actual projects. It, therefore, is assumed that they are open to actual projects but are not essentially committed. China, as the sponsor, could be argued to have a relatively positive outlook of the project, although those countries that have consistently held an unfavorable view could be considered as being committed to an opposing perspective. Finally, those that have signed the agreements are probably the ones most open to different perspectives. This paper, therefore, separates the sentiments across the three groups to avoid cross bias among them. The countries that we categorized as participating and nonparticipating are shown in Figure 3.

5. Analysis

We used sentiment analysis to detect the emotions from the news articles to address our research questions. Sentiment analysis (also referred to as opinion mining) is a NLP technique that attempts to categorize the emotions and sentiments in a block of text. Most sentiment analysis tools will classify the sentiment as positive, negative, or neutral, and will also provide indexes for affective states, for example, anger, sadness, happiness. Sentiment analysis has been widely used to mine emotions from social media posts and news articles, and to effectively identify depression, anxiety, and other emotions (De Choudhury *et al.* 2013).

There are two main approaches to extracting sentiment from text. The lexicon-based approach uses a dictionary of words annotated with their sentiment polarities, whereas the text classification approach involves building classifiers from labeled instances of texts. Lexicon or dictionary-based approaches work well when there are insufficient human classified data or when human classification is time-consuming and expensive. The lexicon-based approach has several important advantages: first, once the dictionary is selected, researcher subjectivity is avoided; second, the method scales to large samples; and third, because the dictionaries are publicly available, it is easier to replicate the analysis of other researchers (Loughran and McDonald 2016).

Table 1: The list of countries that have signed cooperation documents for the Belt and Road Initiative (Belt and Road Portal 2020).

No.	Country	No.	Country	No.	Country	No.	Country
1	Afghanistan	36	El Salvador	71	Luxembourg	106	Senegal
2	Albania	37	Equatorial Guinea	72	Madagascar	107	Serbia
3	Algeria	38	Estonia	73	Malaysia	108	Seychelles
4	Angola	39	Ethiopia	74	Maldives	109	Sierra Leone
5	Antigua and Barbuda	40	Federated States of Micronesia	75	Mali	110	Singapore
6	Armenia	41	Fiji	76	Malta	111	Slovakia
7	Austria	42	Gabon	77	Mauritania	112	Slovenia
8	Azerbaijan	43	Gambia	78	Moldova	113	Solomon Islands
9	Bahrain	44	Georgia	79	Mongolia	114	Somalia
10	Bangladesh	45	Ghana	80	Montenegro	115	South Africa
11	Barbados	46	Greece	81	Morocco	116	South Sudan
12	Belarus	47	Grenada	82	Mozambique	117	Sri Lanka
13	Benin	48	Guinea	83	Myanmar	118	Sudan
14	BiH (Bosnia and Herzegovina)	49	Guyana	84	Namibia	119	Suriname
15	Bolivia	50	Hungary	85	Nepal	120	Tajikistan
16	Brunei	51	Indonesia	86	New Zealand	121	Tanzania
17	Bulgaria	52	Iran	87	Niger	122	Thailand
18	Burundi	53	Iraq	88	Nigeria	123	Togo
19	Cambodia	54	Israel	89	Niue	124	Tonga
20	Cameroon	55	Italy	90	North Macedonia	125	Trinidad Tobago
21	Cape Verde	56	Ivory Coast	91	Oman	126	Tunisia
22	Chad	57	Jamaica	92	Pakistan	127	Turkey
23	Chile	58	Kazakhstan	93	Panama	128	Uganda
24	Comoros	59	Kenya	94	Papua New Guinea	129	Ukraine
25	Costa Rica	60	Kiribati	95	Peru	130	United Arab Emirates
26	Croatia	61	Korea	96	Philippines	131	Uruguay
27	Cuba	62	Kuwait	97	Poland	132	Uzbekistan
28	Cyprus	63	Kyrgyzstan	98	Portugal	133	Vanuatu
29	Czech Republic	64	Laos	99	Qatar	134	Venezuela
30	Djibouti	65	Latvia	100	Republic of Congo	135	Vietnam
31	Dominic	66	Lebanon	101	Romania	136	Yemen
32	Dominica	67	Lesotho	102	Russia	137	Zambia
33	East Timor	68	Liberia	103	Rwanda	138	Zimbabwe
34	Ecuador	69	Libya	104	Samoa		
35	Egypt	70	Lithuania	105	Saudi Arabia		

We used the lexicon-based approach in this study due to its inherent advantages. It would be time-consuming and impractical to manually classify the sentiment in the news articles to create a large enough training dataset. Moreover, the manual classification approach would introduce human bias. There are several sentiment lexicons publicly available. We used a popular lexicon called the NRC lexicon (Mohammad and Turney 2013), which consists of 14,182 words, 2,317 positive, and 3,338 negative.

We preprocessed the data by removing stop words, punctuation, numbers, white spaces, and other words that would not convey sentiment. The sentiment analysis was done on the preprocessed data. The sentiment analysis output was a count of positive and negative sentiment words by month and by the three regional categories (China, participating countries, and nonparticipating countries). We computed a positive, negative, and overall sentiment index from this output by month and by regional category. We computed an index of positive sentiment by normalizing the number of positive sentiment words by the total number of words in that month. The negative sentiment index was computed in the same way as the positive sentiment index. An overall net sentiment index was computed by subtracting the count of negative sentiment words from the count of positive sentiment words and normalizing

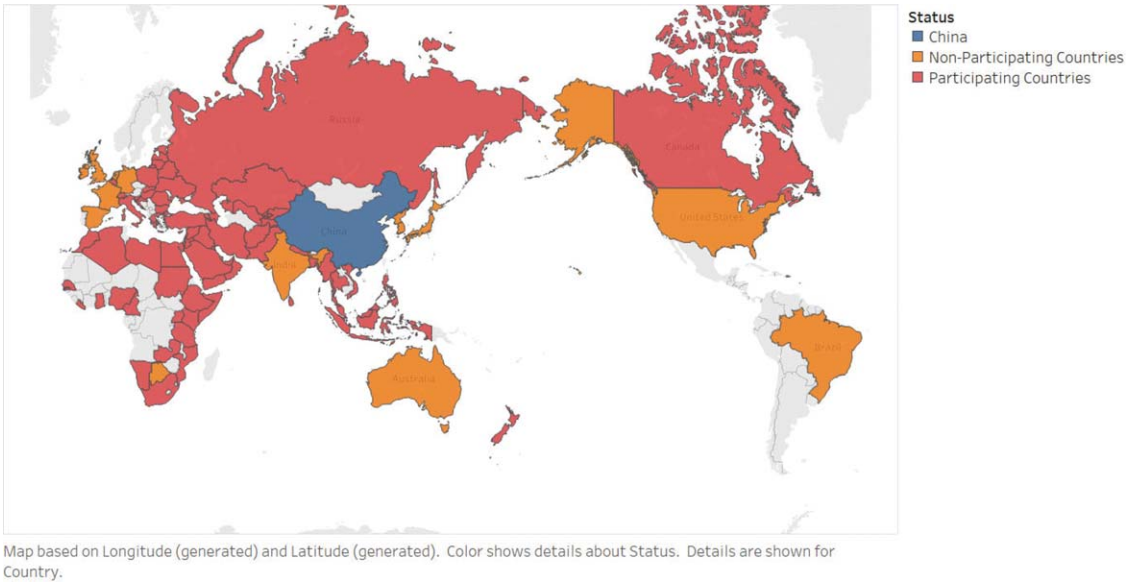

Figure 3: Participating and nonparticipating countries. Note: countries that we were not able to verify the status of have been grayed out.

this difference by the total number of words. We used the methodology described above to determine positive and negative sentiment of the BRI consistently across China, the participating countries, and the nonparticipating countries. We used the R software to perform the analysis.

6. Results and Discussion

The overall sentiment aggregated across all the countries that were part of our analysis is shown in Figure 4. As can be noted from a visual inspection of the graph, the overall positive sentiment is greater than the overall negative sentiment, thus, which leads to a net positive overall sentiment for BRI. The positive sentiment seems to be declining over time in Figures 4–7 that is, overall, China, participating countries, and nonparticipating countries, whereas the negative sentiment has held steady. The graph also shows peaks in overall positive sentiment during 2016.

Trend lines were fit to the data by using time series regressions to evaluate the trend in sentiments in a statistical manner. The dependent variables were the overall, positive, and negative sentiments, and the independent variable was time. The parameters of the time series regressions for the trend lines shown in Figures 4–7 are shown in Table 2.

The regression results show statistically significant ($p < 0.001$) negative slopes for overall sentiment and positive sentiment across all three groups: China, participating countries, and nonparticipating countries. The regression model p values were also statistically significant at the 99% level. The slopes for the negative sentiment for China and the nonparticipating countries were not statistically significant, which indicates no significant change in the negative sentiment. Thus, the statistical analysis corroborates in a more rigorous manner what can be gleaned by a visual inspection of the graphs. A further statistical test was conducted to see if the mean sentiments were different among the three groups: China, participating countries, and nonparticipating countries. The results of the analysis of variance for difference of means are shown in Table 3.

The analysis of variance results show that there was a statistically significant ($p < 0.001$) difference in all three means among the three groups. Post-hoc analysis was performed by using the Tukey HSD (honestly significant difference) multiple comparison tests. The Tukey tests showed a statistically significant ($p < 0.001$) difference in the mean sentiments among all three pairs (China versus participating countries, China versus nonparticipating countries, and participating versus nonparticipating countries).

One reason for the overall positive sentiment being greater than the overall negative sentiment could be that China has created an overall positive outlook on the program. This can be confirmed by the fact that in fewer than seven years since the project was initially proposed by President Xi Jin Ping of China, by some counts, there are 118 projects planned (Hielscher and Ibold 2020), spread across 138 countries (Belt and Road Portal 2020). However, these figures are difficult to verify because the projects are of varying types across multiple countries, and their status is not clear. There are some projects that have been agreed to on paper but have not progressed beyond the planning

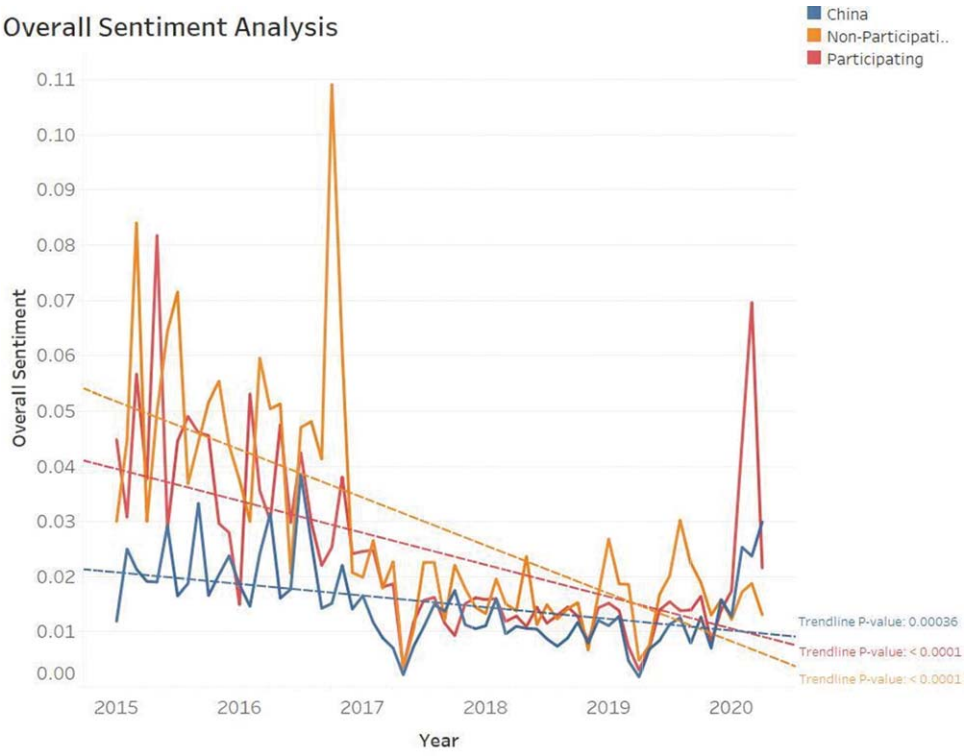

Figure 4: Overall sentiment across all countries.

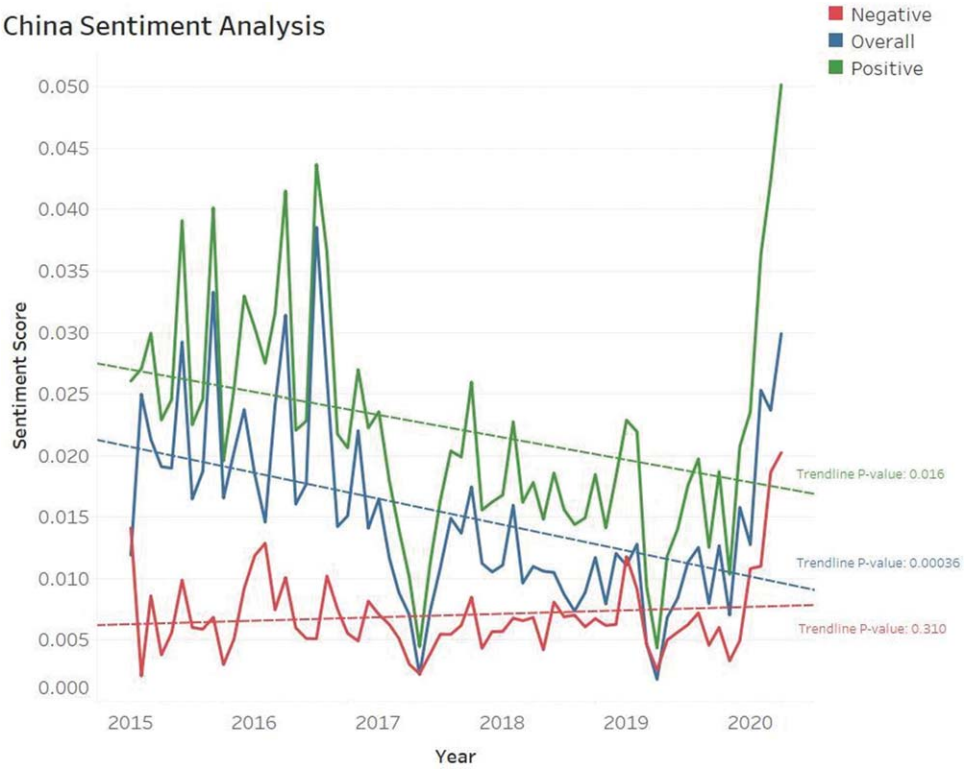

Figure 5: Sentiment analysis of news articles from China.

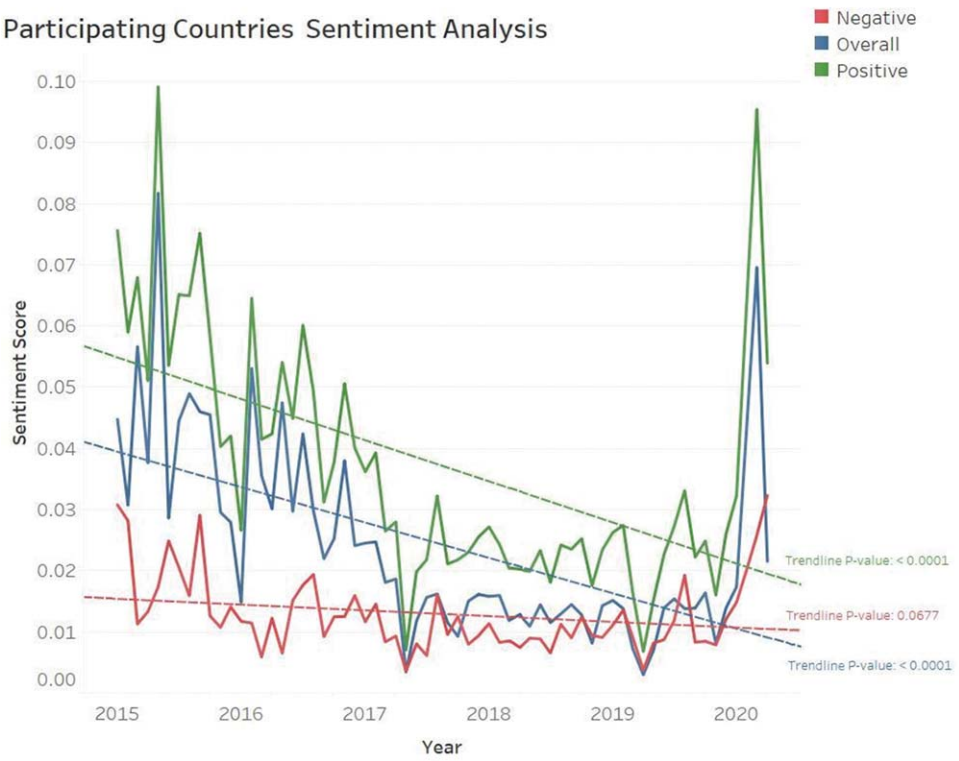

Figure 6: Sentiment analysis of news articles from participating countries.

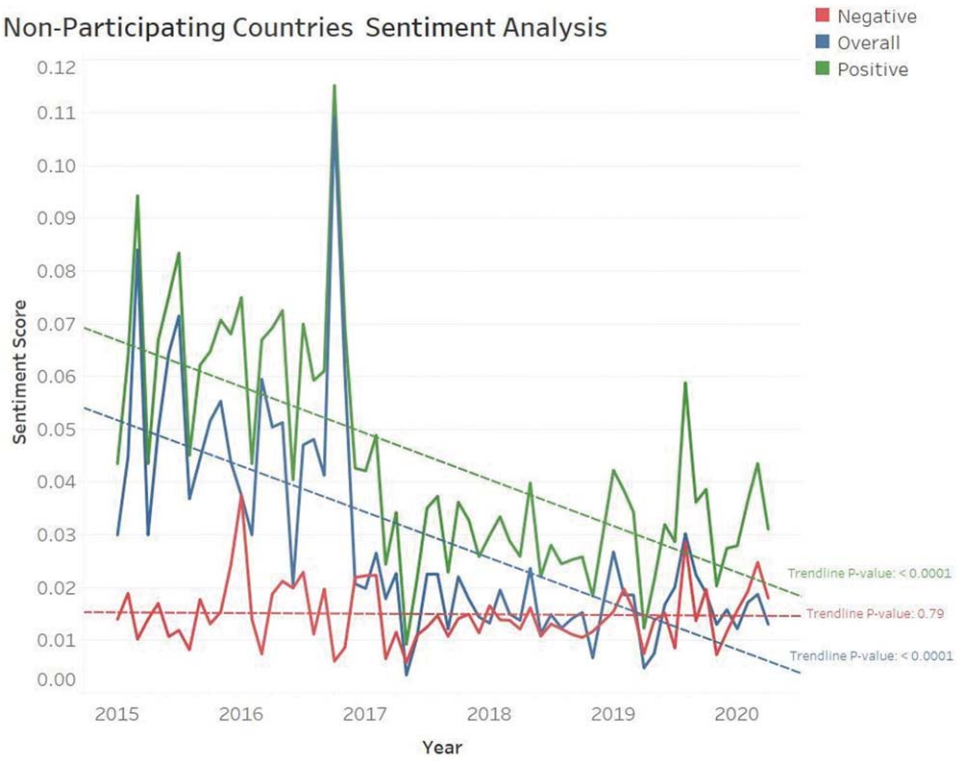

Figure 7: Sentiment analysis of news articles from nonparticipating countries.

Table 2: Time series regression (trend line) parameters.

Dependent Variable	Group	Slope	Intercept	R^2	Regression Model *p* Value
Overall sentiment	China	–0.00021 (6.37 e-05)***	0.0216	0.1544	0.0021***
	Participating	–0.00068 (7.78 e-05)***	0.0441	0.5725	4.6 e-12***
	Nonparticipating	–0.0007 (0.00014) ***	0.0519	0.3125	4.9 e-06***
Positive sentiment	China	–0.00023 (7.11 e-05)***	0.0288	0.1504	0.0024***
	Participating	–0.0009 (9.22 e-05)***	0.0615	0.5998	6.25 e-13***
	Nonparticipating	–0.00073 (0.00014)***	0.0677	0.3233	2.63 e-06***
Negative sentiment	China	−0.00002 (1.9 e-05)	0.0072	0.0193	0.2939
	Participating	–0.00017(3.75 e-05)***	0.0013	0.2721	2.26 e-05***
	Nonparticipating	–3.6 e-05 (4.37 e-05)	0.0158	0.012	0.41598

Standard errors are reported in parentheses.
***Indicates significance at the 99% level

Table 3: The analysis of variance test for comparison of mean sentiments.

	F Statistic	*p*
Overall sentiment	13.78	2.79 e-06***
Positive sentiment	26.5	8.98 e-11***
Negative sentiment	43.41	5.07 e-16***

***Indicates significance at the 99% level.

stage or have been scaled down (Chandran 2019). Other projects have been canceled and then restarted under different terms, for example, the Malaysian East Coast Rail Link (Fook 2019) and the Sri Lankan Hambantota Port Development Project (Patrick 2017). Finally, some projects have been credited to the BRI but may actually not be part of it (e.g., Mumbai Metro Line 4: India has consistently refused to join BRI). There is no denying that there is considerable excitement about the BRI program, and the more than 7,000 news articles that we collected are a testament to that. Below, we discuss some possible reasons why the overall positive sentiment seems to be declining over time, whereas the overall negative sentiment has held relatively steady over time.

One of the biggest criticisms of the BRI is the debt trap that it is accused of creating for the participating countries. The unique selling point of China as a financier was its no-strings-attached approach to lending money, unlike other financial organizations, for example, the World Bank. However, this has led to some unviable projects being undertaken and funded at very high interest rates (Wibisono 2019). The resultant debt has caused a backlash in many democratic countries with the freedom for stakeholders to voice their opinions (Balding 2018). These experiences seem to have contributed to the project's declining sentiment (Holland 2018; Rakhmat and Indramawan 2019).

Initial "euphoria" can best be witnessed by the highest positive sentiment in the eventually participating countries during May 2015. A review of the articles published during that period reveals the outreach by the Chinese government with regard to the project and the positive aspects highlighted by them. Another interesting result is that, in general, across all countries, the variation in positive sentiment (SD [Standard Deviation] = 0.017) is greater than the variation in negative sentiment (SD = 0.006). Variation in the overall sentiment seems to be highest among nonparticipating countries (SD = 0.023) as opposed to China (SD = 0.008) and the participating countries (SD = 0.015).

We have seen peaks in positive sentiment after the announcement of any new national agreement or project, which thus causes significant variation. The negative sentiment, however, seems to be steady across the period. For example, August 2019, which is a peak of both positive and negative sentiment for participating and nonparticipating countries, has announcements and updates of several national agreements (Nepal, Myanmar, Saudi Arabia, Mali, San Marino, Botswana, Thailand, Malaysia, Cambodia, Bulgaria, Turkey, Morocco, Pakistan, Iran, Kazakhstan, Nigeria, Russia, Uzbek, and the Philippines), along with the usual cautionary note, perhaps with a higher tone (e.g., Secretary Bolton's (Sputnik, 2019) and General David Petraeus' (The Australian, 2019) remarks). All three regional categories show a significant dip in positive and negative sentiment for May 2017 and April 2019. May 2017 and April 2019 were periods of Belt and Road Forums for International Cooperation. These periods had the highest number of articles (approximately 1000 in both instances), but they seemed to have a generally neutral and/or factual tone to the reporting.

6.1 Key Drivers of Sentiment

We now analyze and discuss the key drivers of the sentiments by investigating the prominent peaks and valleys in the sentiment across time. The following analysis uses word clouds to assist in identifying the key drivers of sentiments. These clouds highlight the words that drew the sentiment. The analysis follows the timeline with the corresponding word clouds. The following features are applied to the formation of the word clouds:

1. All words are processed in lower case
2. Common words, conjunctions, and symbols were ignored, e.g., and, is, "." (period),! (exclamation) etc.
3. More frequently mentioned words have a larger font.
4. Similar frequency words have the same color

China February 2016

6.1.1 China, February 2016

February 2016 was when both positive and negative (mixed) sentiments peaked in the Chinese media. These sentiments highlighted the competition faced by the Chinese in the global railroad infrastructure markets, particularly from the long-established Japanese. The positive sentiment seemed to have originated from its success in outbidding the Japanese in the contract to construct a high-speed railway linking Jakarta and Bandung in Indonesia. This positive event from the Chinese perspective seemed to have been balanced by the news of the Japanese succeeding in its bid to build a high-speed line between Mumbai and Ahmedabad in India. The success of the Japanese was seen as a threat from a strategic perspective by the Chinese.

Non-Participating Countries October 2016

6.1.2 Nonparticipating Countries, October 2016

October 2016 saw a spike in positive sentiment in articles from nonparticipating countries. This upsurge in sentiment was due to limited extreme data (highly opinionated articles) that month that skewed the results.

Participating Countries **Non-Participating Countries**

December 2016

6.1.3 Participating Countries and Nonparticipating Countries, December 2016

December 2016 saw increased negative sentiments in both participating and nonparticipating countries. This negativity had to do with the regional conflict between India (nonparticipating) and Pakistan (participating), and the implications of the BRI program and its regional component, the China-Pakistan Economic Corridor project on the two countries. The possible joining of Bangladesh to this program was also reported with concern by the Indian side.

Participating Countries August 2017

6.1.4 Participating Countries, August 2017

During August 2017, there was a peak in negative sentiment from the participating countries. The words "health," "thailand," and "minister" appear in this word cloud. China hosted the belt and road high-level meeting for health cooperation in Beijing, attended by ministers from 30 different countries, along with 300 health officials and World

Health Organization leaders. However, three key ASEAN (Association of Southeast Asian Nations) leaders were not invited to the summit: the prime ministers of Thailand, Singapore, and Brunei. China's choice not to invite the aforementioned prime ministers is the source of much negative sentiment from participating countries. Many believe the decision to omit the ASEAN leaders was in response to delays in the Thai-Chinese high-speed railway project despite the two nations conducting 18 rounds of (unsuccessful) negotiations.

China October 2017

6.1.5 China, October 2017

October 2017 illustrates a spike in positive sentiment from Chinese sources. China hosted a two-day "EU-Eurasia-China business summit" that focused on promoting the growth of the BRI into eastern areas where it had not yet started construction. During the summit, the Greek president Prokopis Pavlopoulos stated his support for becoming economically involved in the BRI. Positive sentiment in this month was primarily due to the potential growth of the BRI as a result of the successful summit.

China January 2019

6.1.6 China, January 2019

There was a peak in negative sentiment in January 2019. A Chinese article published on the seventh shares the results of a survey of respondents from ASEAN countries (Kim et al., 2019), where 70 percent of the participants agreed with the statement, "Southeast Asian countries should be cautious in negotiating with China on its flagship Belt and Road Initiative to avoid being trapped in unsustainable debt." A Chinese news article (Wilson, 2019) also expressed negative sentiment in response to Australia's reluctance to sign BRI agreements, despite their neighbor New Zealand's good relationship with China.

Non-Participating February 2019

6.1.7 Nonparticipating Countries, February 2019

February 2019 saw a peak in negative sentiment from nonparticipating and participating countries. The word cloud features "security," "terrorist," and "taliban." The word cloud and peaks in negative sentiment are in response to an attack from terrorist Masood Azhar-led Jaish-e-Mohammed (Jem) on an Indian CRPF (Central Reserve Police Force) convoy at Pulwama. The attack killed 40 people and injured many more. Many articles from participating and nonparticipating sources in August focused on the rising tensions with China and discussed the potential negative impacts of the BRI.

Non-Participating Countries August 2019

6.1.8 Nonparticipating Countries, August 2019

In August 2019, there was a peak in mixed sentiment from nonparticipating nations and a negative sentiment from participating countries. Australian sources from this time expressed heavy negative sentiment in response to the choice of New Zealand to participate in the BRI. An Indian article discussed the upcoming "BRI, China Pakistan Economic Corridor, and TransRegional Integration" conference in Pakistan with positive sentiment and explained how India could contribute to Asia's success as a continent if it participates in the BRI. While acknowledging the potential benefits that BRI could bring in, an article from Myanmar (Thein, 2019) advocated caution to ensure that the local population benefited from the projects.

6.1.9 China, Participating Countries, and Nonparticipating Countries, March 2020

During March 2020, there was a peak in both types of sentiment from all the groups. Many articles from China and their allies discussed how the COVID-19 pandemic will affect BRI's international progress. Many of these articles share the positive sentiment that, because COVID-19 may have a temporary impact on the BRI's performance, the setback will be minor. Some articles (Jianguo, 2020; Zheng & Lo, 2020) even discuss how China demonstrated good emergency preparedness and call the chain of events a success.

The sentiments around BRI could well be considered a proxy for the general perception of Chinese foreign policy global endeavors. China has increasingly been asserting its intent to take over the leadership mantle from the United States (Disis and He 2021; The Policy Planning Staff 2020). The BRI program has both economic as well as political consequences for the countries involved as well as those not involved in it. In China's own neighborhood, BRI-supporting countries, such as Indonesia, Malaysia, Myanmar, Pakistan, Philippines, Sri Lanka, Thailand, and Vietnam, have diverse sentiments about the viability and implications of the program (Chao 2021; Chin 2021; Cox *et al.* 2018; Moramudali 2016; Mursitama and Ying 2021; Punyaratabandhu and Swaspitchayaskun 2021; Qianqian and Yijun 2021; Soong and Aung 2021; Vu, Soong, and Nguyen 2021). Because the implications of this program go beyond transportation efficiency, stakeholder perceptions of this program have takeaways for global entities (commercial and political) in the way that they strategize their future. The size of the program and the constituent national projects have created financial dependence on China in many of the vulnerable economies.

The much-discussed and analyzed case of the Sri Lankan port of Hambantota, which ended up being leased to China for ninety-nine years in lieu of debt, is an example of this (Carrai 2018). If the trend of increasing economic and the resultant strategic ties between China and BRI host countries continues, it will undoubtedly have implications for the United States and other countries in the region that have uneasy relations with China, such as Australia, India, and Japan. These concerns are likely to be shared with other countries beyond Asia, in Europe and Africa. With BRI, China is also entering the Balkan region (Vangeli 2020), with implications for countries in that region. This entry leads to a three-way struggle for influence among Europe, Russia, and China, with the latter two developing a closer relation. The scope of BRI goes beyond infrastructure into the digital domain (Digital BRI) with the push by China for its 5 G telecommunication technologies and hardware (Bartholomew 2020). The influence gained by China can help it open new markets for its 5 G equipment both today as well as tomorrow.

COVID-19 has caused some slowdown in the BRI projects that may well be temporary because practically all the affected nations expect the pandemic to eventually subside (although not clear when). In fact, the pandemic itself

has created an opportunity for China to conduct health diplomacy in the form of a Health Silk Road (Chow-Bing 2020). China has been the go-to source for critical personal protective equipment and a likely source for vaccines. This explains the positive sentiment with regard to BRI, with "cooperation" being the key term for China and the participating countries in March 2020.

7. Conclusion

The results of our study can help in policy decisions. Countries can use the results of our study and adapt our analysis to understand how their public is viewing the BRI initiative and use this information in making policy decisions. Our study is among the first to use sentiment analysis, a NPL technique, to explore the global perception of the BRI. Our findings of a declining and increasingly volatile positive sentiments and a steady and/or stable negative sentiment sheds light on how the global perspective on BRI has changed over time. The spike in interest around the period when BRI-related forums were hosted in Beijing, China, indicates the efficacy of those initiatives by the Chinese government. However, the declining positive sentiment seems to suggest a more cautious perspective that could be the result of the outcomes of the current projects or the geopolitical environment. The fact that the sentiment is still overall positive bears testament to the fact that countries appreciate this initiative of President Xi. This positive sentiment is a takeaway for what could be a successful foreign policy.

The US withdrawal from Afghanistan has opened the doors for China's entry offering support for the war-torn country in the form of infrastructure development under the BRI. Afghanistan is rich in mineral wealth and can provide a connection to central Asia from the port of Gwadar in Pakistan, which has been developed as a part of the BRI and the China-Pakistan Economic Corridor. China has stepped up as one of the first countries to develop relations with the Taliban government in Afghanistan. These developments will no doubt have implications for the BRI program. Those committed to either supporting or rejecting the BRI-affiliated projects will probably continue to do so. However, it would be interesting to see the sentiments emerging from countries that were open to but did not actually initiate any projects. Our research makes it possible for a repeat of this analysis by using the latest news articles to show the change if any in attitudes to the BRI as well as to China.

One of our study's limitations is that the sentiment analysis does not lead to causal inferences. A further limitation is that our results are dependent on the lexicon we picked for the sentiment analysis, although the lexicon we picked is a popular and comprehensive one. Further research could be performed with different lexicons to assess if the results change significantly with the choice of the lexicon. Although our in-depth analysis of the drivers of the sentiment sheds light on the causes, further research needs to be performed to explore in detail the causes for the changes in sentiments over time. Another limitation is that we have only considered English language news articles in our analysis. An analysis of local language news articles is currently infeasible with the current technology because lexicons for local languages are not available. But, this could be an avenue for future research as the technology improves and local lexicons become available. Other avenues for further research include drilling down to the individual country level and exploring changes in sentiment and further exploring the change in sentiment around certain major BRI-related events, particularly project cancellations.

References

Adrien, M. 2019. "OFNRS - Observer, Analyser et Conseiller." Accessed January 23, 2021. https://observatoirenrs.com/.

Ali, Z., Ö. Gökçe, M. Binark, and A. D. Gidreta. 2020. "China-Pakistan Economic Corridor and Technicians of Opinion in Pakistani Twittersphere: A Thematic Content Analysis." *Asya Araştırmaları Uluslararası Sosyal Bilimler Dergisi* **4**, no. 1: 9–28.

Arifon, O., Z. A. Huang, Y. Zheng, and A. Zyw Melo. 2019. "Comparing Chinese and European Discourses Regarding the 'Belt and Road Initiative'." *Revue Française Des Sciences de L'information et de la Communication* **17**.

Balding, C. 2018. "Why Democracies Are Turning against Belt and Road." *Foreign Affairs* **97**, no. 5.

Bartholomew, C. 2020. "China and 5G." *Issues in Science and Technology* **36**, no. 2: 50–57. Accessed January 23, 2021. https://www.proquest.
com/scholarly-journals/china-5g/docview/2452125915/se-2.

Belt and Road Portal. 2020. "List of countries that have signed cooperation documents with China for the Belt and Road Initiative." Accessed January 23, 2021. https://www.yidaiyilu.gov.cn/info/iList.jsp?tm_id=126&cat_id=10122&info_id=77298

Bērziņa-Čerenkova, U. A. 2016. "BRI instead of OBOR – China edits the English Name of its most ambitious international project." Accessed January 23, 2021. http://liia.lv/en/analysis/bri-instead-of-obor-china-edits-the-english-name-of-its-most-ambitious-international-project-532

Bloomberg. 2019. "Analysis |China's Belt and Road is getting a reboot. Here's why." *Bloomberg News*. Accessed https://www.washingtonpost.com/business/chinas-belt-and-road-is-getting-a-reboot-heres-why/2019/08/14/6725fe4e-be63-11e9-a8b0-7ed8a0d5dc5d_story.html

Carrai, M. A. 2018. "China's Malleable Sovereignty Along the Belt and Road Initiative: The Case of the 99-Year Chinese Lease of Hambantota Port." *New York University Journal of International Law and Politics* **51**: 1061.

Chandra, J. K., E. Cambria, and A. Nanetti. 2020. "One Belt, One Road, One Sentiment? A Hybrid Approach to Gauging Public Opinions on the New Silk Road Initiative." Paper Presented at the 2020 International Conference on Data Mining Workshops (ICDMW), Sorrento, Italy, November 17–20.

Chandran, N. 2019. "Fears of excessive debt drive more countries to cut down their Belt and Road investments." Accessed January 23, 2021. https://www.cnbc.com/2019/01/18/countries-are-reducing-belt-and-road-investments-over-financing-fears.html.

Chao, W.-C. 2021. "The Philippines' Perception and Strategy for China's Belt and Road Initiative Expansion: Hedging with Balancing." *The Chinese Economy* **54**, no. 1: 48. doi: 10.1080/10971475.2020.1809817

Chatzky, A., and J. McBride. 2019. 05/21/2019). "China's massive Belt and Road initiative." Accessed January 23, 2021. https://www.cfr.org/backgrounder/chinas-massive-belt-and-road-initiative.

Chin, K. F. 2021. "Malaysia's Perception and Strategy toward China's BRI Expansion: Continuity or Change?" *The Chinese Economy* **54**, no. 1: 9–11. doi: 10.1080/10971475.2020.1809814

Chow-Bing, N. 2020. "*COVID-19, Belt and Road Initiative and the Health Silk Road: Implications for Southeast Asia*." Paper Presented at the 4th NACAI International Symposium 21 November 2020 University of Yangon (Centennial).

Cox, M., T. S. M. Majid, Y. Jie, J. Yan, and H. Hamzah. 2018. "China's Belt and Road Initiative (BRI) and *Southeast* Asia." *CIMB ASEAN Research Institute* **47**.

Das, S., G. Medina, L. Minjares-Kyle, and Z. Elgart. 2018. "Social Media Hashtags Associated with Bike Commuting: Applying Natural Language Processing Tools (No. 18-03545)." In *Paper Presented at the Transportation Research Board 97th Annual Meeting*, Washington DC, January 7–11, 2018.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz. 2013. "Predicting Depression via Social Media." In *Proceedings of the International AAAI Conference on Web and Social Media* **7**, no. 1: 128–137.

Disis, J., and L. He. 2021. "China is rehearsing for when it overtakes America." *CNN Business*. Accessed January 23, 2021. https://www.cnn.com/2021/01/26/economy/china-xi-economy-intl-hnk/index.html.

Evans-Cowley, J. S., and G. Griffin. 2012. "Microparticipation with Social Media for Community Engagement in Transportation Planning." *Transportation Research Record: Journal of the Transportation Research Board* **2307**, no. 1: 90–98. doi: 10.3141/2307-10

Fook, L. L. 2019. China-Malaysia relations back on track? Accessed January 23, 2021. http://hdl.handle.net/11540/10256.

Griffiths, R. T. 2017. *Revitalising the Silk Road: China's Belt and Road Initiative*. Leiden, The Netherlands: Hipe Publications.

Gu, Y., Z. S. Qian, and F. Chen. 2016. "From Twitter to Detector: Real-Time Traffic Incident Detection Using Social Media Data." *Transportation Research Part C: emerging Technologies* **67**: 321–42. doi: 10.1016/j.trc.2016.02.011

Hielscher, L., and S. Ibold. 2020. "Belt and Road initiative." Accessed January 23, 2021. https://www.beltroad-initiative.com/projects/

Holland, T. 2018. "Why borrowers on China's Belt and Road will go from Euphoria to depression." *South China Morning Post*. Accessed https://www.scmp.com/week-asia/opinion/article/2138499/why-borrowers-chinas-belt-and-road-will-go-euphoria-depression.

Hurley, J., S. Morris, and G. Portelance. 2019. "Examining the Debt Implications of the Belt and Road Initiative from a Policy Perspective." *Journal of Infrastructure, Policy and Development* **3**, no. 1: 139–75. doi: 10.24294/jipd.v3i1.1123

Indermit Gill, S. V. L., and L. Mathilde. 2019. "Winners and losers along China's Belt and Road." Accessed January 23, 2021. https://www.brookings.edu/blog/future-development/2019/06/21/winners-and-losers-along-chinas-belt-and-road/

Jianguo, W. (2020, 03/06/2020). Coronavirus an opportunity for BRI healthcare cooperation. Global Times. Accessed through Factiva 04/23/2020.

Kim, J., S. Lewis, and C. Fernandez. (2019, 01/07/2019). Southeast Asia wary of China's Belt and Road project, sceptical of U.S.: survey. Macau Business Daily. Accessed through Factiva 03/11/2019.

Li, J., Y. Yao, Y. Xu, J. Li, L. Wei, and X. Zhu. 2019. "Consumer's Risk Perception on the Belt and Road Countries: Evidence from the Cross-Border e-Commerce." *Electronic Commerce Research* **19**, no. 4: 823–40. doi: 10.1007/s10660-019-09342-x

Li, Y., Y. Zhang, L. A. Tiffany, R. Chen, M. Cai, and J. Liu. 2021. "Synthesizing Social and Environmental Sensing to Monitor the Impact of Large-Scale Infrastructure Development." *Environmental Science and Policy* **124**: 527–40. https://doi.org/10.1016/j.envsci.2021.07.020. doi: 10.1016/j.envsci.2021.07.020

Liu, Y., Y. Li, and W. Li. 2019. "Natural Language Processing Approach for Appraisal of Passenger Satisfaction and Service Quality of Public Transportation." *IET Intelligent Transport Systems* **13**, no. 11: 1701–1707. doi: 10.1049/iet-its.2019.0054

Loughran, T., and B. McDonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* **54**, no. 4: 1187–230. doi: 10.1111/1475-679X.12123

Malik, T. H. 2020. "The Belt and Road Initiative (BRI) Project Legitimisation: The Rhetor's Innovation and the US Response." *Asian Journal of Comparative Politics* doi: 10.1177/2057891120959476

Mark, J. J. 2019. "Silk Road." Accessed January 23, 2021. https://www.ancient.eu/Silk_Road/

Mehrotra, S., and S. Roberts. 2018. "Identification and Validation of Themes from Vehicle Owner Complaints and Fatality Reports Using Text Analysis." In 97th Annual Meeting of the Transportation Research Board, Washington, DC.

Mohammad, S. M., and P. D. Turney. 2013. "Crowdsourcing a Word–Emotion Association Lexicon." *Computational Intelligence* **29**, no. 3: 436–465. doi: 10.1111/j.1467-8640.2012.00460.x

Moramudali, U. 2016. "Sri Lanka's Debt and China's money." *The Diplomat.* Accessed January 23, 2021. https://bandapost.org/wp-content/uploads/pdf/Dec%20Srilanka%20debt%20and%20China%20money.pdf.

Mursitama, T. N., and Y. Ying. 2021. "Indonesia's Perception and Strategy toward China's OBOR." *The Chinese Economy* **54**, no. 1: 35–47. doi: 10.1080/10971475.2020.1809816

Napitupulu, A., M.A. Embi, and B. Briando. 2020. "Public Sentiment Analysis on the Existence of Foreign Worker during the Covid 19 Pandemic." Paper Presented at the International Conference on Law and Human Rights. Virtual Conference, October 26–27, 2020.

Niu, W., and L. Wu. 2019. "Sentiment Analysis and Contrastive Experiments of Long News Texts." Paper Presented at the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China December 20–22.

Patrick, A. 2017. China-Sri Lanka Strategic Hambantota Port Deal. National Maritime Foundation, Accessed January 23, 2021. http://www.maritimeindia.org/View%20Profile/636276610966827339.pdf

Punyaratabandhu, P., and J. Swaspitchayaskun. 2021. "Thailand's Perception and Strategy toward China's BRI Expansion: Hedging with Cooperating." *The Chinese Economy* **54**, no. 1: 69–69. doi: 10.1080/10971475.2020.1809819

Qianqian, L., and L. Yijun. 2020. "The China-Pakistan Economic Corridor: The Pakistani Media Attitudes Perspective." *Technology in Society* **62**: 101303. https://doi.org/10.1016/j.techsoc.2020.101303. doi: 10.1016/j.techsoc.2020.101303

Rakhmat, M. Z., and D. Indramawan. 2019. Belt and Road Initiative: Challenging South and Southeast Asia – Analysis. Accessed January 23, 2021. https://www.eurasiareview.com/15112019-belt-and-road-initiative-challenging-south-and-south-east-asia-analysis/

Rithmire, M., and Y. Li. 2019. "Chinese Infrastructure Investments in Sri Lanka: A Pearl or a Teardrop on the Belt and Road?" Harvard Business School Case 719-046, January 2019. Accessed January 23, 2021. https://www.hbs.edu/faculty/Pages/item.aspx?num=55410

Soong, J.-J., and K. H. Aung. 2021. "Myanmar's Perception and Strategy toward China's BRI Expansion on Three Major Projects Development: Hedging Strategic Framework with State-Market-Society Analysis." *The Chinese Economy* **54**, no. 1: 20–34. doi: 10.1080/10971475.2020.1809815

Sputnik. (2019). Beijing Opposes Bolton's Remarks on Dangers of Chinese Investments - Foreign Ministry. Sputnik News Service.

Teo, H. C., A. Campos-Arceiz, B. V. Li, M. Wu, and A. M. Lechner. 2020. "Building a Green Belt and Road: A Systematic Review and Comparative Assessment of the Chinese and English-Language Literature." *Plos ONE* **15**, no. 9: e0239009. doi: 10.1371/journal.pone.0239009

The Australian. (2019, 08/14/2019). Petraeus crystallises challenges. The Australian. https://www.theaustralian.com.au/

The Policy Planning Staff. 2020. "The elements of the China Challenge." Accessed January 23, 2021. https://www.state.gov/wp-content/uploads/2020/11/20-02832-Elements-of-China-Challenge-508.pdf.

Thein, A. Z. P. (2019, 08/02/2019). The Belt and Road Initiative in Myanmar: staring down the dragon. Frontier Myanmar. Accessed through Factiva 03/11/2019.

UNESCO. 2019. "About the Silk Road | SILK ROADS." Accessed https://en.unesco.org/silkroad/about-silk-road.

Vangeli, A. 2020. "China's Belt and Road in the Balkans in the Post-COVID-19 Era." European Institute of the Mediterranean. *Mediterranean Yearbook 2020.* Accessed January 23, 2021. https://www.iemed.org/observatori/arees-danalisi/arxius-adjunts/anuari/med.2020/China_Belt_Road_Balkans_Anastas_Vangelis_IEMed_YearBook2020.pd.

Vu, V.-H., J.-J. Soong, and K.-N. Nguyen. 2021. "Vietnam's Perceptions and Strategies toward China's Belt and Road Initiative Expansion: Hedging with Resisting." *The Chinese Economy* **54**, no. 1: 56–68. doi: 10.1080/10971475.2020.1809818

Whitfield, S. 2007. "Was There a Silk Road?" *Asian Medicine* **3**, no. 2: 201–13. https://doi.org/10.1163/157342008X307839. doi: 10.1163/157342008X307839

Wibisono, A. N. 2019. "China's 'Belt and Road Initiative'." *In Sri Lanka: Debt Diplomacy in Hambantota Port Investment. Mandala: Jurnal Ilmu Hubungan Internasional* **2**, no. 2: 222–245.

Wilson, K. (2019, 01/24/2019). Canberra clings to outdated alliance. China Daily-Global Edition. Accessed through Factiva 01/23/2021.

Zheng, S., and K. Lo. (2020, 03/05/2020). Coronavirus: China keen to promote its success in controlling epidemic. South China Morning Post. Accessed through Factiva 04/23/2020.

Zhou, J., J. Wang, Q. Li, Y. Wen, Y. Zhang, and P. Lin. 2021. "A Public Opinion Analysis System Based on Emotion Analysis in Linyi." *Journal of Physics: Conference Series* **1757**, no. 1: 012116. doi: 10.1088/1742-6596/1757/1/012116

WWW.JBDTP.ORG

ISSN: 2692-797

JBDTP Professional Vol. 1, No. 1, 2022 DOI: 10.54116/jbdtp.v1i1.19

FOUR-CLASS EMOTION CLASSIFICATION PROBLEM USING DEEP LEARNING CLASSIFIERS

Miaojie Zhou
School of Graduate Professional Studies, Penn State University, Malvern, PA
mjz5304@psu.edu

Satish Mahadevan Srinivasan
School of Graduate Professional Studies, Penn State University, Malvern, PA
sus64@psu.edu

Abhishek Tripathi
Department of Accounting and IS, The College of New Jersey, Ewing, NJ
tripatha@tcnj.edu

ABSTRACT

Social media sites and blogs generate a vast amount of emotionally rich data in the form of tweets, status updates, blog posts, etc. Such textual data represent emotions expressed by an individual or a group of people on any given topic. By analyzing the emotions within these textual data, we can get an idea about how individuals or communities express their views. Analytical techniques are widely used for analyzing emotions within these texts. However, due to the training datasets' imbalanced nature, the supervised classifiers fail to classify the different emotion classes. As a result, these classifiers demonstrate a poor performance in identifying emotions within the texts. Here, using a constructed heterogeneous training dataset from well-known training datasets we have trained two deep learning models namely the Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) to address a four-class emotion (Anger, Sadness, Happy, and Surprise) classification problem. By appropriately tuning the deep learning classifiers' hyperparameters, our study reveals that the CNN classifier has slightly better performance (77%) than the RNN classifier (76%) for a four-class emotion classification problem.

Keywords *Emotion classification, Deep learning, Supervised classifiers, 10-Fold validation, Word embeddings*

1. Introduction

Human beings have mostly expressed their emotions either by speech or through written texts. With the emergence of social media and blogging sites, individuals and communities have found a way to freely express their opinions, feelings, and thoughts on various topics through texts. Irrespective of the number of characters, texts often hold a wealth of information on how individuals or communities communicate their thoughts, emotions (happiness, anxiety, and depression), and feelings within their network. By analyzing the corpus of texts from the social media and

blogging site, one can learn not only the emotions of individuals but also the emotions of larger groups (such as a certain country, state etc.). More commonly expressed, emotions include *anger*, *disgust*, *fear*, *happiness*, *surprise*, *sadness*, *tensed*, etc. For example, the text "I felt quite *happy* and lighthearted; I put on the shoes and danced and jumped about in them" expresses a happy emotion. Not only one, but more than one emotion can be expressed within a text. Since texts lack structure and size, determination of emotions, i.e., *Emotion classification*, is a very challenging task.

Even though sentiment analysis has been widely studied in the field of Data Mining and Machine Learning, it does not address the wide range of emotions that are associated with human behavior. Moreover, it is important to know the exact emotion behind a topic rather than a generic sentiment. Since several different emotions are expressed within a text (sentence), it becomes necessary to analyze each sentence within a document to determine the overall emotion. Ghazi, Inkpen, and Szpakowicz (2010) found that emotion expressions tend to be the most informative in an expressive sentence, so emotion classification is practically important to text summarization. Nowadays, the popularity of short messages within social media and blogging sites are replacing the traditional electronic document. Mining emotions within these short messages is another challenge for emotion classification. Our study is significant for researching the use of deep learning in the short, textual data and the applicability and practical use across domains in actual life, such as the social network posts and movie review analysis. Coviello *et al.* (2014) found that online social networks may magnify the intensity of global emotional synchrony. Determining the emotional category on IMDB datasets accurately predicts the human's preference/interest in movies (Liu 2020), which would further affect film studio and cinema's decision-making in next quarter (Liu 2020).

Unlike conventional texts, short messages are peculiar in structure and size. Adding to that is the language used by people within these texts to express their emotions which is very different from the digitized documents (Ling and Baron 2007). A major challenge is also posed by the availability of many features within the texts. There are also challenges associated with manually classifying the texts into different emotion types. Manually annotating the texts may be ambiguous at times and does not guarantee complete accuracy (Hasan, Rundensteiner, and Agu 2014). Also, the inherent nature of the different types of emotions makes it very difficult to differentiate between them. According to the Circumplex model, there are 28 affect words or emotions. In this model, several emotions are clustered so close to each other that it becomes very hard to differentiate between them. There is always a high probability of mislabeling the emotions that are clustered so close to each other. For example, sadness and depression are two emotions that are very close to each other, that it is hard to differentiate between them (Russell 1980). All these factors together inhibit the classifiers from learning the critical features that can enable it to identify emotions within the text.

Previous studies have focused on exploring the potentiality of different classification techniques using the bag-of-words (BOW) and/or the *n*-grams as features (Aman and Szpakowicz 2007; Diman, Inkpen, and Szpakowicz 2010; Chaffar and Inkpen 2011; Hasan, Rundensteiner, and Agu 2014; Badshah *et al.* 2016). These studies have used traditional machine learning classifiers, but none have explored the deep learning classifiers' potentiality for emotion classification.

In this paper, we have demonstrated both the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) classifiers' potentiality for the four-class emotion classification problem. We have achieved an average of 76–77% accuracy when classifying 277 instances in the testing dataset. According to the observational results, CNN slightly outperformed RNN in classifying all the four emotion types. Still, they both misclassified a significant number of instances from the surprise class to the happy class.

The rest of the paper is organized as follows. In section 2, we report the survey of the literature. In section 3, we detail the materials and methods employed in this study. In section 4, we present the results from this study and discuss our findings. Finally, in section 5, we conclude the paper and outline the future direction of our research.

2. Related Work

While classifying emotions based on textual data is a relatively new research area, it has attracted lots of attention. Bhowmick, Basu, and Mitra (2010) observed the same relative performance exhibited by humans and machines on different data sets in the emotion classification task, so the high accuracy achieved by machine learning or deep learning could be trusted. Nowadays, approaches employing Deep neural networks have been widely studied for emotion classification in textual data. This technological advancement significantly outperforms other off-the-shelf models (Chatterjee *et al.* 2019). CNNs have been a popular choice in several different works. Simple CNN with slight hyperparameter tuning demonstrated excellent results on multiple benchmarks, including the fine-grained Stanford Sentiment Treebank (SST) for binary classification (Kim 2014). Kalchbrenner, Grefenstette, and Blunsom (2014) have discussed using a Dynamic CNN (DCNN) for Twitter sentiment prediction. According to them, the

DCNN can handle varying lengths of input sentences and is easily applicable to any language due to the usage of Dynamic k-Max Pooling, which reduced the error rate by 25% (Kalchbrenner, Grefenstette, and Blunsom 2014). Acharya *et al.* (2018) have proposed a complex 13-layer CNN architecture for emotion detection in EEG signals.

On the other hand, RNNs are designed to handle sequence problems and have gained much attention over time. Lai *et al.* (2015) introduced a recurrent CNN that automatically judges and captures the key components in texts that can help boost the experiments' accuracy. Abdul-Mageed and Ungar (2017) have proposed a core model of Gated RNNs (GRNNs) and a modern variation of RNN for classifying emotions in several dimensions with high accuracies. Kratzwald *et al.* (2018) have reported that both RNN and sent2affect (transfer learning) consistently outperform the traditional machine learning algorithms across six benchmark datasets.

Wang *et al.* (2016) have proposed a regional CNN-LSTM model to predict the VA ratings of texts in the SST corpus. Zhang *et al.* (2019) have proposed a Coordinated CNN- LSTM-Attention (CCLA) model using the Soft-Max regression classifier on four datasets including Movie Review Data (MR), Large movie review (IMDB), TREC question dataset (TREC), and Subjectivity dataset (SUBJ). Socher *et al.* (2013) have introduced Recursive Neural Tensor Network (RNTN) for the famous SST dataset. They have reported 85.4% accuracy for sentiment classification. Irsoy and Cardie (2014) have proposed a deep recursive neural network constructed by stacking multiple recursive layers very similar to the conventional deep feed-forward networks. As a result, they achieved 50% accuracy for the task of fine-grained sentiment classification (five-classes of SST dataset; Irsoy and Cardie 2014).

Several other studies have been reported relating to textual-based emotion classification. Jabreel and Moreno (2019) have proposed a novel method, Binary Neural Network (BNet) for emotion classification on Twitter data (SemEval2018 Task 1: E-c multi-label emotion classification) and have reported an accuracy score of 59%. Zheng *et al.* (2014) have trained a deep belief network (DBN) and have achieved an accuracy of 86.91% and 87.62% in the experiments of DBN and DBN-HMM models, respectively. Zhou *et al.* (2016) propose a BLSTM architecture that helped capture long-term sentence dependencies and introduced a combined model BLSTM-2DCNN, which achieves 52.4% accuracy on SST binary classification and fine-grained classification tasks. Hamdi, Rady, and Aref (2020) utilized CNN streams and the pretrained word embeddings (Word2Vec) and achieved 84.9% accuracy from the Stanford Twitter Sentiment dataset. Zhang, Lee, and Radev (2016) presented the Dependency Sensitive CNNs (DSCNNs) that outperforms traditional CNNs and achieved 81.5% accuracy in the sentiment analysis of Movie Review Data (MR) proposed by Pang and Lee (2005). Zhou *et al.* (2015) proposed a novel model called C-LSTM for sentence representation utilizing both CNN and LSTM to achieve 49.2% in five-class classification tasks.

In the next section, we discuss the materials and methods employed in this study.

3. Materials and Methods

3.1 Dataset

Here we use a synthetic dataset SYN80 constructed by combining the Alm's dataset (Alm 2008; Chaffar and Inkpen 2011) and Aman's dataset (Aman and Szpakowicz 2007; Chaffar and Inkpen 2011). Instances from both the Alm's and Aman's datasets were combined and reshuffled. After reshuffling, 80% of the instances were randomly picked without replacement to create the synthetic dataset SYN80. The SYN80 dataset contains instances, each represented as *{text, emotion-class}*. We ensured that all the instances in this dataset have a single annotation of an emotion class (Srinivasan and Ramesh 2018). In SYN80, both with resampling and without resampling, the performance of the traditional classifiers was significantly boosted with few exceptions for a four-class emotion classification problem (Srinivasan and Ramesh 2018). Table 1 lists the class-wise distribution of the number of instances in the SYN80 dataset.

3.2 Models

The deep neural networks have the ability to learn and model nonlinear and complex relationships, which is important because, in the context of the emotion classification problem, many of the relationships between inputs and outputs are nonlinear as well as complex (Kalchbrenner, Grefenstette, and Blunsom 2014; Abdul-Mageed and Ungar 2017; Chatterjee *et al.* 2019; Jabreel and Moreno 2019; Zhang *et al.* 2019; Hamdi, Rady, and Aref 2020). Therefore,

Table 1: Class-wise distribution of the number of instances in the SYN80 dataset.

Class	Happy	Anger	Sadness	Surprise
Number of sentences	739	133	325	186

in this study, we have explored both the CNNs and RNNs architecture and have compared the results from these two classifiers. To optimize the hyperparameters, we employed the grid search technique. As discussed in literature (Khorrami *et al.* 2016; Lakomkin, Bothe, and Wermter 2018; Al Machot *et al.* 2019; Ghosal *et al.* 2019), upon performing the grid search, we determined the optimal numbers for the filters and layers for both the CNN and RNN models. The grid search functionality was implemented in Python using the GridSearchCV class in the Scikit-learn library. An exhaustive list of values was provided for tuning the hyperparameters. The performance of the models was compared at different stages to finalize the values for the filters and layers for both CNN and RNN.

3.2.1 CNNs

CNNs is a network that employs convolution operation in at least one layer that works well, especially in image data or a grid of values. Convolution is a specialized kind of mathematical operation that can be used in one-dimension, two-dimension, and multi-dimension. The convolution layer abstracts or scales the input matrix to a feature map using multiplication or dot product. As mentioned in the related work section, CNN is an improved and developed neural network. The CNN architecture employed in this research consists of five different types of layers. (1) The embedding layer that consists of pretrained weights for the words and maps each input word to a 300-dimensional vector. (2) Convolutional layer that consists of filters (kernels) which slide across the embedding layer; this layer is used to obtain the feature map. (3) MaxPooling layer or down-sampling layer that performs the maxpooling operation to reduce the dimension of output neurons and computational intensity, thus preventing overfitting; the maxpooling operation selects only the maximum value in each feature map. (4) Dense (fully connected) layer that has a full connection to all the activations in the previous layer. (5) And flatten layer in Keras that reshapes the tensor to have a shape that is equal to the number of elements contained in the tensor. The architecture of the CNN model and the number of neurons in each layer is shown below (Table 2).

The CNN network consists of 3 convolution layers with MaxPooling after each layer. For the first two convolutional layers, the filters were set to 150, kernel size was specified to 2, padding used was 'same,' and the activation function used was 'ReLU.' However, for the third convolutional layer the filters were set to 128. The pool size was set to 5x5 for the first two MaxPooling layers. The third MaxPooling layer does global pooling with a size of 24. Then we added a flatten layer whose output was then fed to a Dense layer consisting of 50 neurons. Since this CNN is classifying four categories, the output layer was set to four outputs. For the loss function, we used the cross-entropy, and Adam optimizer was set to minimize the loss.

Here we use two activation functions: (1) ReLU and (2) SoftMax. As highly recommended, we have used an activation function after every convolutional layer. The leaky rectifier linear unit (LeakyReLU) was used as an activation function for the convolutional layers 2, 4, and 6 to add nonlinearity and sparsity in the network structure. We also used the SoftMax activation function in the output layer to boost the performance of the multi-class classification.

3.2.2 RNNs

RNNs are networks that process a sequence of values and have a chain-like structure which means that connections between nodes form a directed graph along a temporal sequence. RNNs have the capability to scale to much longer sequences. According to the literature review, RNN is widely used in emotion classification problems due to its temporal dynamic property. It uses internal memory to process the sequence of inputs and can process sequences of variable length.

Table 2: The details of the CNN structure used in this research.

Layers	Type	Number of Neurons (Output Shape)
1	Embedding	(600,300)
2	Convolution	(600,150)
3	MaxPooling	(120,150)
4	Convolution	(120,150)
5	MaxPooling	(24,150)
6	Convolution	(24,128)
7	MaxPooling	(1,128)
8	Flatten	128
9	Dense (fully connected)	50
10	Dense (fully connected)	4

Table 3: The details of the RNN structure used in this research.

Layers	Type	Number of Neurons (Output Shape)
1	Embedding	(600,300)
2	LSTM	128
3	Dense (Fully Connected)	64
4	Dropout	64
5	Dense (Fully Connected)	4

One of the most important layers in RNN is the LSTM layer, which has shown success in various other domains. We use one LSTM layer after the embedding layer to processes the text from left to right. Deep neural networks tend to run the risk of overfitting, especially when the training dataset is small. Consequently, we use the dropout parameter within the recurrent layer, which is equivalent to randomly dropping out connections between the recurrent LSTM cells. We also use a dropout layer between the output of a 64-layer dense network and the output layer. The architecture of the RNN model and the number of neurons in each layer are shown below (Table 3).

A fiev-layer RNN has been employed in this study. The dimension of word embeddings used here is 300, the number of hidden units in the LSTM layer is 128. Within the LSTM layer, we set the dropout and the recurrent dropout parameter to 0.2. We have used ReLU as the activation function in the first dense layer with 64 hidden units. For regularization, we employ the Dropout operation with a dropout rate of 0.5. The final output layer (dense layer) consists of four outputs since we address a four-class classification problem. We also use SoftMax as the output layer. The loss function used here is the cross-entropy and Adam optimizer, which is set to minimize the loss.

In the next section, we discuss the experimental procedures performed in this study.

3.3 Experimental Designs

3.3.1 Preprocessing

The preprocessing begins with shuffling the instances in the SYN80 dataset to maintain randomness in classes before the model is trained. The textual instances were corrected for spelling mistakes. Then, all the stop words and special characters were removed. All the textual instances were converted into lower case. The instances were then padded to a length of 600 characters. Instances that were longer than 600 characters were trimmed. Next, we performed tokenization and retained the top 1500 words for our analysis. We used the Global Vectors for Word Representation (GloVe; Pennington, Socher, and Manning 2014) to initialize the weights for these words and then created the embedding matrix.

3.3.2 Ten-fold cross-validation

A 10-fold cross-validation approach was followed in this study. First, the preprocessed dataset was split into the train and test set following the ratio 80:20, respectively. When training the model, we performed 10-fold cross-validation using the training dataset. For performing the 10-fold cross-validation technique, we initially partitioned the training dataset into 10-folds. Across the 10 different experiments performed, in each experiment, 9-folds of the dataset were used for training the model, and the remaining 1-fold was used for validating the model. Finally, the model was tested on the test dataset. In the testing phase, we performed 10 different experiments using 10 different models and then averaged the resulting metric accuracy. For each experiment, we used the same test dataset but across different models. Here we report the validation and test accuracy, sensitivity(recall), F1-score, and precision all averaged across 10 different experiments.

3.3.3 Performance measures

In this study, we have tried to address a four-class classification problem. Here, we report the values for *Accuracy*, *Recall*, *Precision*, and *F1-score*. *Accuracy* is the ratio of the number of truly predicted samples to the total number of samples in the test dataset. *Precision* is the fraction of relevant instances among the retrieved instances and *Recall* is the fraction of the relevant instances that have been retrieved over the total number of relevant instances. *F1-score* is the harmonic mean of precision and recall. All the above discussed performance measures were computed using the confusion matrix.

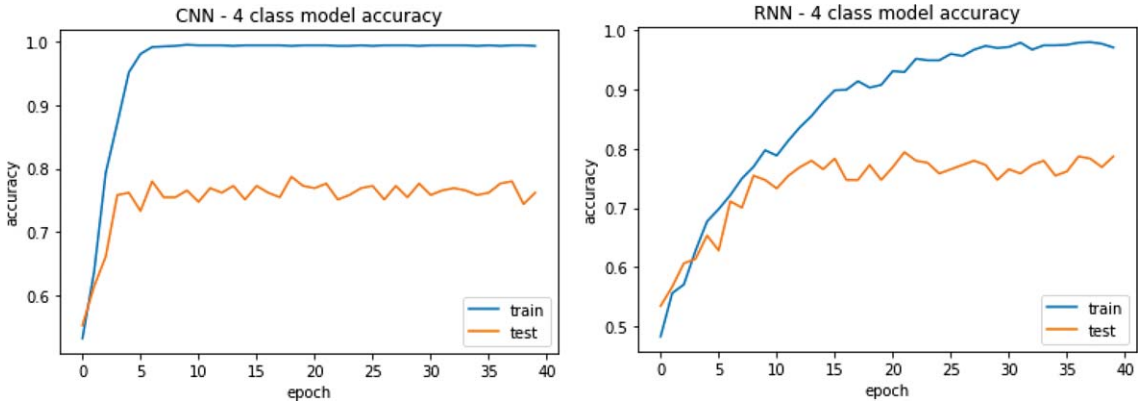

Figure 1: CNN (left) and RNN (right) learning curve for four-class emotion classification.

Table 4: Averaged performance measure (accuracy) for both the CNN and ANN on SYN80 dataset.

	Model	Validation Accuracy	Test Accuracy
(1)	CNN	75.68%	77%
(2)	RNN	73.79%	76%

Both the CNN and RNN networks were created using the *Keras* package in Python. To summarize, this study conducts a series of experiments based on the steps outlined below:

1. All the textual instances were cleaned/preprocessed as outlined in section 3.3.1. Using the shuffle functionality, we randomized the instances within the SYN80 dataset. The resultant dataset was then split into the ratio of 80:20; 10-fold cross-validation was performed on the 80% (training) dataset.
2. The training dataset instances were transformed into a feature vector and the embedding matrix was constructed using the steps outlined in section 3.3.1.
3. Both the CNN and RNN network was trained using the training dataset, and their performance across 10-fold cross-validation was recorded as outlined in section 3.3.2.
4. Finally, we test our models using the test dataset and report the performance measures (see section 3.3.3) across the validation and test dataset averaged over 10 different experiments as discussed in section 3.3.2.

In the next section, we present the results and discussions from this study.

4. Result and Discussion

A four-class emotion classification was performed using both the CNN and RNN. we noticed that the CNN model's test accuracy tends to be stable after 10 iterations (see Figure 1, left). Therefore, we choose to set 10 epochs for the CNN model. The RNN model established the test accuracy after 15 iterations (see Figure 1, right).

The overall average prediction accuracy on validation and test dataset for both the CNN and RNN are reported in Table 4. CNN reported an average of 75.68% validation accuracy slightly better than the RNN classifier (see Table 4). The CNN reported an average of 77% accuracy for the four-class emotion classification in the test dataset. Based on the validation and test accuracy for both the CNN and RNN, we can conclude that there is no evidence of overfitting. It is also clearly evident that both the classifiers performed significantly better in the test dataset than in the validation dataset (see Table 4). A total of 100 iterations of training was performed to train the CNN models. However, it took 150 iterations to train the RNN models. It is also important to note that the time taken to train the RNN models was significantly larger than the time taken to train the CNN models (see Figure 1).

Srinivasan and Ramesh (2018) have reported a baseline performance of the supervised classifiers including k-nearest neighbor (kNN; k = 1, 3, 5, 7), J48 Classifier (C4.5), Classification and Regression Trees (CART), Naïve Bayes Multinomial (NBM), Random Forest (RF), and Sequential Minimal Optimization (SMO) on the SYN80 dataset for a four-class emotion classification problem. The test accuracies on the SYN80 dataset without

Table 5: Averaged performance measure (precision, recall and F1-score) for CNN and RNN on the validation dataset.

	Model	Class	Class Name	Precision	Recall	F1-Score
(1)	CNN	0	Anger	0.806	0.694	0.739
		1	Sadness	0.661	0.776	0.712
		2	Happy	0.849	0.856	0.85
		3	Surprise	0.654	0.528	0.578
(2)	RNN	0	Anger	0.717	0.7	0.704
		1	Sadness	0.679	0.748	0.703
		2	Happy	0.832	0.87	0.851
		3	Surprise	0.652	0.432	0.51

Table 6: Averaged performance measure for CNN and RNN on the test dataset.

				Predicted				
	Model	Class Name	Actual	0	1	2	3	Class Accuracy
(1)	CNN	Anger	0	22.8	4.1	2.8	3.3	0.9426
		Sadness	1	2	42.7	7.4	2.9	0.8747
		Happy	2	1.1	13.5	125.9	6.5	0.8422
		Surprise	3	2.6	4.8	12.4	22.2	0.8827
(2)	RNN	Anger	0	23	3.5	5.3	1.2	0.9303
		Sadness	1	3.2	41.1	7.4	3.3	0.8755
		Happy	2	0.9	12	128	6.1	0.8368
		Surprise	3	5.2	5.1	13.5	18.2	0.8758

resampling were 61.61%, 55.39%, 53.51%, 53.58%, 55.24%, 68.33%, 74.48%, 73.61%, and 76.72% for the classifiers 1NN, 3NN, 5NN, 7NN, C4.5, CART, NBM, RF, and SMO, respectively (Socher *et al.* 2013). The test accuracies for CNN (77%) and RNN (76%) obtained in this study are comparable to the performance of the SMO classifier. When compared against traditional classifiers such as kNN, C4.5, CART, NBM, and RF the performance of the CNN and RNN are slightly superior. Here it is important to note that the preprocessing step did not involve resampling the SYN80 dataset.

In the validation dataset, both the CNN and RNN demonstrated a high F1-score for the *happy* emotion. Both CNN and RNN demonstrated a significantly better average F1-score (above 70%) for the *anger* emotion with CNN significantly outperforming the RNN classifier. However, both the classifier recorded a significantly low F1-score for the surprise emotion compared to the other emotion classes with CNN significantly outperforming the RNN (see Table 5). This is a very striking observation as both the emotion classes, i.e., *anger* and *surprise*, had significantly smaller number of instances in the *SYN80* dataset (see Table 1). To best reason this observation, we believe that the instances belonging to the *anger* class have a rich set of features that can discriminate these instances with respect to the other classes. For the *surprise* class, the RRN classifier resulted in a higher number of false negatives than CNN, which suggests that the features within the instances belonging to the *surprise* class are not clearly discriminative (see Table 5).

The average class accuracy across each emotion class over 10 different models using a single test dataset is reported in Table 6. We also report the averaged confusion matrix of CNN and RNN on the single test dataset over 10 different models. In average 32% of the instances (13.5 out of 42) belonging to the *surprise* emotion has been misclassified as *happy* by the RNN classifier (see Table 6). On the other hand, CNN also misclassified 29.5% of the *surprise* instances (12.4 out of 42) on an average to *happy*. This observation on the test dataset is very consistent with the observations, i.e., low recall and low precision, in the validation dataset (see Tables 5 and 6).

5. Conclusion

In this study, we have demonstrated the deep neural networks' potentiality for the four-class emotion classification problem. Both the CNN and RNN classifiers were explored in this study. On the SYN80 dataset, we achieved an average of 77% accuracy while trying to classify 277 instances belonging to four different emotion types. CNN

slightly outperformed RNN in classifying all the four emotion types. However, both the classifiers misclassified a significant number of instances from the surprise class to the happy class. This observation can be attributed to the fact that both the emotion types, i.e., surprise and happy, are very similar to each other, which is also evident from the circumflex model, and the instances belonging to the surprise type clearly lacks distinctive features that can help the models to discriminate this emotion against the happy emotion. In this study, we have limited our investigations to only four classes of emotions. In the future, we plan to consider additional emotion types and investigate the potentiality of the deep neural networks for emotion classification.

References

Abdul-Mageed, Muhammad, and Lyle Ungar. 2017. "Emonet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 718–28.

Acharya, U., Shu Lih Oh Rajendra, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. 2018. "Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals." *Computers in Biology and Medicine* **100**: 270–78. doi: 10.1016/j.compbiomed.2017.09.017

Alm, Ebba Cecilia Ovesdotter. 2008. *Affect in Text and Speech*. Urbana*:* University of Illinois at Urbana-Champaign, ProQuest Dissertations Publishing, 1–119.

Al Machot, F., A. Elmachot, M. Ali, E. Al Machot, and K. Kyamakya. 2019. "A Deep-Learning Model for Subject-Independent Human Emotion Recognition Using Electrodermal Activity Sensors." *Sensors* **19**, no. 7: 1659, 1–14. doi: 10.3390/s19071659

Aman, Saima, and Stan Szpakowicz. 2007. "Identifying Expressions of Emotion in Text." In: Matoušek V., Mautner P. (eds) *Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg 4629: 196–205.

Badshah, Abdul Malik, Jamil Ahmad, Mi Young Lee, and Sung Wook Baik. (2016) "Divide-and-conquer based ensemble to spot emotions in speech using MFCC and random forest." In *the proceedings of the 2nd international integrated conference & concert on convergence*, 1-8. *arXiv*:1610.01382.

Bhowmick, Plaban Kumar, Anupam Basu, and Pabitra Mitra. 2010. "Classifying Emotion in News Sentences: When Machine Classification Meets Human Classification." *International Journal on Computer Science and Engineering* **2**, no. 1: 98–108.

Chaffar, Soumaya, and Diana Inkpen. 2011. "Using a heterogeneous dataset for emotion analysis in text." in *Canadian conference on artificial intelligence*, LNAI 6657, 62–67.

Chatterjee, Ankush, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. "Understanding Emotions in Text Using Deep Learning and Big Data." *Computers in Human Behavior* **93**: 309–17. doi: 10.1016/j.chb.2018.12.029

Coviello, Lorenzo, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. "Detecting Emotional Contagion in Massive Social Networks." *PLoS One* **9**, no. 3: e90315, 1–6. doi: 10.1371/journal.pone.0090315

Diman, Ghazi, Diana Inkpen, and Stan Szpakowicz. 2010. "Hierarchical versus Flat Classification of Emotions in Text." *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 140–6.

Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz. 2010. "Hierarchical Versus Flat Classification of Emotions in Text." *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, **1**(1), 140–6.

Ghosal, D., N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. 2019. "Dialoguegcn: A Graph Convolutional Neural Network for Emotion Recognition in Conversation." *arXiv:1908.11540*. doi: 10.48550/ARXIV.1908.11540, 1–11.

Hamdi, Eman, Sherine Rady, and Mostafa Aref. 2020. "A Deep Learning Architecture with Word Embeddings to Classify Sentiment in Twitter." *International Conference on Advanced Intelligent Systems and Informatics*, 115–25.

Hasan, Maryam, Elke Rundensteiner, and Emmanuel Agu. 2014. "Emotex: Detecting Emotions in Twitter Messages." *ASE Bigdata/SocialCom/Cybersecurity Conference*, Stanford University, 1–10.

Irsoy, Ozan, and Claire Cardie. 2014. "Deep Recursive Neural Networks for Compositionality in Language." *Advances in Neural Information Processing Systems* **27**: 2096–104.

Jabreel, Mohammed, and Antonio Moreno. 2019. "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets." *Applied Sciences* **9**, no. 6: 1123, 1–16. doi: 10.3390/app9061123

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. 2014. "A Convolutional Neural Network for Modelling Sentences." *arXiv*:1404.2188v1. doi: 10.48550/ARXIV.1404.2188, 1–11.

Khorrami, P., T. Le Paine, K. Brady, C. Dagli, and T. S. Huang. 2016. "How Deep Neural Networks Can Improve Emotion Recognition on Video Data." *2016 IEEE International Conference on Image Processing (ICIP)*, 619–23. IEEE.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." *arXiv*:1408.5882v2. doi: 10.48550/ARXIV.1408.5882, 1–6

Kratzwald, Bernhard, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. "Deep Learning for Affective Computing: Text-Based Emotion Recognition in Decision Support." *Decision Support Systems* **115**: 24–35. doi: 10.1016/j.dss.2018.09.002

Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao. 2015. "Recurrent Convolutional Neural Networks for Text Classification." *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, **1**(1), 2267–2273.

Lakomkin, E., C. Bothe, and S. Wermter. 2018. "GradAscent at EmoInt-2017: Character-and Word-Level Recurrent Neural Network Models for Tweet Emotion Intensity Detection." *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, *arXiv:1803.11509*. doi: 10.18653/v1/W17-5222"10.18653/v1/W17-5222, 169–74.

Ling, Rich, and Naomi S. Baron. 2007. "Text Messaging and IM: Linguistic Comparison of American College Data." *Journal of Language and Social Psychology* **26**, no. 3: 291–98. doi: 10.1177/0261927X06303480

Liu, Bing. 2020. "Text Sentiment Analysis Based on CBOW Model and Deep Learning in Big Data Environment." *Journal of Ambient Intelligence and Humanized Computing* **11**, no. 2: 451–8. doi: 10.1007/s12652-018-1095-6

Pang, Bo, and Lillian Lee. 2005. "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales." *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05) arXiv:cs/0506075*. doi: 10.3115/1219840.1219855, 115–24.

Pennington, Jeffrey, Richard Socher, and ChristopherD. Manning. 2014. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* **39**, no. 6: 1161–78. doi: 10.1037/h0077714

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–42.

Srinivasan, Satish M., and Prashanth Ramesh. 2018. "Comparing Different Classifiers and Feature Selection Techniques for Emotion Classification." *International Journal of Society Systems Science* **10**, no. 4: 259–84. doi: 10.1504/IJSSS.2018.095595

Wang, Jin, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 225–30.

Zhang, Rui, Honglak Lee, and Dragomir Radev. 2016. "Dependency Sensitive Convolutional Neural Networks for Modeling Sentences and Documents." *Proceedings of the 2016 Conference of the North America Chapter of the Association for Computational Linguistics: Human Language Technologies. arXiv:1611.02361*. doi: 10.18653/v1/N16-1177"10.18653/v1/N16-1177, 1512–21.

Zhang, Yangsen, Jia Zheng, Yuru Jiang, Gaijuan Huang, and Ruoyu Chen. 2019. "A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model." *Chinese Journal of Electronics* **28**, no. 1: 120–26. doi: 10.1049/cje.2018.11.004

Zheng, Wei-Long, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. 2014. "EEG-Based Emotion Classification Using Deep Belief Networks." *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Zhou, Peng, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. "Text Classification Improved by Integrating Bidirectional LSTM with Two-Dimensional Max Pooling." *arXiv:1611.06639v1*. doi: 10.48550/ARXIV.1611.06639, 1–11.

Zhou, Chunting, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. "A C-LSTM Neural Network for Text Classification." *arXiv:1511.08630*. doi: 10.48550/ARXIV.1511.08630, 1–10.

WWW.JBDTP.ORG

ISSN: 2692-7977

JBDTP Professional Vol. 1, No. 1, 2022 DOI: 10.54116/jbdtp.v1i1.16

STRATEGIES FOR DEMOCRATIZATION OF SUPERCOMPUTING: AVAILABILITY, ACCESSIBILITY, AND USABILITY OF HIGH PERFORMANCE COMPUTING FOR EDUCATION AND PRACTICE OF BIG DATA ANALYTICS

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Margaret Brennan-Tonetta
Rutgers University
mbrennan@njaes.rutgers.edu

Yana Samuel
Northeastern University
samuel.y@northeastern.edu

Pradeep Subedi
University of Utah
pradeep.subedi@utah.edu

Jack Smith
Marshall University
smith1106@marshall.edu

ABSTRACT

There has been an increasing interest in and growing need for high performance computing (HPC), popularly known as supercomputing, in domains such as textual analytics, business domains analytics, forecasting, and natural language processing (NLP), in addition to the relatively mature supercomputing domains of quantum physics and biology. HPC has been widely used in computer science (CS) and other traditionally computation intensive disciplines but has remained largely siloed away from the vast array of social, behavioral, business, and economics disciplines. However, with ubiquitous big data, there is a compelling need to make HPC technologically and economically accessible, easy to use, and operationally democratized. Therefore, this research focuses on making two key contributions, the first is the articulation of strategies based on availability, accessibility, and usability (AAU) concepts for the demystification and democratization of HPC, based on an analytical review of Caliburn, a notable supercomputer at its inception. The second contribution is a set of principles for HPC adoption based on an experiential narrative of HPC usage for textual analytics and NLP of social media data from a first-time user perspective. Both, the HPC usage process and the output of the early-stage analytics are summarized. This research study synthesizes expert input on HPC democratization strategies and chronicles the challenges and opportunities from a multidisciplinary perspective, of a case of rapid adoption of supercomputing for textual analytics and NLP. Deductive logic is used to identify strategies which can lead to efficacious engagement, adoption, production, and sustained usage for research, teaching, application, and innovation by researchers, faculty, professionals, and students across a broad range of disciplines.

Keywords *High performance computing, Education, Technology access, Supercomputing, Democratization, Big data, Artificial intelligence, Textual analytics, NLP*

1. Introduction

Because HPC stands at the forefront of scientific discovery and commercial innovation, it is positioned at the frontier of competition—for nations and their enterprises alike…
(Ezell and Atkinson 2016)

Big data and artificial intelligences (AIs) are having a significant impact on business, work, governance, social interaction, and education. While big data and AIs hold tremendous potential for value creation and transformation of human life, their potential can only be realized through appropriate technological implementations. Specifically, significantly more powerful and scaled up data processing, networking, and data storage capabilities are required to harness the vast promise of AIs and big data. This necessitates the usage of technologies, which are popularly termed as "supercomputing," and also as "high performance computing" (HPC), referring to the use of supercomputers or high performance computers for computing complex, voluminous, or iteration intensive calculations and analytics. The term "supercomputing" (or supercomputers) has been treated as being synonymous with HPC (or high performance computers), and has been parsimoniously described as a computer or a cluster of computers with far greater computing-memory-storage capabilities than a general computer, and as being "characterized by large amounts of memory and processing power" (George 2020). So also, HPC has been varying defined as being a "combination of processing capability and storage capacity" that can efficiently create solutions for "difficult computational problems across a diverse range of scientific, engineering, and business fields" (Ezell and Atkinson 2016), and also as being "massively parallel processing (MPP) computers" (Bergman *et al.* 2019). HPC can be classified as being homogeneous or heterogeneous, based on the use of similar or dissimilar processors (or memory, or similar HPC components) respectively, in its array of processors, such as homogeneous HPC with CPU arrays, and heterogeneous HPC with CPU and GPU arrays (Gao and Zhang 2016). Heterogeneous HPC can be used to improve effectiveness, speed, and also to gain additional energy savings. High performance computers can therefore be viewed as organized systems of high-powered, parallel structured computational capabilities, including extreme and diverse processing capabilities, general or task varied and scalable memory, scalable storage, grid or network or cloud based, and appropriate capabilities management interfaces with the potential to help solve vast and complex problems.

1.1 The Critical Need: Multidisciplinary HPC Applications

The compelling need for fostering HPC and HPC education has been well recognized by industry, government, and academia. However, most of these efforts have been largely isolated streaks, albeit with some progression, to a restricted set of traditionally computational domains. Given the explosive growth in quantity, diversity, complexity, granularity, and acceleration of data generation, it has become impossible to meaningfully depend on desktops, servers or standalone computers to create competitive value, and an increasingly large number of disciplines have begun HPC evaluation and adoption processes, to ensure that they remain competitive in progressively data and computation intensive environments (Fiore *et al.* 2018). The already steep trend towards HPC engagement and adoption can be expected to become stronger with the advent of new technologies and the identification of new opportunities (Bergman *et al.* 2019). An investigation by the Council on Competitiveness discovered that the vast majority of United States corporations with HPC capabilities had significant concerns about being able to hire persons with "sufficient HPC training," and that there are no easy solutions because "… there aren't enough faculty, researchers, educators, and professionals with the HPC skills and knowledge to fulfill the demand for talented individuals" (Lathrop 2016). There has been a sustained call over the past decade for opening up access to HPC/SC resources: "In the past decade high performance computing has transformed the practice and the productivity of science. Now this analytical power must be opened up to industry, to improve decision making, spur innovation and boost competitiveness" (Moran and O'Dea 2013). HPC has been treated as a critical catalyst for "inter- and trans-discipline breakthroughs" impacting the development of science and innovation globally (Mosin 2017). As illustrations, extant research has called for machine learning applications on varying types of information across a wide range of domains, include behavioral finance, social media analytics, textual data visualization and pandemic sentiment analysis, all of which are best implemented at scale using HPC resources (Samuel 2017; Samuel *et al.* 2018; Conner *et al.* 2019, 2020; Rahman *et al.* 2020). HPC is now a global phenomenon, and competitive advantage in many domains is associated with multidisciplinary HPC capabilities.

1.2 The Future of HPC Impact: Pervasive and Ubiquitous

Big data-driven AI holds the keys to global value creation (Samuel 2021; Samuel *et al.* 2022). Given the rapid growth of AI and the associated need to process big data, HPC has become one of the critical drivers of success for

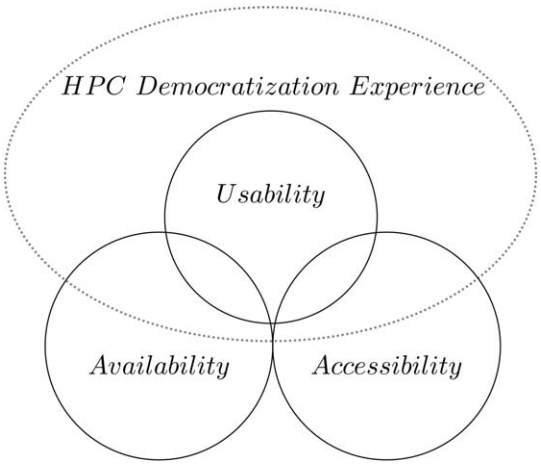

Figure 1: The HPC democratization concept.

research institutions, corporations, and nations. Ubiquitous AI and big data applications imply an equally expansive HPC requirement and influence. HPC ubiquity is a certainty, although end users will most likely not be required to interface with the technological complexity of HPC systems, just as electricity and internet are ubiquitous, without end users having to interface with systems managing electricity generation equipment or the movement of data packages and transmission protocols for the internet. Institutions, corporations, and countries which enable their constituents and stakeholders with easily usable HPC capabilities will possess a significant competitive advantage. HPC-driven competitive advantages will impact individuals, businesses, society, and nations, bearing the potential for significant socioeconomic impact. It is therefore of paramount importance to look closely at strategies for catalyzing future HPC thought leadership and competitiveness.

1.3 The Need: Democratized Supercomputing!

Democratized supercomputing, in the context of this study, refers to the opening up of HPC resources, freeing it from restrictive domain boundaries, and making it seamlessly available to all on an as-needed basis. Presently, even where HPC is available, it still remains inaccessible to many. Furthermore, even where access is provided, usability is restricted due to operational and skills barriers. Our quasi-phenomenological perspective based on the motivations driving the conceptualization, development, and deployment of Caliburn supercomputer at Rutgers, the State University of New Jersey, and an analysis of HPC usage for textual analytics and natural language processing (NLP) indicate that there are three key strategies that are necessary for democratizing HPC across domains: 1) availability strategy, HPC resources need to be built and distributed to ensure fair availability; 2) accessibility strategy, just the mere fact that HPC infrastructure exists in an institution or at a location does not ensure its accessibility, and therefore deliberate steps need to be taken to align and distribute available HPC resources in a manner so as to ensure HPC accessibility; and 3) usability strategy, availability and accessibility ensure that end users are empowered to access HPC resources; and yet these alone do not democratize or catalyze HPC utilization without the necessary dimension of ease-of-use. An effective HPC democratization initiative must include the three strategies of HPC Availability, HPC Accessibility, and HPC Usability, to ensure that capabilities are developed to achieve a maximized spectrum of benefits from HPC (Figure 1).

The rest of this paper is organized as follows. First, the study clarifies the multidisciplinary context, provides theoretical lenses from Information Systems (IS), and anchors HPC democratization discussion to the theories of technology adoption and usage. This is followed by an analytical and reflective narrative of the motivations and process for the acquisition and deployment of Caliburn, a supercomputer at Rutgers University. Availability, Accessibility, and Usability (AAU) strategies are then elaborated upon in subsequent sections. This is followed by a case analysis of HPC usage for NLP and key principles for sustained usage. The paper concludes with notes on implications, limitations, and a motivational conclusion.

2. HPC Engagement, Adoption, and Sustained Usage. Theoretical and Applied Considerations

2.1 The Future of HPC Relevance: Ubiquitous, Multidisciplinary, Transdisciplinary, and Interdisciplinary

Extant research meaningfully distinguishes between" multidisciplinarity," "interdisciplinarity," and "transdisciplinarity," wherein "multidisciplinarity draws on knowledge from different disciplines but stays within their boundaries," while interdisciplinarity "analyzes, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole" and transdisciplinarity "integrates the natural, social and health sciences in a humanities context, and transcends their traditional boundaries" (Choi and Pak 2006; Alvargonzález 2011). We believe that such distinction is valuable. However, since this study does not delve into the nature of disciplinary research, but rather emphasizes the need for HPC to be used across disciplines, simultaneously drawing on knowledge and integrating lessons learned from the past, irrespective of discipline, hence we employ the word "multidisciplinary" in its broadest sense, inclusive of the properties of interdisciplinarity and transdisciplinarity.

2.2 HPC Engagement: Theoretical Basis

As with all technologies, there are critical drivers for HPC engagement, adoption and sustained usage. IS research studies have provided extensive insights into user engagement, adoption and sustained usage of technologies. The seminal, and in many senses foundation setting, technology acceptance model (TAM) theory validated and popularized the concepts of perceived usefulness and ease of use of technologies (Davis 1989). Subsequent studies updated TAM, including a broader theoretical basis with a unified perspective leading to the "unified theory of acceptance and use of technology" (UTAUT), which "highlights the importance of contextual analysis in developing strategies for technology implementation" (Venkatesh *et al.* 2003). Extant research has also demonstrated that the challenges of technology adoption and usage are subject to information facets, information complexity, equivocality of information, and information overload (Samuel 2016; Samuel and Pelaez 2017). Some technologies, such as blockchains, are pertinent to specific user categories, where generally dominant drivers of technology acceptance may have less relevance than factors such as "security, privacy, transparency, trust and traceability aspects" (Grover *et al.* 2019).

2.3 HPC Usage: From Theory to Practice

The generic usability and acceptance model (GUAM), in contrast to UTAUT, provides a significantly better explanation of "behavioral intention (72%) and technology use (63%)" for learning innovations (Obienu and Amadin 2021). This demonstrated that domain or discipline sensitive models have the potential to outperform generic adoption models like TAM or UTAT, due to variations on technological features and characteristics of user groups. Factors such as gender and social characteristics have also been shown to influence user engagement with technologies, such as the indication by prior research that women tend to weigh ease of use more strongly than men who tend to focus on perceived usefulness of the technology (Venkatesh and Morris 2000). Supportive technologies, such as anthropomorphic chatbots with human-like NLP and communication capabilities, can have a meaningfully positive impact on user perception of the usefulness of technology, and such supportive mechanisms can also support perceived ease of use (Rietz *et al.* 2019). Additional theories need to be evaluated to maximize the theoretical basis for HPC and supercomputing engagement. For example, flow theory which refers to the "the holistic sensation that people feel when they act with total involvement" and the state of "flow" where people experience becoming "absorbed in their activity," akin to Chess players and gamers whose intelligences are fully engaged and focused (Csikszentmihalyi and Csikszentmihalyi 1992; Koufaris 2002; Csikszentmihalyi 2014). On the applied side, numerous AI research initiatives highlight the need for HPC in advancing research in multiple areas, such as natural language generation in the context of social media analytics (Garvey *et al.* 2021; Samuel *et al.* 2021). The present study presents an analysis of a first time users' experience with HPC for textual analytics and NLP, to identify some useful principles that will help potential HPC users to move from theory and strategy to applications and practice.

3. Caliburn, a Story of Strategic Value Creation through ACI AAU

The story of Caliburn, the first supercomputer at Rutgers University and in the state of New Jersey, begins in 2011 with the creation of the Rutgers Discovery Informatics Institute (RDI^2) by Dr. Manish Parashar, Distinguished Professor, Computer Science. His motivation was to establish a comprehensive and internationally competitive multidisciplinary Computational and Data-enabled Science and Engineering (CDS&E) institute at Rutgers University

that could catalyze and nurture the integration of research and education with advanced computing infrastructure (ACI). Parashar structured RDI^2 to provide ACI resources that were available, accessible and usable by offering technologies and expertise to academic researchers and companies that want to take advantage of ACI resources, but do not have the financial resources or expertise necessary to acquire these human and hardware resources. It was his vision to have a national level ACI at Rutgers University available to all.

3.1 Sensing the Need

The importance and immediacy of having such ACI competency at Rutgers was further accentuated by the growing role of computation and data in all areas of science, engineering, and business, as well as current and future trends in ACI. These included disruptive hardware trends, ever-increasing data volumes, complex application structures and behaviors, and new first-order concerns such as fault-tolerance and energy efficiency. These trends are a result of the continued quest towards extreme scales in computing and data that is necessary to drive innovations in science, engineering, and other data intensive fields.

3.2 Innovation

The CI developed by RDI^2 is innovative and provides researchers with global linkages to the national and international CI (e.g., XSEDE, OSG, OOI, LHC, iPlant, PRACE, EGI, etc.) that connects Rutgers with observational instruments, data streams, experimental tools, simulation systems, and individuals distributed across the globe. Overall, the impact of RDI^2 was a revolutionary advance in the scale and effectiveness of science and engineering research conducted at Rutgers and by academia and industry throughout the state.

3.3 Strategic Vision

As a next step, it was critical that Rutgers develop and implement a bold strategic vision for an ACI ecosystem that was competitive at the national and international levels (Berman *et al.* 2013). This ecosystem had to provide researchers with cutting-edge computing and data handling capabilities, and students with necessary ACI exposure and training. In 2013, RDI^2 initiated a university-wide ACI strategic planning process with input from faculty across many disciplines at Rutgers. This resulted in a comprehensive plan, "Accelerating Innovation Through Advanced Cyberinfrastructure: A Strategic Vision for Research Cyberinfrastructure at Rutgers" (Berman *et al.* 2014). The plan called for strategic investment in ACI to drive innovation, improve research capabilities and productivity, and enhance faculty competitiveness. Two specific findings of the Rutgers ACI strategic plan were the need to deploy a nationally competitive advanced cyberinfrastructure and to establish a central Office of Advanced Research Computing at the University.

3.4 Availability

Deploying a nationally competitive ACI required infrastructure investments in computing, mass storage, and high-speed/bandwidth digital communication that could provide state-of-the-art capacities and capabilities for Rutgers researchers and offer a competitive advantage among Big Ten peer institutions and beyond. However, the level of investment required to achieve this goal was significant. Fortunately, in 2013, the State of New Jersey announced the Higher Education Equipment Leasing Fund to support investments in cutting-edge equipment at the state's higher education institutions. RDI^2 submitted a proposal entitled "Rutgers University Advanced Compute & Data Cloud" to establish a statewide ACI resource at Rutgers that could have far-reaching benefits for higher education institutions, industry, and state government. The state recognized the tremendous impact that this capability could have for its innovation economy, and the proposal was approved. RDI^2 was awarded $10 million to purchase cutting edge ACI systems, the largest award given through this program.

3.5 Accessibility

After two years of design and installation, Caliburn and companion system ELF were deployed in 2016 at Rutgers. The rationale for the two systems was the recognition of the potential limitations for many researchers who were inexperienced in a complex ACI such as Caliburn, or do not need its high-level capabilities. Thus, ELF, which had significantly more computing capability than what was currently available but not as complex or powerful as Caliburn, could be used by researchers as a first step in building experience and understanding of more sophisticated ACI. These systems provide a balanced advanced computational and data environment that contains a large-scale high-end compute engine, as well as significant co-located storage with embedded analytics capabilities. The Caliburn system was designed by SuperMicro in collaboration with the RDI^2 technical team. The latest in energy

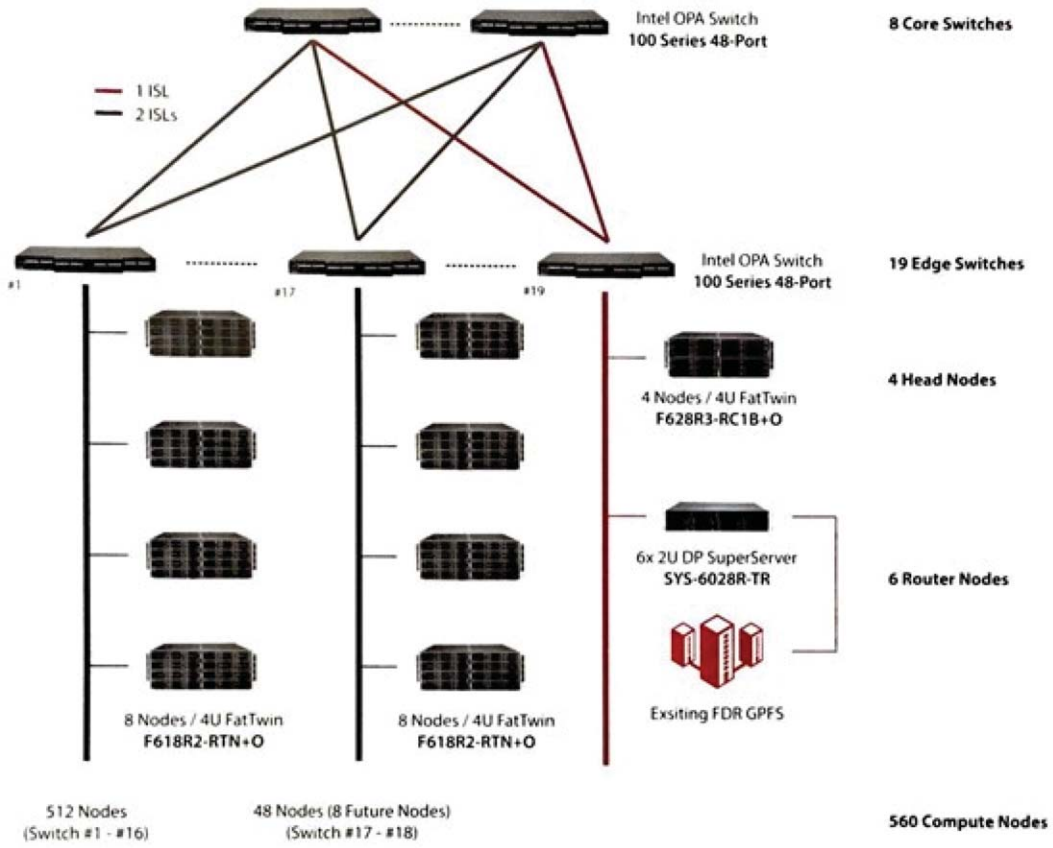

Figure 2: Overview of Caliburn architecture.

Table 1: Summary of Caliburn technical specifications

Caliburn	
Compute nodes	704
Cores	23,616
Memory	176 TB
Flash storage	200+ TB

efficient and space saving technologies were utilized, resulting in Caliburn using 90% less energy than standard indoor cooling units and is 80% space efficient. See Figure 2 for a diagram of the Caliburn architecture.

The Caliburn and ELF platforms are seamlessly accessible as a cloud service, providing researchers, students, industry and government across the state, with on-demand and pervasive access to these capabilities for research and instruction. Caliburn/ELF are also connected with high-speed networking to key national and international research/educational facilities. The overall platform is unique, and the most powerful academic system in the state. When it was commissioned in Summer 2016, Caliburn was ranked on the Top 500 list of computer systems worldwide as #2 system among the United States Big 10 institutions and #8 among all United States academic institutions, #50 among academic institutions globally, and #166 among all computer systems worldwide (Parashar 2019). Table 1 provides a quick overview for Caliburn's technical specifications.

3.6 Usability

To further have ACI as available, accessible, and usable as possible at the university, RDI^2 began planning the creation of the first central Rutgers Office of Advanced Research Computing (OARC) in 2015. In Spring 2016, the

university's first Associate Vice President for Advanced Research Computing was appointed to continue the growth and development of university-wide ACI. The functions of this office are to provide strategic leadership, coordinate investments in ACI and related expertise, and catalyze and nurture cyberinfrastructure-enabled multidisciplinary research, all aimed at fostering a community of excellence in computing and data, empowering research, learning, and societal engagement and providing a competitive advantage to the Rutgers community and throughout the state and region. In 2019, management of the Caliburn and ELF systems was moved to OARC. These and other ACI resources are broadly available through this office.

3.7 Caliburn: Extending AAU with a Futuristic Vision

While Caliburn has been used for a wide range of multidisciplinary projects, one of the noteworthy initiatives which embodies the availability-accessibility-usability paradigm is the "Caliburn Supercomputing Awards." This initiative provides a pathway for scholars and academics outside of Rutgers to apply for HPC allocations, and thus expands the reach and impact of Caliburn; this is an example of a useful mechanism for the democratization of HPC: scholars and academics from institutions without HPC capabilities are now empowered in their research and thought leadership, which would otherwise be lacking. The process which was established by RDI^2 and continues with OARC, starts with OARC inviting proposals for "the allocation of computing resources on Caliburn," providing "high performance computing capabilities to academic researchers across the state to accelerate research programs that use or develop highly scalable computing applications." OARC also invites applicants in a second "startup" category, and these startup proposals "are provided as means to have full access with a limited time usage allocation" and are encouraged to be structured such that "they can be converted into awarded allocations during the next call for proposal cycle." Applications are limited to academic institutions in New Jersey under this program; this is something that OARC can go beyond, subject to availability of resources after meeting state-level needs, to provide a measure of access to meet the HPC needs of individual residents of New Jersey, nonacademic institutions, and academic outside of the state as well. For example, it would be of great value to New Jersey residents and students, if public libraries in New Jersey were empowered to provide interactive HPC demonstrations and interaction opportunities locally. Another avenue would be to explore relationships with specific institutions outside of the state to foster democratization of HPC resources.

4. The Expanding HPC Landscape: Notable Initiatives and Case Analyses

There are numerous existing efforts which in some form address the AAU paradigm. However, the focus is mostly on facilitating availability, and accessibility to a lesser extent. Usability tends to be left for the end user to wrestle with, with the help of "user guides," often leading to a loss of time and effort. The following section provides a brief overview illustrating some of the prominent HPC efforts.

4.1 HPC across Institutions and Disciplines

Advanced Cyberinfrastructure Research & Education Facilitators (ACI-REF): The Clemson-led ACI-REF program (NSF #1341935) advanced research computing through a network of Cyberinfrastructure (CI) Facilitators. This team is now co-leading Campus Research Computing Consortium (CaRCC). CaRCC is a NSF Research Coordination Network (NSF #1620695) and a follow-on to the ACI-REF project that addresses the huge growth in demand for local research computing, by sharing, collaborating, and developing best practices for research-facing, system-facing, software-facing, and stakeholder-facing CI professionals. CaRCC does not do large-scale workforce development for CI professionals itself, but most CaRCC institutions have now participated in Neeman's Virtual Residency Program. The ACI-REF Virtual Residency Program (VRP) was initiated by Henry Neeman, University of Oklahoma. He was a collaborator on the original ACI-REF proposal, and started the ACI-REF Virtual Residency Program (VRP) with an NSF CC-IEE grant (NSF #1440783) to provide national-scale CI Facilitator training. All the original ACI-REF institutions have participated in the VRP (subsection reference: Neeman *et al.*, 2016, 2018).

4.2 Facilitation of CI-Driven Research

XSEDE Campus Champions (CCs): There are more than 700 CCs, and these numbers are growing, at over 300 United States institutions helping their local researchers use CI, especially large scale/advanced computing. Most CCs perform CI facilitation activities, and CCs usually peer-mentor each other. The Champion community has: (a) a very active mailing list, where CCs exchange ideas and help each other solve problems; (b) regular conference calls for learning what's happening both among CCs and in national CI; and (c) major participation at national conferences like PEARC (e.g., 23% of PEARC'20 attendees were CCs). Many CCs have also participated in the VRP.

The Society of Research Software Engineering (SRSE) has 29 participating universities, supports Research Software Engineers (RSEs), focusing on reproducibility, reusability, and accuracy. The goal is to foster career paths for academic RSEs and ensure that they are recognized and rewarded. The United States Research Software Engineer Association (US-RSE) is the United States counterpart for SRSE. It has over 700 members. The United States Research Software Sustainability Institute (URSSI) has been funded by NSF from 2017 to 2021. Their goal is to design an institute on research software and to build the RSE community, in order to (i) improve how individuals and teams function, and (ii) advance research software and the STEM research it supports. Other efforts include initiatives such as the Supercomputing in Plain English (SIPE) workshop, which is an annual workshop on supercomputing (HPC) at Oklahoma University, run by Henry Neeman, using plain English to introduce fundamental issues of supercomputing as they relate to Computational and Data-enabled Science & Engineering. Internet2 and EDUCAUSE also have programs to help enable CI Facilitators (subsection reference: Neeman *et al.* 2016, 2018).

4.3 NLP Case: Multidisciplinary HPC Education and Productivity

In addition to the analysis and study of institutional level initiatives, it is important to factor in individual perspectives of HPC usage from a multidisciplinary perspective. This subsection summarizes the key conceptual factors and implications for HPC usage from the lens of a first time HPC user for textual analytics (TAn) and NLP on social media data. TAn and NLP have been widely used for social media data analytics and a broad range of natural language sense-making efforts, including research on COVID-19, stock market, and public perception (Kretinin *et al.* 2018; Samuel *et al.* 2020a, 2020b). The narrative is based on a Caliburn allocation award for a TAn and NLP project; this narrative does not focus on the findings of the core research, but rather on the process employed by a new HPC user, and the associated learning curve. The analysis highlights how the Caliburn award served as an excellent example of CI availability and facilitation of capability, but was lacking in sufficient usability support for a non-Computer Science (CS) user.

4.3.1 Caliburn Usage Case: HPC for Making Sense of TAn and NLP

The goals and motivation for HPC engagement were to make sense of a large data file of social media data, consisting of over 7 million records, which needed to cleaned, explored, summarized, and analyzed for general and public sentiment insights. The analysis was initiated using R and Python, and associated software packages and libraries. This task, which was initiated in 2019, was beyond the capabilities of a high-powered desktop with 64 GB of RAM and mandated the use of HPC resources. A Caliburn HPC allocation award made this analysis possible from 2020, and the second phase of the project continues into 2021. The sentiment analysis and custom advanced data visualization methods used for analysis necessitated the installation of new packages and libraries on the HPC system.

4.3.2 Navigating Caliburn: Initiation

The HPC engagement process involved remote access of Caliburn with a two-factor authentication process. Rutgers OARC had a very clear step-wise process for this, and the initial access process was smooth and efficient, thus indicating that the availability of CI HPC resources were well supported by an efficient Accessibility strategy. The challenges occurred on the usability level, and although issues such as data transfer were self-resolved by the researcher's own efforts, issues with running necessary software remained. The major challenges faced were on key dimensions of usability: (a) interface, the command line drive interface led to a long learning curve, and this could be mitigated by the use of open-source solution such as OpenOndemand; (b) software installation, while Caliburn had existing tools, it was a laborious and iterative process to figure out optimal ways to install all required R packages and Python libraries; and (c) exporting and saving the analysis, especially the data visualizations in the required format. A qualitative estimate indicates that about 75% of man-hours invested into the analysis were spent of resolving usability issues; this indicates a significant challenge for new users and non-CS users of HPC. Two kinds of HPC engagement were utilized: running live jobs (smaller subsets of data, relatively low computational requirement) and running batch jobs (the intended purpose of HPC, using larger datasets, with higher computational requirements). In running batch jobs, an additional issue arose: development of standalone script for running complete analytical processes; this requires a new mindset as compared to live-interactive analytics processes and is described in further detail in the Caliburn usage process subsection below.

4.3.3 Caliburn Usage Process: Tactical Summary

In our case, we utilized two kinds of HPC engagement: running live jobs and running batch jobs. Running live jobs involved "asking" Caliburn for access to a HPC node, where once access was provided, it was like using a Linux

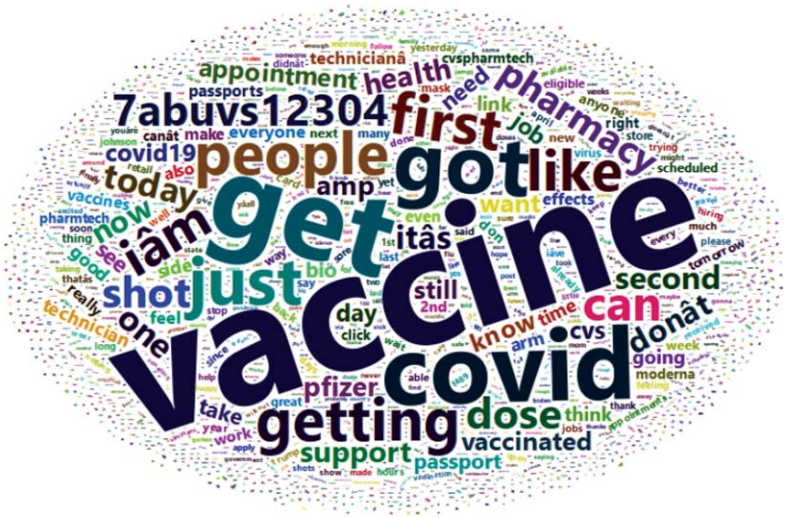

Figure 3: WordCloud on COVID-19 vaccine data.

machine via terminal. This use case is particularly suitable for analyses that require frequent user-intervention and employs smaller subsets of data with relatively low computational requirements. In contrast to live jobs, running batch jobs can be further divided into small scale jobs and large-scale jobs. Both batch cases require the development of a standalone script for running complete end-to-end analytical processes; this requires a new mindset as compared to live-interactive analytics processes. While the small-scale batch job is similar to running a live job from the perspective of storage and computational requirement, the benefit is that once the script is written, it does not require constant monitoring and interaction. Large-scale batch jobs are mainly designed for larger datasets with higher computational requirements, and it usually engages multiple nodes of the HPC machine. However, these large-scale batch jobs come with a steep learning curve on the efficient use of multiple nodes and cores of a HPC machine, and on end-to-end output-inclusive scripts that are required to run parallel analytics processes. Caliburn was powerful enough to process a few million tweets fairly quickly. Large scale social media data analysis has become a powerful research paradigm, and has been used for a broad range of analyses, including public sentiment analytics (Ali *et al.* 2021). The processed data was then used for a variety of topical analyses on Caliburn and on local machines, for running exploratory, predictive and machine learning models, and visualizations, including word-clouds on COVID-19 and vaccine data, as illustrated in Figure 3.

4.3.4 Caliburn Usage Case: Deductive Implications

The above described HPC usage analysis for TAn and NLP provided interesting principles and insights for HPC adoption. 1) Effective HPC Availability and Accessibility strategies led to the use of Caliburn for business analytics purposes, employing TAn and NLP methods. This research would not have been possible with the Rutgers-OARC's Caliburn award. 2) Usability strategies were limited to benefit expert HPC users and were not ready to support non-CS and new HPC users. 3) OARC support staff were well trained and provided critical assistance in addressing the barriers and issues with data transfers, software installation, and script required to enable TAn and NLP tools; this mitigated usability issues but still led to a significant increase in man-hours used for the analysis. 4) The absence of a GUI (graphic user interface) and user-friendly software installation process caused delays and discouraged additional creative research efforts, thus limiting the value created through for core research objectives. And 5) a successful HPC Usability strategy must implement appropriate user education and critical changes to the HPC process and system, to enable smooth and high impact multidisciplinary research.

4.4 HPC Education and Productivity Principles

User education is critical for the success of any HPC Usability strategy. Education in HPC is an actively developing field with specialists from various disciplines participating in diverse initiatives. The development of specialized and purpose specific educational programs is therefore critical for meeting the ever-increasing demand for skilled HPC users (Connor *et al.* 2016). It has been shown that addressing the skill set gap is critical for meeting the needs in CI research workforce as well and interdisciplinary teaming helps foster the learning process in technological

domains (Choi and Kim 2017). Extant research has shown that interdisciplinary faculty are essential for successful implementation of HPC instruction (Neumann *et al.* 2017). Globally, emphasis on hands-on experiences and communications with international faculty is gaining significant prominence in HPC education (Sancho 2016). Several researchers call for a "holistic approach" to HPC training and education rather than focusing on a particular HPC ecosystem (Chaudhury *et al.* 2018). CI productivity can be maximized with an effective implementation of HPC Usability strategy.

5. HPC Democratization Strategies: AAU

This leads us to summarizing a key contribution of this study: the articulation of AAU strategies for the demystification and democratization of HPC.

5.1 Availability

This has been the first step in expanding HPC usage, and institutions across the nation and globally have been at the forefront of acquiring and developing supercomputing capabilities. An effective availability strategy consists of acquiring and developing CI, such that it caters to stakeholder needs for the current phase, while being scalable to accommodate larger workloads, and flexible to be developed for diverse workloads. In the age of cloud computing, needless to say, availability is not restricted by geography but bounded by network and access protocols. Once capacities were developed across many institutions, it was observed that underutilization was a problem due to accessibility issues. Multiples measures exist for the scale and scope of CI, which are essentially a summary of HPC technological components such as processors, memory, storage, and structure. Sufficient Availability is critical for HPC democratization. However, it is a costly error to believe that Availability will lead to maximized usage and optimal output.

5.2 Accessibility

While some CI is built to cater to a very limited and specialized group of stakeholders, the Accessibility strategy is defined in reference to HPC capabilities at public organizations and academic institutions, where there is a need to cater to a broader need for HPC resources, as well as a responsibility to maximize the investment dollars. An effective Accessibility strategy consists of frameworks and processes that maximize the engagement of multidisciplinary stakeholders, plan for expanded user categories with prioritization of core stakeholders, such that underutilization of allocation is minimized. Accessibility strategies should be augmented with appropriate tracking and reporting of CI productivity and reach to evaluate the success of availability strategy. The minimization of underutilization of allocation of resources would serve as an indicator of success of an Accessibility strategy. There have been numerous creative efforts at expanding Accessibility and sharing of available HPC resources. Similar to Availability, expanded and equitable Accessibility is necessary for HPC democratization. However, Accessibility must lead to the productive, efficient, and effective state of Usability to ensure the highest likelihood of maximized usage and optimal output.

5.3 Usability

There is a difference between optimal allocation of CI resources and optimal usage of CI resources. Maximizing the allocation of resources would be a measure for effectiveness of an Accessibility strategy, but that does not ensure optimal utilization and productivity at the end user level. Productivity maximization at the end user level not only requires good Availability and Accessibility strategies, but also a robust Usability strategy. Based on lessons learned from Caliburn, usage experiences, technology engagement theories, and a broader HPC landscape review, we describe a Usability strategy to maximize end-user level productivity. An effective HPC Usability strategy consists of a well-designed HPC system with multidisciplinary orientation, easy to use human interfaces, expert usage support, and general and discipline-specific applied HPC education. Productivity maximization at the end user level would serve as an indicator of success for Usability strategies. Basic measures, for example, could use averages of ratios of actual consumption to total allocation per user, across users in a category. Such ratios and measures would also serve as a check on over-allocation (excess availability at an individual end-user level) as well. An effective Usability strategy has been missing across many CI initiatives and addressing this strategy can lead to a remarkable increase in HPC productivity without additional expensive investments into increasing HPC availability. Based on our conversations with numerous CI experts and campus leaders, although there remains a need to improve Availability and Accessibility, yet the critical pathway to the highest likelihood of maximized usage and optimal output

is through the development and expansion of Usability factors; enhanced multidisciplinary Usability is the key for sustaining and ensuring the highest return on HPC investments.

6. Implications

HPC as an evolving technological, economical, and social multidisciplinary paradigm will have significant implication for human society, and this topic by itself will require a fair amount of research. This section does not attempt to discuss all potential implications, but is restricted to select issues most relevant to the current narrative, namely, the use of HPC productivity optimization strategies, HPC education as a key to HPC democratization and special issues and equal opportunity concerns with multidisciplinary HPC.

6.1 The Future of HPC Expansion: Decreasing Depth and Increasing Breadth

Moore's law predicted the doubling of transistors every two years, and the processor industry has experienced this curve till it reached the limits of physical properties; therefore, Moore's law will no longer be relevant to the future of new technologies, to the same extent that it has been in the past (Moore 1965, 1995; Schaller 1997; Theis and Wong 2017). HPC has entered into the early stages of a post-Moore era, and we have also seen significant advances in forms of massively parallel and hybrid forms of scalable computing. The underlying technologies have proven their worth and are well understood, leading to the greatest need: satisfy a broad range of domain-specific big data and voluminous algorithmic processing. We posit that the future is going to be relatively more strongly driven by an expansion the breadth of HPC applications across domains, rather than intensifying the vertical implementation of hardware improvements for marginal benefits in size and speed. This is supported by a growing demand for newer HPC applications, and relatively dwarfed demand for more sophisticated hardware. The success of cloud-based computing services such as Amazon Web Services (AWS) attest this perspective on current trends. Therefore, HPC availability and Usability strategies must be revised to cater to a broad range of disciplines, many of which will be traditionally non-computation-intensive disciplines. Development of HPC Usability strategies, and HPC education modules in particular, will have a significant impact on HPC democratization and productivity.

6.2 The Future of HPC Education: Modular, Applied, and Multidisciplinary

The goal here is to very briefly provide an impetus for HPC education, mostly focused on multidisciplinary curricular "modularization" and democratization. Notable initiatives include the "CyberAmbassadors" program, which is a 2017 CyberTraining project, focused on developing an open-source curriculum on interpersonal communication and mentoring skills for CI professionals. Similarly, SIGHPC Education Chapter is a virtual chapter of ACM's Special Interest Group in HPC and has merged with the IHPCTC (International HPC Training Consortia). They focus on developing best practices for HPC training, but don't provide such training themselves. The Linux Clusters Institute (LCI) holds workshops on HPC system administration, at introductory, intermediate, and advanced levels. These workshops have been extremely successful, typically attracting 20–40 HPC system administrators per workshop. LCI focuses on system-facing CI professionals, not researcher-facing. The Carpentries is an international volunteer organization that has run 2300+ hands-on workshops on research computing skills that so far have served 56,000+ researchers at 250+ institutions worldwide. They have demonstrated the effectiveness of "training the trainers" of researchers in informal education at large scale, with an emphasis on technical skills and pedagogy, and not on training CI Facilitators. The Coalition for Academic Scientific Computation (CASC): Members of this non-profit organization are primarily United States institutional CI leaders, plus some national CI leaders. CASC's focus closely aligns with CI leadership. CASC is not budgeted for or positioned to take on a major teaching or training role, but significant peer mentoring emerges from CASC activities. Science Gateways Community Institute (SGCI): the SGCI offers workforce development via internships, mentoring, and travel funding to conferences for graduate and undergraduate students, connections to the Young Professional Network, and support for gateway-related career paths (subsection reference: Neeman *et al.* 2016, 2018). In spite of many such initiatives, there is a critical need to evaluate Usability strategy-oriented education. A number of these initiatives cater to education pertaining to Availability and Accessibility and not to the proportionately higher need for multidisciplinary Usability training. This leaves a huge gap in educational needs being addressed for implementing effective HPC Usability strategies.

6.3 HPC Education: Special Issues and Equal Opportunity

Some of the challenges in the practice of HPC will be associated with addressing bias, such as gender bias in technology resulting in a low representation of women in HPC practice. Although there is very limited research done on the issue of gender and HPC, it is safe to assume that attracting and retaining women in HPC practice is going to be a challenge. This

is already so for the fields of STEM and especially CS, which is affected the most by a significant underrepresentation of women (Ehrlinger *et al.* 2018). Studies conducted on the topic of female underrepresentation found that less than 20% of the technological workforce is estimated to be women" (Frachtenberg and Kaner 2020). The adverse implications of having so few women in HPC are numerous and significant (Frantzana 2019). Extant research has emphasized the need to develop gender-specific learning strategies which accommodate women learners in technology disciplines (Samuel *et al.* 2020c). These issues will need to be addressed as aspects of the Usability strategy to ensure fair and balanced HPC democratization without bias, facilitating equal opportunities to persons in all categories.

7. Conclusion

This study identifies three key strategies for HPC democratization and important principles for catalyzing multidisciplinary HPC productivity. This research thus provides critical ideas and motivations for promoting HPC based research, applications and innovation in traditionally non-CS and non-computation-intensive disciplines and domains. We believe that this is a vital need for the current decade, and anticipate that this study will contribute to the body of knowledge that will influence HPC education policy in the future. We boldly propose and call for an increased emphasis on Accessibility and Usability strategies: every institution of higher education must ensure some measure of access to HPC resources through partnerships and networks, such as XSEDE, for their faculty and students. The responsibility for such efforts must be shared between those who "own" CI resources and those who need it.

"HPC is becoming a major driver for innovation offering possibilities that currently we cannot even evaluate or think about" (Puertas-Martín *et al.* 2020). We have emphasized that the dimension which needs the most attention is multidisciplinary HPC education within the HPC Usability strategy. We posit that undergraduate and graduate programs across disciplines must contain courses with HPC concepts and application modules, such as HPC lessons in IS courses for undergraduate business programs and in MIS courses for graduate business and relevant MS programs. Furthermore, workshops and interactive virtual education modules can be used for topic and discipline specific training. An appropriate HPC-Usability strategy and forward looking HPC education modules will ensure demystification of HPC and popularize its usage for innovation and value creation, by a broader range of students from multiple disciplines, and thus nurture the future HPC and AI workforce.

Furthermore, it is obvious that institutions and corporations with fewer resources, lacking research and technological infrastructure development funding, are at a disadvantage when it comes to HPC usage for research, teaching, and practice. An invigorated vision to democratize HPC reach into the smallest of institutions, going beyond boxed-in notions of traditionally bounded HPC domains, will maximize the return on investment for CI resources, as well as promote the noble and forward looking cause of better educating a robust future technological workforce.

Acknowledgments: *We would like to thank Dr. Manish Parashar for his input and support for this research. We would also like to thank the OARC at Rutgers University for granting the two Caliburn allocation awards which facilitated this research and the RUCI lab (Rutgers Urban and Civic Informatics Lab; https://rucilab.rutgers.edu/about-ruci/).*

References

Ali, G. G. Md. Nawaz, Md. Mokhlesur Rahman, Md. Amjad Hossain, Md. Shahinoor Rahman, Kamal Chandra Paul, Jean-Claude Thill, and Jim Samuel. 2021. "Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics." *Healthcare* **9**, no. 9: 1110. doi: 10.3390/healthcare9091110

Alvargonzález, David. 2011. "Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences." *International Studies in the Philosophy of Science* **25**, no. 4: 387–403. doi: 10.1080/02698595.2011.623366

Bergman, Keren, Tom Conte, Al Gara, Maya Gokhale, Mike Heroux, Peter Kogge, Bob Lucas, Satoshi Matsuoka, Vivek Sarkar, and Olivier Temam. 2019. "Future high performance computing capabilities: Summary report of the Advanced Scientific Computing Advisory Committee (ASCAC) subcommittee." Technical Report. Washington, DC: United States Department of Energy Office of Science.

Berman, Helen, Margaret Brennan, and Manish Parashar. 2013. "Accelerating Innovation through Advanced Cyberinfrastructure: The Urgent Need for Strategic Planning for Cyberinfrastructure at Rutgers." In *Rutgers Discovery Informatics Institute Report*. Rutgers: The University of New Jersey.

Berman, Helen, Manish Parashar, Don Smith, and Margaret Brennan. 2014. "Accelerating Innovation through Advanced Cyberinfrastructure: A Strategic Vision for Research Cyberinfrastructure at Rutgers." In *Rutgers Discovery Informatics Institute Report*. Rutgers: The University of New Jersey.

Chaudhury, Bhaskar, Akshar Varma, Yashwant Keswani, Yashodhan Bhatnagar, and Samarth Parikh. 2018. "Let's HPC: A Web-Based Platform to Aid Parallel, Distributed and High Performance Computing Education." *Journal of Parallel and Distributed Computing* **118**: 213–32. doi: 10.1016/j.jpdc.2018.03.001

Choi, Bernard C. K., and Anita W. P. Pak. 2006. "Multidisciplinarity, Interdisciplinarity and Transdisciplinarity in Health Research, Services, Education and Policy: 1. Definitions, Objectives, and Evidence of Effectiveness." *Clinical and Investigative Medicine* **29**, no. 6: 351–64.

Choi, Ji-Eun, and Hale Kim. 2017. "Vertically Integrated Projects (VIP) at Inha University: The Effect of Convergence Project Education on Learning Satisfaction." *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 436–43. IEEE.

Conner, Cherilyn, Jim Samuel, Andrey Kretinin, Yana Samuel, and Lee Nadeau. 2019. "A Picture for the Words! Textual Visualization in Big Data Analytics." In *Northeast Business and Economics Association: NBEA Annual Proceedings-46*, Rhode Island, USA, 37–43,

Conner, Cherilyn, Jim Samuel, Myles Garvey, Yana Samuel, and Andrey Kretinin. 2020 "Conceptual Frameworks for Big-Data Visualization: Discussion on Models, Methods and Artificial Intelligence for Graphical Representations of Data". In *Handbook of Research for Big Data: Concepts and Techniques*. USA: Apple Academic Press.

Connor, Carolyn, Amanda Bonnie, Gary Grider, and Andree Jacobson. 2016. "Next Generation HPC Workforce Development: The Computer System, Cluster, and Networking Summer Institute." *2016 Workshop on Education for High-Performance Computing (EduHPC)*, 32–9. IEEE.

Csikszentmihalyi, Mihaly. 2014. "Play and Intrinsic Rewards." In *Flow and the Foundations of Positive Psychology*, 135–53. San Diego: Springer.

Csikszentmihalyi, Mihaly, and IsabellaSelega Csikszentmihalyi. 1992. *Optimal Experience: Psychological Studies of Flow in Consciousness*. Cambridge: Cambridge University Press.

Davis, Fred D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly* **13**, no. 3: 319–40. doi: 10.2307/249008

Ehrlinger, Joyce, E. Ashby Plant, Marissa K. Hartwig, Jordan J. Vossen, Corey J. Columb, and Lauren E. Brewer. 2018. "Do Gender Differences in Perceived Prototypical Computer Scientists and Engineers Contribute to Gender Gaps in Computer Science and Engineering?" *Sex Roles* **78**, no. 1–2: 40–51. doi: 10.1007/s11199-017-0763-x

Ezell, Stephen J., and Robert D. Atkinson. 2016. "The Vital Importance of High-Performance Computing to us Competitiveness." *Information Technology and Innovation Foundation* **28**: 1–58.

Fiore, Sandro, Mohamed Bakhouya, and WaleedW. Smari. 2018. "On the Road to Exascale: Advances in High Performance Computing and Simulations—An Overview and Editorial." *Future Generation Computer Systems* **82**: 450–458.

Frachtenberg, Eitan, and Rhody Kaner. 2020. "Representation of women in high-performance computing conferences." Technical report, EasyChair.

Frantzana, Athina. 2019. "Women's Representation and Experiences in the High Performance Computing Community." https://era.ed.ac.uk/bitstream/handle/1842/36127/Frantzana2019.pdf?sequence=1

Gao, Yuxiang, and Peng Zhang. 2016. "A Survey of Homogeneous and Heterogeneous System Architectures in High Performance Computing." *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, 170–75. IEEE.

Garvey, Myles D., Jim Samuel, and Alexander Pelaez. 2021. "Would You Please like my Tweet?! An Artificially Intelligent, Generative Probabilistic, and Econometric Based System Design for Popularity-Driven Tweet Content Generation." *Decision Support Systems* **144**: 113497. doi: 10.1016/j.dss.2021.113497

George, Nathan. 2020. "Literature Review and Implementation Overview: High Performance Computing with Graphics Processing Units for Classroom and Research Use." Preprint, submitted May, 2020. *arXiv preprint arXiv:2005.07598*.

Grover, Purva, Arpan Kumar Kar, Marijn Janssen, and P. Vigneswara Ilavarasan. 2019. "Perceived Usefulness, Ease of Use and User Acceptance of Blockchain Technology for Digital Transactions–Insights from User-Generated Content on Twitter." *Enterprise Information Systems* **13**, no. 6: 771–800. doi: 10.1080/17517575.2019.1599446

Kretinin, Andrey, Jim Samuel, and Rajiv Kashyap. 2018. "When the Going Gets Tough, the Tweets Get Going! An Exploratory Analysis of Tweets Sentiments in the Stock Market." *American Journal of Management* **18**, no. 5: 23–36.

Koufaris, Marios. 2002. "Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior." *Information Systems Research* **13**, no. 2: 205–23. doi: 10.1287/isre.13.2.205.83

Lathrop, Scott. 2016. "A Call to Action to Prepare the High-Performance Computing Workforce." *Computing in Science & Engineering* **18**, no. 6: 80–3. doi: 10.1109/MCSE.2016.101

Moore, Gordon E. 1965. "Cramming More Components onto Integrated Circuits." *Proceedings of the IEEE* **86**, no. 1 (1998): 82–85.

Moore, Gordon E. 1995. "Lithography and the Future of Moore's Law." In *Integrated Circuit Metrology, Inspection, and Process Control IX*, volume 2439, 2–17. International Society for Optics and Photonics.

Moran, Nuala, and Joanne O'Dea. 2013. "Prace Special Report Supercomputers For All—The Next Frontier for High Performance Computing." Brussels, Belgium: Science | Business Publishing Ltd. Accessed September 9, 2021, https://www.genci.fr/sites/default/files/prace_report_october_2013.pdf.

Mosin, S. 2017. "The State of the Art Trends in Education Strategy for Sustainable Development of the High Performance Computing Ecosystem." In *Russian Supercomputing Days*, 494–504. San Diego: Springer.

Neeman, Henry, Hussein M. Al-Azzawi, Aaron Bergstrom, Zoe K. Braiterman, Dana Brunson, Dirk Colbry, Eduardo Colmenares, et alet al. 2018. "Progress Update on the Development and Implementation of the Advanced Cyberinfrastructure Research & Education Facilitators Virtual Residency Program." *Proceedings of the Practice and Experience on Advanced Research Computing*, 1–7,

Neeman, Henry, Aaron Bergstrom, Dana Brunson, Carrie Ganote, Zane Gray, Brian Guilfoos, Robert Kalescky, et alet al. 2016. "The Advanced Cyberinfrastructure Research and Education Facilitators Virtual Residency: Toward a National Cyberinfrastructure Workforce." *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, 1–8.

Neumann, Philipp, Christoph Kowitz, Felix Schranner, and Dmitrii Azarnykh. 2017. "Interdisciplinary Teamwork in HPC Education: Challenges, Concepts, and Outcomes." *Journal of Parallel and Distributed Computing* **105**: 83–91. doi: 10.1016/j.jpdc.2016.12.025

Obienu, A. C., and F. I. Amadin. 2021. "User Acceptance of Learning Innovation: A Structural Equation Modelling Based on the GUAM Framework." *Education and Information Technologies* **26**, no. 2: 2091–33. doi: 10.1007/s10639-020-10341-x

Parashar, Manish. 2019. "CALIBURN: Advanced Cyberinfrastructure Report." *Rutgers Discovery Informatics Institute Report*. Rutgers: The University of New Jersey.

Puertas-Martín, Savíns, Antonio J. Banegas-Luna, María Paredes-Ramos, Juana L. Redondo, Pilar M. Ortigosa, Ol'ha O. Brovarets', and Horacio Pérez-Sánchez. 2020. "Is High Performance Computing a Requirement for Novel Drug Discovery and How Will This Impact Academic Efforts?" *Expert Opinion on Drug Discovery* **15**, no. 9: 981–85. doi: 10.1080/17460441.2020.1758664

Rahman, Md. Mokhlesur, G. G. Md. Nawaz Ali, Xue Jun Li, Kamal Chandra Paul, and Peter H. J. Chong. 2020. "Twitter and Census Data Analytics to Explore Socioeconomic Factors for Post-Covid-19 Reopening Sentiment." https://arxiv.org/pdf/2007.00054

Rietz, Tim, Ivo Benke, and Alexander Maedche. 2019. "The Impact of Anthropomorphic and Functional Chatbot Design Features in Enterprise Collaboration Systems on User Acceptance." In *Proceedings of the 14th International Conference on Wirtschaftsinformatik*, Siegen, Germany, February 24–27.

Samuel, Jim. 2016. "An Analysis of Technological Features Enabled Management of Information Facets." PhD diss., CUNY Academic Works.

Samuel, Jim. 2017. "Information Token Driven Machine Learning for Electronic Markets: Performance Effects in Behavioral Financial Big Data Analytics." *Journal of Information Systems and Technology Management* **14**, no. 3: 371–83. doi: 10.4301/S1807-17752017000300005

Samuel, Jim. 2021. "A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies." *SSN: Scholars.org*. Accessed December, 2021. https://scholars.org/contribution/call-proactive-policies-informatics-and.

Samuel, Jim, Md. Rahman Gg Ali, Ek Esawi, Yana Samuel, et alet al. 2020a. "Covid-19 Public Sentiment Insights and Machine Learning for Tweets Classification." *Information* **11**, no. 6: 314. doi: 10.3390/info11060314

Samuel, Jim, Rajiv Kashyap, and Andrey Kretinin. 2018. "Going Where the Tweets Get Moving! An Explorative Analysis of Tweets Sentiments in the Stock Market." *Proceedings of the Northeast Business & Economics Association*, NJ, USA

Samuel, Jim, Rajiv Kashyap, Yana Samuel, and Alexander Pelaez. 2022. "Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations." *International Journal of Information Management* **65**: 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, Jim, Ratnakar Palle, and Eduardo Soares. 2021. "Textual Data Distributions: Kullback Leibler Textual Distributions Contrasts on GPT-2 Generated Texts with Supervised, Unsupervised Learning on Vaccine & Market Topics & Sentiment." Preprint, submitted June, 2021. *SSRN*, http://ssrn.com/abstract=3856396.

Samuel, Jim, and Alexander Pelaez. 2017. "Informatics in Information Richness: A Market Mover? An Examination of Information Richness in Electronic Markets." *Proceedings of the 8th International Conference on Society and Information Technologies, ICSIT*, FL, USA

Samuel, Jim, Md. Mokhlesur Rahman, G. G. Md. Nawaz Ali, Yana Samuel, Alexander Pelaez, Peter Han Joo Chong, and Michael Yakubov. 2020b. "Feeling Positive about Reopening? new Normal Scenarios from Covid-19 us Reopen Sentiment Analytics." *IEEE Access* **8**: 142173–90. doi: 10.1109/ACCESS.2020.3013933

Samuel, Yana, Jean George, and Jim Samuel. 2018. "Beyond STEM, How Can Women Engage Big Data, Analytics, Robotics and Artificial Intelligence? An Exploratory Analysis of Confidence and Educational Factors in the Emerging Technology Waves Influencing the Role of, and Impact Upon, Women." In *2018 NEDSI Annual Conference* (47th) (p. 359)

Sancho, Maria-Ribera. 2016. "BSC Best Practices in Professional Training and Teaching for the HPC Ecosystem." *Journal of Computational Science* **14**: 74–7. doi: 10.1016/j.jocs.2015.12.004

Schaller, Robert R. 1997. "Moore's Law: Past, Present and Future." *IEEE Spectrum* **34**, no. 6: 52–9. doi: 10.1109/6.591665

Theis, Thomas N., and H.-S. Philip Wong. 2017. "The End of Moore's Law: A New Beginning for Information Technology." *Computing in Science & Engineering* **19**, no. 2: 41–50. doi: 10.1109/MCSE.2017.29

Venkatesh, Viswanath, and Michael G. Morris. 2000. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior." *MIS Quarterly* **24**, no. 1: 115–39. doi: 10.2307/3250981

Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. "User Acceptance of Information Technology: Toward a Unified View." *MIS Quarterly* **27**, no. 3 425–78. doi: 10.2307/30036540

WWW.JBDTP.ORG

ISSN: 2692-797

JBDTP Professional Vol. 1, No. 1, 2022 DOI: 10.54116/jbdtp.v1i1.17

In Search of Pedagogical Approaches to Teaching Business Ethics in the Era of Digital Transformation

Ethné Swartz
Feliciano School of Business,
Montclair State University,
swartze@montclair.edu

Rashmi Jain
Feliciano School of Business,
Montclair State University
jainra@montclair.edu

Margaret Brennan-Tonneta
New Jersey
Agricultural Experiment Station,
Rutgers University
mbrennan@rutgers.edu

Marina Johnson
Feliciano School of Business,
Montclair State University
johnsonmari@montclair.edu

Stanislav Mamonov
Feliciano School of Business,
Montclair State University
mamonovs@montclair.edu

Matthew Hale
Seton Hall University
Matthew.Hale@shu.edu

J.D. Jayaraman
New Jersey City University
jjayaraman@njcu.edu

Abstract

The authors explore the challenges in teaching business ethics in an era of digital transformation, provide an understanding of the limitations of traditional ethics approaches, and explore emerging approaches that may more effectively deal with the ethical complexities of the new digital era. Building on previous research conducted during December 2018 and January 2019 with regard to the skills required for jobs in the Big Data field, the authors argue that business ethics must be an essential skill for those working in this field. Ethical frameworks in Big Data and information management, including Universalist, Integrative Social Contract Theory, and Care Theory as well as agency, disciplinary, contextual, and outcomes considerations are discussed. The authors posit that traditional ethical frameworks, such as Universalist approaches, are no longer sufficient to guide decision-making in an era of digital transformation and the "datafication" of society. Business educators have a duty to cultivate ethical mindsets in their students and the adoption of "responsible innovation" principles such as those developed in the science and technology innovation literature.

Keywords *digital transformation, business ethics, Care Theory, Integrative Social Contract Theory, pedagogy, datafication*

1. Introduction

In January 2020, *The New York Times* published a leaked memorandum in which Facebook executive, Andrew Bosworth, discussed the outcome of the 2016 US election, quoting the political philosopher, John Rawls, to justify why Facebook's position on political advertising is fair and just:

> The philosopher John Rawls reasoned that the only moral way to decide something is to remove yourself entirely from the specifics of any one person involved, behind a so called "Veil of Ignorance." That is the tool that leads me to believe in liberal government programs like universal healthcare, expanding housing programs, and promoting civil rights. It is also the tool that prevents me from limiting the reach of publications who have earned their audience, as distasteful as their content may be to me and even to the moral philosophy I hold so dear. ("Lord of the Rings, 2020" 2020)

Rawls published in the 1970s, and he is finding a new life as researchers and others search for answers to the vexing problem of what is the "moral" or "right" thing to do as new technologies emerge in all aspects of business, but particularly in Big Data, e-commerce, and artificial intelligence (AI). Given the relatively new ethical challenges arising from the digital transformation of society and business, research on Big Data and ethics is in the early stages of development (Kuc-Czarnecka and Olczyk 2020). Scholars have argued that teaching ethics should not focus on what students should think but engage students in discussions of often nonobvious implications of data science and AI (Heggeseth 2019). At a societal level, this concern is clearly visible in the public debates about the responsibility and power of technology corporations; the role of governments; and the responsibility of individuals as citizens, employees, and consumers. The seemingly ethical argument advanced by Facebook's Bosworth highlights how moral philosophy can be used to justify or account for actions that may have significant negative societal outcomes (Kim *et al.* 2021).

Teaching students how to evaluate such arguments is key to what educators need to do in the classroom, exposing the incubation of negative outcomes, particularly in the design of interactions of technology and humans. Tim Wu (2016), author of "The Attention Merchants" highlights "a sense of attentional crisis" that the human race is experiencing. He uses the term "homo distractus" to describe our species as now characterized by a short attention span and compulsively checking our devices. Further reinforcing the crisis, Wu quotes William James who opinioned that, at the end of our lives, we end up with experiences of "what we paid attention to," whether by choice or by default. He further emphasizes the role of ethics as addressed through the design of online interaction on social media. Design is what sets the terms of any online interaction, which he calls the "agenda-setter." Wu (2016) calls for a "human reclamation project," a key component of which is technologists who are devoted to a different ethic. He highlights the need of the hour, more tools that are designed to serve their owner's interests and less driven by other agendas. Instead of helping users achieve their goals, the design is often used to exploit users' weaknesses. Designers and companies must put users first and design tools that work for them, not against them. "Very few things are more important now to the future of humanity than design ethics," Wu says. "Design is the determinant, along with your will. But design creates the way you exercise choices" (Schwab 2018).

As companies shift from generating data to relying on data and data combinations to create business value (Berinato 2019; Goes 2014), we seek to revisit the responsibility of educators to ensure pedagogy equips students' ethical judgment vis-a-vis the societal implications of technology development. Zwitter (2014) argues that certain principles of contemporary philosophy of ethics might require changes in "philosophy, professional ethics, policy-making, and research." Berinato (2019) suggests that analytics projects add value when a team "asks smart questions, wrangle relevant data and uncover insights. Second, it must figure out—and communicate—what those insights mean for the business. The ability to do both is extremely rare …" and requires a variety of capabilities, including project management, data analysis, data wrangling, design, and storytelling. In this paper, we suggest that asking "smart" questions and understanding what those answers mean to the business must be coupled with understanding the consequences of Big Data analysis for the individuals whose data are aggregated for our business purposes. Furthermore, our pedagogical approaches must highlight whether the decisions we make for business purposes may have detrimental individual or societal outcomes.

The need for ethics in data science has been frequently noted (Floridi and Taddeo 2016; Schwartz 2011; Fung 2015). For example, Fung (2015) recommends that every data science and analytics team should have onboarding training that covers the ethics of using data and exposes data scientists and engineers to the legal obligations and regulations of using data. Schwartz (2011) suggests that organizations should develop policies and designate a team of individuals to govern information management and analytics processes to align with ethics guidelines, laws, and regulations. It has also been noted that organizations that practice data science should provide ethical training and

participative ethical assessments to analyze ethical issues, but it is not clear that organizations have the breadth and depth of knowledge to effectively offer this training (Saltz and Dewar 2019). For business educators, how to integrate such critical thinking into curricula is a constant challenge and worthy of debate.

The research question addressed here is the following: How can business schools best educate students in data-related courses to ask "smart" questions by using ethical guidelines that they acquire while studying information management, Big Data, and analytics. Indeed, the National Academies of Sciences, Engineering and Medicine (National Academies of Sciences, Engineering, and Medicine 2018) recommend that disciplines should adopt a code of ethics as part of professional practice and that these codes be re-evaluated in line with new knowledge and developments. We argue that the time has arrived for business ethics practitioners and educators to heed this advice. The ubiquitous nature of Big Data (Kuc-Czarnecka and Olczyk 2020) and its critical link with creating value for businesses (Bazerman 2020) makes it imperative for a re-evaluation of pedagogical approaches to business ethics. We argue that there is a normative ethical case for including ethical approaches in the study of Big Data. However, we also argue that there is a business or "self-interested" approach to ensure that our students understand the ethical implications of Big Data processes to both individuals and society.

2. Background to the Research

This paper builds on research that we conducted as members of the New Jersey Big Data Alliance during December 2018 and January 2019 to develop a "New Jersey Big Data Workforce Roadmap" that could ensure a skilled workforce in the State of New Jersey is prepared for current and future technological changes (Johnson *et al.* 2021). In this research, we examined specific technological changes, such as AI, machine learning, and large-scale automation, that are impacting New Jersey's major industry clusters (e.g., health care, logistics, food, financial services, clean energy, and advanced manufacturing), and the resultant skills needed for a competitive workforce. We found the fastest growing skills across these industries include predictive analysis, machine learning, and data visualization.

Also of interest was the significant demand for a workforce with "hybrid" skills, those that combine technological expertise along with softer skills, such as communication, teamwork, research, and problem solving, among others. Of importance, we provided a recommendation that stresses the need for lifelong training to maintain appropriate technology skills, which are rapidly changing, as a precondition for successfully participating in the data-driven economy. Our research concluded, "most companies have not realized the full potential of these technological advances due to a number of barriers, including talent shortages. By educating, training, and facilitating access to individuals with advanced computing and analytics skill sets, New Jersey can provide a competitive advantage for its employers" (Johnson *et al.* 2021). However, training in business ethics and applying ethical considerations through critical thinking skills was not a part of that research.

Although the initial intent for this paper was to conduct a systematic review of the extant frameworks in the research publications that focus on pedagogy and approaches to teaching ethics in Big Data and information management, we decided that an initial position paper on the topic would be more appropriate. We also decided to narrow the scope of this position paper to business ethics pedagogy and its link with Big Data, and to use an emergent approach to wrestle with the critical issues related to business ethics and digital transformation.

3. Pedagogical Approaches to Teaching Business Ethics: What Is the State of Play and What Is Missing?

Business school curricula typically include an ethics requirement as part of both undergraduate and graduate degrees. Such curriculum content typically resides in the core business curriculum (AACSB 2004) and can also be interwoven throughout the curriculum (Godwyn 2015).

Business ethics education, according to Association for the Advancement of Collegiate Schools of Business (AACSB) (2004), typically focuses on four themes. First, the responsibility of business to society (wealth creation, job creation, consideration of stakeholder interests, etc.). Second, business schools educate students about the importance of ethical leadership for effective management (ensuring that students understand the need to develop ethics decision-making skills and the relationship of normative ethics to business ethics). Third, students learn about frameworks that can assist with making ethical decisions (consequentialist, deontological, and virtue ethics are typically used) and guide ethical behavior. Fourth, to augment the shaping of ethical behavior, schools have incorporated corporate governance as an important facet of ethics training to ensure that students understand national and international legislation, guidelines, and professional codes of conduct.

Business school curricula cover these themes in core classes and in courses on Big Data, AI, information systems, and technology in business. Instructors use a mix of approaches, such as case studies, simulations, group discussions, and guest lectures, to improve students' understanding of ethics (Sexton and Garner 2020; Rutherford *et al.* 2012). However, Godwyn (2015) conducted qualitative research among business school students and educators across multiple continents and found that attitudes (and approaches) often conflicted. Godwyn (2015) found that, in many cases, educators who cared deeply about the underlying values and responsibilities of business often faced assumptions by students who demeaned ethics classes as unimportant, which led her to contend that:

> Individuals manifest definitions of ethical behavior that fluctuate depending on the group or groups with which they are currently identifying. Using concepts introduced by Hannah Arendt and Emile Durkheim, I argue that because of the social solidarity ritualized in part by identification with the ethical values associated with the business world, business ethics and the resulting behavior often remain hidden and evade critical examination.

In echoing this concern, the AACSB report (AACSB 2004) notes that a persistent failure in business ethics is the "development of 'moral courage.' . . . Examples abound of individuals with 'solid' values who failed to do the right thing because of constraints imposed by authority structures and unethical corporate cultures."

Students drawn from generations such as Millennials and Gen Z have noted these failures across all four of the AACSB ethics themes. For example, in the sphere of ethical leadership at a societal level, students from these generations have been vocal participants in Black Lives Matter protests in 2020. In January 2021 Manchester United footballer, Marcus Rashford, aided by social media posts, exposed the abuse of UK government funding (for food aid) by a major food corporation during the pandemic (Campbell and Weale 2021). In both cases, young people used social media to expose leadership failures. However, social media use has also had negative consequences, and the limits to how business ethics are taught by faculty and received by students are more visible now because of the pace of change as well as a more fine-grained understanding of technology developments (and attendant problems). For instance, exploring factors that can affect trust in AI systems, Kumar *et al.* (2020) propose a framework for trustworthy AI that suggests that ethics of algorithms, ethics of data, and ethics of practice represent three distinct areas of concern. Besides AACSB, through Accreditation Board for Engineering and Technology (ABET), the science and engineering community has tried to explore the challenges of ethics in data science. "We have a responsibility to build a better world and that means arming students of applied science, computing, engineering, and engineering technology with the real-world skills and the moral courage to step into high-stakes environments with the clarity to know when something is right and when it feels wrong" (Milligan 2018).

The magnitude of societal implications of the ethical challenges in data science have become increasingly complex and incomprehensible in many ways over the years. For example, let us compare the responsibilities of engineers and designers for engineering disasters, such as the defective fuel system design of the Ford Pinto car (Ford Motor Company, Dearborn, MI, USA) in the late sixties through the mid-seventies, with that of Cambridge Analytica's obtaining of 87 million individual Facebook profiles by Cambridge Analytica and subsequently selling this information to 2016 election campaigns. In Ford's case, the company recalled 1.5 million cars built during 1971–1976 and the case led to significant regulatory changes about car safety. In the latter case, the company filed for bankruptcy during mid-2018, and this led to several lawsuits against Facebook and Cambridge Analytica (London, United Kingdom), but regulatory changes are still being argued in Congress. Monitoring of customer behavior on the World Wide Web and exploiting that data for selling services and products (and selling the customer data to other companies) have become a rampant business. Technology companies have developed a new normal for how we learn, watch television, drive, communicate, shop, and even express our feelings.

Educators need to teach ethics of accountability. For instance, an educator might encourage students to question who is responsible when a social media company shares information used for much deeper political purposes. Alternatively, how does one deal with the ethics of a driverless car that kills a pedestrian or the tangible effects of AI on the future of American jobs, transportation, or warfare? Responsible AI that is designed to benefit and not harm people calls for ethical rationality, just as we expect of our engineered structures and products.

A recent AACSB International-sponsored initiative called Management Curriculum for the Digital Era, led by Stevens Institute of Technology, NJ, brought together 100 institutions of higher education to investigate and identify business disruption due to digital transformation and how the academic community is preparing to address this. One of the task forces within this initiative focused on data science. The report of this task force noted that there is a different philosophical approach to how undergraduate courses present ethics in analytics and data science, as opposed to graduate courses. For instance, although programs at both levels offer emphasis on concepts such as algorithms, machine learning, and Big Data, undergraduate programs have a much heavier focus on the ethics associated with the use of these technologies, in addition to decision support systems. Although 10 of 25 undergraduate programs reported including ethics and human factors in their curriculum, only 7 of 32 graduate programs reported the same (Lyytinen *et al.* 2020).

Table 1: Elements of innovation and impact questions.

Elements of Innovation	Indicative Impact Questions
Product	What are the risks and/or benefits, what are the impacts we know this will have, what do we already know, and what might we want to know?
Process	What standards have been applied and how do we measure risks and/or benefits; who is in control and participation, and how do we know we are right?
Purpose	Why is the innovation or technology being developed? Who benefits and are those motivations transparent and in the public interest? What do the developers gain and what are the alternatives?

Table 2: Codes of ethics principles.

Principle	Guidance
Beneficence	Do good (promote individual and community well-being and preserve trust in trustworthy agents)
Nonmaleficence	Avoid harm (also by protecting security, privacy, dignity, and sustainability)
Autonomy	Promote the capabilities of individuals and groups (also by protecting civic and political freedoms, privacy, and dignity)
Justice	Be fair, avoid discrimination, and promote social justice and solidarity
Control	Knowledgeably control entities, goals, process, and outcomes that affect people
Transparency	Communicate your knowledge of entities, goals, process, and outcomes, in an adequate and effective way, to the relevant stakeholders
Accountability	Assign moral, legal, and organizational responsibilities to the individuals who control entities, goals, processes, and outcomes that affect people

Over the past decade, there has been an increasing awareness of the need for responsible or ethical development of technologies, and the history of this movement in terms of research policy has been covered by various authors (Stilgoe *et al.* 2013; Owen *et al.* 2012). Supported by science foundations and government science councils in the United Kingdom and the United States, the more recent focus in developing technologies with some "foresight" recognizes that innovation creates unintended externalities, which renders the conventional response of governing such consequences through regulation insufficient. Stilgoe *et al.* (2013) suggest that considering a regulatory framework before implementation might be required so that "data before market" can inform implementation. In the science and technology domains, there is a shift from governance of risks toward governance of innovation, and responsible innovation requires that we ask questions about the impacts across the dimensions of product, process, and purpose of the innovation (Stilgoe *et al.* 2013). These elements are in Table 1.

Stilgoe *et al.* (2013) provide four dimensions for a deliberative framework to govern innovation: anticipation, reflectivity, inclusion, and responsiveness. Some of these elements are found in the guidance of the European Commission's Ethics Guidelines for Trustworthy AI (European Commission 2019) and they recur in the themes that Loi *et al.* (2020) discern in their evaluation of codes of ethics made public by 20 leading technology companies and summarized in Table 2.

4. The Limits of Universalist Approaches—Rise of Professional Codes?

In most professional disciplines, including business, the primary ethical frameworks used in the classroom draw on Universalist approaches. By Universalist, we refer to approaches to ethics that adhere to Western philosophies grounded in the idea that there are norms that hold true and transcend historical periods (Evanoff 2004). These broadly include a teleological approach that centers ethical decision-making around providing a positive outcome to some defined population or community. The Utilitarian conception of "the greatest good for the greatest many" is a prime example of an ethical and moral philosophy that conceptualizes "good" and "ethical" as what brings the most "utility" to the largest community. However, the focus on outcomes allows for morally defensible arguments that boundaries can be set on the community in question. As such, what is "good" and "ethical" for a

company, a community, or a nation state could become, what is good for "my" company, community, or nation state.

In contrast to the teleological approaches, deontological approaches also posit a standard set of norms that hold across historical periods but argue that there exists a set of universal moral truths (in general, do not kill, steal, or hurt) that transcend not just time but community. It is always wrong to violate a universal truth in this conception of ethics and morality. However, the difficulty with this approach is that the subset of universal moral truths is admittedly small (as noted, killing and stealing and perhaps the more amorphous hurting). This often results in the concept that all other decisions and acts outside of these universal moral truths are open to interpretation. For instance, lying is generally bad, but because one can lie without intent to hurt, kill or steal, lying may be acceptable under some circumstances.

Another important approach to ethics in the professional education setting is the importance of ethical decision-making. The basic conception is that professional students need practice in confronting ethical dilemmas and in developing a process by which they can confront and process underlying assumptions. The goal and hope of this approach are that, by practicing and solving ethical dilemmas in the classroom, students will be better positioned to solve those ethical dilemmas in the real world.

In recent years, researchers have raised concerns about the limits of these Universalist approaches to ethics for use in business, and specifically Big Data (Zwitter 2014; Evanoff 2004; Donleavy 2007). For Zwitter (2014), Big Data changes our assumptions about free will, power, and individuality, whereas Donleavy (2007) argues that ethics is and should be concerned with relationships rather than atomistic individuals. These contextual elements are missing from Universalist approaches (Evanoff 2004). However, classical frameworks developed by Kant (deotology), Hobbes (teleology) and more recently the work of Rawls (ethical decision-making), all assume that valid ethical stances presuppose universal application.[1] In addition, (Donleavy (2007) posits that Universalist frameworks can be exploited or abused by dictators ("ends justify the means"). Instead, recent approaches such as the Integrative Social Contract Theory (ISCT) and the Care Theory build on the essence of Universalist frameworks when considering context and including the realities of competition.

Expressing similar concerns about the limits to Universalist theories and, in particular, utilitarian frameworks, Taylor (2016) discusses the complexity that faces the growing responsible data movement in relation to data-sharing practices to tackle development problems in lower and middle income countries. Utilitarian perspectives assume that an objective "litmus test" can be developed to evaluate under what conditions digital data can be shared for humanitarian purposes and to advance the "common good." Taylor (2016) considers the specific case of lower and middle income countries where mobile telephone operators have been cautious to share data. She provides a nuanced discussion of specific cases that involved European operators who are reluctant to share call detail records and uncovers reasons for their reluctance. Taylor (2016) discusses the assumptions made by different actors in the responsible data movement, the (underlying) disciplinary worldviews, and their incentives.

There are competing claims between the Big Data companies and the responsible data movement due to fundamental issues such as power, rights, and legal responsibilities, and the issue of the nature of knowledge versus the nature of data. Taylor (2016) endorses the perspective of Purtova (2015) that digital data are a "system resource comprising an ecosystem of people, platforms and profiles," which makes data as a public good argument difficult to sustain. In line with Purtova (2015), Taylor (2016) believes that knowledge generated from digital data can be transformed and turned into a public good. Taylor (2016) cites the fact that the World Economic Forum adopted this distinction by referring to a "personal data ecosystem" to explain how knowledge produced through personal data is a commercial process that occurs through the contact that companies have with individuals. This development fits with the reshaping of individual data ownership and with compensating people for the use of their data (Berners-Lee 2019).

One important observation we make is that various disciplinary worldviews weigh the costs and benefits of using a dataset differently. Data scientists use a framework that moves from identifying risks and harms of using specific types of data and then evaluating the potential for these events to occur. Next, beneficiaries are identified and only then are the positive (or negative) effects that might result from the use of the data are determined. In contrast, social scientists move from identifying the context of a problem that needs solving and then finding the data to help solve the problem, as summarized:

- Identify risks of data usage and evaluate potential
- Apply to solving problems
- Identify social problems to solve
- Find data to solve problems

[1]In addition to Evanoff (2004), we refer readers to the work of Bauman (1993) for a more comprehensive review of classical approaches to ethics. Baumanm Z. 1993. *Postmodern Ethics*. Oxford: Blackwell.

ISCT requires prior consent of the contract parties to an engagement or transaction. Donleavy (2007) discusses the problems with prior consent as riddled with assumptions and contradictions. For example, contracts carry normative force because of prevailing power structures (having to agree to an online provider's terms of service even if one does not fully agree, just to gain access to that service); those signing the contract might not fully understand the implications to what they consent; and children are not generally covered well by ISCT.

Although ISCT tries to identify hyper-norms that have been difficult to specify clearly in business contracting, Care Theory fills that vacuum. A feminist ethicist (Noddings 2003) formulated the first Care Theory framework. She proposed that an ethical framework is concerned with justice and that this is merely the superstructure underpinned by a foundation posited to be the care and concern for the subject. Thus, Care Theory draws on the belief that justice is not possible without caring and a sense of compassion that moves on to consider rights, equality, or merit. Donleavy (2007) argues that a framework that merges Care Theory's contextual sensibility and a focus on needs rather than interests and a commitment to dialogue, provides a sound basis for moral deliberation. He ends by asserting that such a framework is participative rather than detached, must support Kant's moral norm of treating people as ends rather than means, and maintains a non-Universalist point of view, which considers local norms, traditions, and understanding.

Researchers in biomedical fields have raised concerns about ethics in their fields (Floridi and Taddeo 2016; Mittelstadt and Floridi 2016) as data sources and aggregated datasets proliferate. Medical and biomedical data are highly sensitive at the individual level and at the group level. When a private equity company such as Blackstone (New York, NY, USA) acquires Ancestry.com (Lehi, UT, USA), ethical foresight is necessary to anticipate the implications that arise from making a profit from a DNA testing database and associated information about individuals and families. What are the interests of the investors? Furthermore, how do those interests conflict or align with those of families who wish to understand their cultural and biological heritage? Will those emotional needs that lead to individuals signing up to become an Ancestry client supersede the profit needs of the investor?

Smart questions also include asking whether the company and the clients using its technology define value the same way. Floridi (Mittelstadt and Floridi 2016) suggests that gaps in perspective occur in the following areas:

1. Informed consent
2. Privacy rights, including anonymization and data protection
3. Ownership of data
4. Epistemology and objectivity, including assessing ethics of Big Data
5. Big Data divides
6. Group-level ethical harms
7. Fiduciary relationships that become data saturated
8. Distinction between commercial and academic practices and harm to subjects
9. Future problems of ownership with data generated from aggregated datasets
10. Meaningful access rights to data subjects who lack resources or knowledge.

Nunan and Di Domenico (2013) echo these concerns as well as Zwitter (2014), who argue that we are often forced to judge by using ordinary moral norms in "unchartered realms." As Blackburn *et al.* (2020) show, from data collected globally, the Fourth Industrial Era society is so digitally connected that power is often distributed and networked. Individual agency is often compromised, and Zwitter (2014) quotes Simon (2016) that individual knowledge and ability to act is one of the most difficult phenomena that we witness for the "governance of sociotechnical epistemic systems." For example, during the summer 2020 Black Lives Matter protests in the United States (and globally), many individuals who felt aligned with the goals of the protests but concerned about the violence, were confused about how to signal their support. Many felt that social media posts by friends and acquaintances left them no choice in how they discussed their support, and some felt it best not to say anything. In contrast, many organizations and corporations saw the opportunity to post messages of support, despite having a history of racism or lacking values aligned with those of the protesters. This "networked agency" (Floridi and Taddeo 2016) becomes a factor when judging the moral responsibility of individuals or agents.

5. Toward a New Pedagogy of Business Ethics

It is important to acknowledge the ongoing polemic contrasting normative and behavioral approaches to teaching ethics (Kim *et al.* 2021; Bazerman 2020). It is clear that we need both approaches in teaching business ethics to

provide a fuller understanding of the complexities of this issue. Students need to understand that normative approaches emphasize prescriptive evaluations of alternative courses of action (Stahl and De Luque 2014; Tenbrunsel and Smith-Crowe 2008; Treviño and Weaver 1994), whereas behavioral ethics approaches focus on understanding factors that influence ethical behavior (Cropanzano and Stein 2009; De Cremer, Mayer, and Schminke 2010; Treviño, Weaver, and Reynolds 2006; Banaji, Bazerman, and Chugh 2003).

De Los Reyes *et al.* (2017) suggest that normative (how) and behavioral approaches (why) offer complementary perspectives and integration can provide business practitioners with guidance to act ethically while understanding that, as humans, there will be behavioral challenges in specific contexts (De Los Reyes *et al.* 2017). In other words, these approaches cultivate a moral code and establish what moral courage is, while encouraging self-reflection and curiosity with regard to outcomes that may be inconsistent with both the accepted definitions of moral code and moral courage.

Thus, how should pedagogy change to incorporate these approaches in curricula? Kuc-Czarnecka and Olczyk (2020) suggest that the behavioral sciences (economics, management, business, political science, and sociology) lag in publishing on the issue of ethics and Big Data and contend that a unique set of competencies are required to understand the issues:

> the contextual and multi-level phenomenon of ethics and Big Data is a demanding research area, requiring extensive knowledge, both philosophical and purely technical. These factors contribute to the relatively low popularity of the issue . . .

Saltz and Dewar (2019) noted that, because data science is a new domain, the full breadth and depth of its ethical challenges have yet to be fully explored, and although the field has grown, it often excludes ethical analysis in both practice and academia. In addition, there is disagreement about what constitutes ethical versus unethical use of data science. They suggest the creation of a framework of the different ethical challenges that a team might encounter when working on a data science project. Furthermore, they suggest that these identified challenges can then be used proactively when executing the project.

Saltz and Dewar (2019) also noted an increasing dialog on the subject of ethics in data science as demonstrated by the significant increase in the number of recently published articles on this subject, with the majority of the identified papers reported to have been recently published, only 8 of the 80 identified articles were published before 2014. The highest concentration of articles was published in information technology focused journals and/or conferences (17), more than five papers published in each of the journals and conferences focused on philosophy and/or ethics, information technology and/or engineering ethics, and Big Data and/or data science, and the remaining seven in law and health.

Insight from the research of Saltz and Dewar (2019) with regard to pedagogy could provide a useful framing for the classroom. They acknowledge four challenges, including the need for an ethics framework, the newness of the field, data-related challenges, and model-related challenges. The latter two themes are identified as two general paths to potentially cause harm:

1. Data-related challenges: the preparation, storage, and dissemination of data could impinge on the privacy or anonymity of the subject or cause bias in the resulting analytics. For example, just because data are available does not make it ethical to use that data.
2. Model-related challenges: incorrectness of a data science model, for example, in which some subjects could be misclassified, resulting in harm. A model might operate correctly, but the objective of the model is inherently unfair to some subjects. Although data science can bring objectivity to decision-making, there is subjectivity within data science modeling that involves making decisions about which algorithm to use, which data sources to use, whether one data point should be used as a proxy for a missing fact, and how to interpret results.

A *New York Times* article (Singer 2018) compares the ethic of the medical profession "First, do no harm" with that of the Silicon Valley "Build it first and ask for forgiveness later!" Now, with the role of social media in question for creating an "alternate reality" for different sets of consumers, there are serious ethical implications for big tech companies. Some universities that helped produce some of Silicon Valley's top technologists are introducing ethics in their design and data science curriculum. Harvard University and the Massachusetts Institute of Technology jointly offered a new course on the ethics and regulation of AI. Stanford University and Cornell University of Texas at Austin now have courses in computer science ethics. Cornell introduced a course in data science that focuses on teaching students "how to deal with ethical challenges."

The Harvard–Massachusetts Institute of Technology courses focus on the ethical, policy, and legal implications of AI. Some jarring examples of bias and stereotyping are covered, such as the spread of algorithmic risk scores that

use data to predict the probability of someone committing a crime based on whether a person was ever suspended from school or how many of his or her friends have arrest records. These courses that teach about the ethical issues in technology development have now become a necessity in the teaching of powerful tools such as machine learning, AI, Deep Learning, and other similar tools that involve computer algorithms that can autonomously learn tasks by analyzing large amounts of data. The fear is that such tools could ultimately alter human society by controlling social communications and responses. The consequences could be irreversible.

6. Conclusion

In this research, we reviewed the prevailing approaches to teaching business ethics and provided specific examples of how these approaches are insufficient in the era of digital transformation. The question that will continue to drive the concerns of educators is how to integrate these findings within a data science curriculum. These ethics concepts could be integrated within existing classes with the creation of key questions for students such as those outlined by Mittelstadt and Floridi (2016). These questions would provide students with a basic toolkit to think about ethical challenges embedded in a data science project.

We contend that faculty should revise traditional teaching approaches to contextualize ethics teaching in the context of the ongoing rapid digital transformation in global business. Emerging technologies such as AI and visualization pose unique ethical challenges that need to be addressed in business ethics curricula. We reviewed the current approaches to teaching business ethics and highlighted the deficiencies in these approaches in addressing the ethical challenges that face an increasingly technology- and data-driven world. We discussed the limits of Universalist approaches to ethics in current business environments in which decision-making is driven more and more by data, technology, and algorithms. We then outlined an approach toward a new pedagogy in business ethics that incorporates both normative and behavioral approaches while addressing the modern-day data- and model-related challenges.

Our aim in this paper was to explore the research question: How can business schools best educate students in data-related (data acquisition, processing, storing, visualizing, and reporting for decision-making) courses to ask "smart" questions by using ethical guidelines that they acquire while studying information management, Big Data, AI, visualization, and analytics? In essence, we tried to make a normative case that traditional Universalist approaches to teaching ethics are inadequate when the subject matters are Big Data and AI. We support this contention primarily by examining previous research on teaching ethics in business schools and by looking at the accreditation standard. We then report on emerging ethical approaches such as the feminist centric ethic of care and integrated social contract theory as possible alternatives to existing approaches to teaching ethics. Throughout this paper, we also attempt to situate the discussions of ethical frameworks as delivered in the classroom, with the ethical dilemmas and ethical problem solving needed in the real world. Taken together, our limited aim is to identify possibilities for improving ethics education within the context of teaching business school students about Big Data and AI.

Our paper has limitations. The exploratory nature of the paper means that we have not empirically tested or presented any statistical data. The intent was to curate literature that addresses what accredited business schools are supposed to teach in ethics classes. It is possible that business schools are already adjusting to the need to alter ethics education in the new world of Big Data and AI. We see little obvious evidence of this based on our own experiences, but, clearly, deeper and more systematic analysis of business ethics education today would be welcome additions to this exploratory work. A second limitation was that we do not attempt to explore the teaching of ethics outside the United States, although we have introduced frameworks that originate in Europe (Kuc-Czarnecka and Olczyk 2020; Floridi and Taddeo 2016; Lyytinen *et al.* 2020; Owen *et al.* 2012; Nunan and Domenico 2013), we have not considered emerging economies. We also recognize that the problems we identify and our suggested solutions may not translate outside of the context of US business schools. Future research should examine how well the frameworks presented here translate to other cultural contexts.

Many business schools are initiating new pedagogical approaches in business ethics and are attempting to address it by incorporating approaches that draw on critical thinking aligned with principles that question assumptions that reveal nonobvious ways that Big Data use incubates negative outcomes. Although this is a start in the right direction, business schools need to go further in revamping their ethics curricula in light of the challenges laid out in our research. Business schools should consider ethics education as a vital part of their curricula and not relegate it to the periphery. Further research into innovative ethics curriculum designs and innovative methods of delivering ethics education will help in revamping business ethics education for the modern world. As proposed earlier, embedding the practical examples of ethical take-aways embedded in internship projects and capstones should be explored further.

References

AACSB. 2004. "Ethics Education in Business Schools." Report of the Ethics Education Task Force to AACSB International's Board of Directors. https://www.aacsb.edu/-/media/publications/research-reports/ethics-education.pdf?la=en.

Banaji, M., Bazerman, M., and Chugh, D. 2003. How (un)ethical are you? *HBR* **81**, no. 12: 56–94.

Bazerman, M. H. 2020. "A New Model for Ethical Leadership." *Harvard Business Review*. **98**, no. 5: 90–97.

Berinato, S. 2019. "Data Science and the Art of Persuasion." *Harvard Business Review* **2019**: 126–137.

Berners-Lee, T. 2019. "I Invented the World Wide Web. Here's How We Can Fix It." *The New York Times*, November 24.

Blackburn, S., L. LaBerge, C. O'Toole, and J. Schneider. 2020. "Digital Strategy during the Coronavirus Crisis." https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-strategy-in-a-time-of-crisis.

Campbell, L., and S. Weale. 2021. "Rashford: something 'Going Wrong' with Free School Meal Deliveries." The Guardian. January 12, 2021.

Cropanzano, R., and J. Stein. 2009. "Organizational Justice and Behavioral Ethics: Promises and Prospects." *Business Ethics Quarterly* **19**, no. 2: 193–233. doi: 10.5840/beq200919211

De Cremer, D., D. Mayer, and M. Schminke. 2010. "Guest Editors' Introduction: On Understanding Ethical Behavior and Decision Making: A Behavioral Ethics Approach." *Business Ethics Quarterly* **20**, no. 1: 1–6. doi: 10.5840/beq20102012

De Los Reyes, G., T. W. Kim, and G. R. Weaver. 2017. "Teaching Ethics in Business Schools: A Conversation on Disciplinary Differences, Academic Provincialism, and the Case for Integrated Pedagogy." *Academy of Management Learning and Education* **16**, no. 2: 314–36. doi: 10.5465/amle.2014.0402

Donleavy, G. 2007. "Towards an Ethical Framework Grounded in Everyday Business Life." *Issues in Social and Environmental Accounting* **1**, no. 2: 199–216. doi: 10.22164/isea.v1i2.14

European Commission. 2019. "Ethics Guidelines for Trustworthy AI."

Evanoff, R. J. 2004. "Universalist, Relativist, and Constructivist Approaches to Intercultural Ethics." *International Journal of Intercultural Relations* **28**, no. 5: 439–58. doi: 10.1016/j.ijintrel.2004.08.002

Floridi, L., and M. Taddeo. 2016. "What is Data Ethics?." Royal Society of London,

Fung, K. 2015. "The Ethics Conversation We're Not Having About Data," Accessed January 31, 2021. https://hbr.org/2015/11/the-ethics-conversation-were-not-having-about-data.

Godwyn, M. 2015. *Ethics and Diversity in Business Management Education*. Berlin: Springer.

Goes, P. 2014. "Editor's Comments: Big Data and Is Research." *MIS Quarterly* **38**, no. 3: iii–viii.

Heggeseth, B. 2019. "Intertwining Data Ethics in Intro Stats." In *Symposium on Data Science and Statistics*. https://drive.google.com/file/d/1GXzVMpb6GVNfWPS6bd9jggtqq1C77Wsc/view.

Johnson, M., R. Jain, P. Brennan-Tonetta, E. Swartz, et al. 2021. "Impact of Big Data and Artificial Intelligence on Industry: Developing a Workforce Roadmap for a Data Driven Economy.." Global Journal of Flexible Systems Management. doi: 10.1007/s40171-021-00272-y

Kim, P. H., S. S. Wiltermuth, and D. T. Newman. 2021. "A Theory of Ethical Accounting and Its Implications for Hypocrisy in Organizations." *Academy of Management Review* **46**, no. 1: 172–91. doi: 10.5465/amr.2018.0161

Kuc-Czarnecka, M., and M. Olczyk. 2020. "How Ethics Combine with Big Data: A Bibliometric Analysis." *Humanities and Social Sciences Communications* **7**, no. 1: 1–9.

Kumar, A., T. Braud, S. Tarkoma, and P. Hui. 2020. "Trustworthy AI in the Age of Pervasive Computing and Big Data." In 2020 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2020,

Loi, M., C. Heitz, and M. Christen. 2020. "A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data." *7th Swiss Conference on Data Science* : 41–6. in

"Lord of the Rings, 2020 and Stuffed Oreos: Read the Andrew Bosworth Memo - The New York Times," The New York Times, January 2020.

Lyytinen, K., H. Topi, and J. Tang. 2020. "Interim Report of MaCuDE Curriculum Analysis."

Milligan, M. 2018. "Technology and the Ethics Gap," *ABET*. Accessed January 31, 2021. https://www.abet.org/technology-and-the-ethics-gap/.

Mittelstadt, B. D., and L. Floridi. 2016. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." In *The Ethics of Biomedical Big Data*, 445–80. Cham: Springer.

National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. doi: 10.17226/25104

Noddings, N. 2003. *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley: University of California Press,

Nunan, D., and M. Di Domenico. 2013. "Market Research and the Ethics of Big Data." *International Journal of Market Research* **55**, no. 4: 505–20. doi: 10.2501/IJMR-2013-015

Owen, R., P. Macnaghten, and J. Stilgoe. 2012. "Responsible Research and Innovation: From Science in Society to Science for Society, with Society." *Science and Public Policy* **39**, no. 6: 751–60. doi: 10.1093/scipol/scs093

Purtova, N. 2015. "The Illusion of Personal Data as No One's Property." *Law, Innovation and Technology* **7**, no. 1: 83–111. doi: 10.1080/17579961.2015.1052646

Rutherford, M. A., L. Parks, D. E. Cavazos, and C. D. White. 2012. "Business Ethics as a Required Course: Investigating the Factors Impacting the Decision to Require Ethics in the Undergraduate Business Core Curriculum." *Academy of Management Learning and Education* **11**, no. 2: 174–86. doi: 10.5465/amle.2011.0039

Saltz, J. S., and N. Dewar. 2019. "Data Science Ethical Considerations: A Systematic Literature Review and Proposed Project Framework." *Ethics and Information Technology* **21**, no. 3: 197–208. doi: 10.1007/s10676-019-09502-5

Schwab, K. 2018. "The future of humanity depends on design ethics, says Tim Wu." Accessed January 31, 2021. https://www.fastcompany.com/90239599/the-future-of-humanity-depends-on-design-ethics-says-tim-wu.

Schwartz, P. M. 2011. "Privacy, Ethics, and Analytics." *IEEE Security and Privacy Magazine* **9**, no. 3: 66–69. doi: 10.1109/MSP.2011.61

Sexton, R., and B. Garner. April 2020. "Student Perspectives of Effective Pedagogical Strategies for Teaching Ethics." *Marketing Education Review* **30**, no. 2: 132–37.

Simon, J. 2016. "Value-Sensitive Design and Responsible Research and Innovation." In *The Ethics of Technology: Methods and Approaches*, edited by Sven Ove Hansson, pp. 219–36. London, UK: Rowman & Littlefield International.

Singer, N. 2018. "Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It." *The New York Times*, February 18.

Stahl, G. K., and M. S. De Luque. 2014. "Antecedents of Responsible Leader Behavior: A Research Synthesis, Conceptual Framework, and Agenda for Future Research." *Academy of Management Perspectives* **28**, no. 3: 235–54. doi: 10.5465/amp.2013.0126

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* **42**, no. 9: 1568–80. doi: 10.1016/j.respol.2013.05.008

Taylor, L. 2016. "The Ethics of Big Data as a Public Good: which Public? Whose Good?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, no. 2083: 20160126. doi: 10.1098/rsta.2016.0126

Tenbrunsel, A. E., and K. Smith-Crowe. 2008. "13 Ethical Decision Making: Where We've Been and Where We're Going." *Academy of Management Annals* **2**, no. 1: 545–607. doi: 10.5465/19416520802211677

Treviño, L. K., and G. R. Weaver. 1994. "Business ETHICS/BUSINESS Ethics: One Field or Two?*." *Business Ethics Quarterly* **4**, no. 2: 113–28. doi: 10.2307/3857484

Treviño, L. K., G. R. Weaver, and S. J. Reynolds. 2006. "Behavioral Ethics in Organizations: A Review." *Journal of Management* **32**, no. 6: 951–990. doi: 10.1177/0149206306294258

Wu, T. 2016. *The Attention Merchants*. New York: Vintage Books, 272–273.

Zwitter, A. 2014. "Big Data Ethics." *Big Data Soc*. doi: 10.1177/2053951714559253

Journal of Big Data:
Theory and Practice

Journal of Big Data: Theory and Practice

The Journal of Big Data: Theory and Practice
publishes two issues per year and special issues on a rolling basis.
Accepted articles are made available online immediately.

http://JBDTP.org