Journal of Methods and Measurement in the Social Sciences,
Vol. 13, No.1, 15-40, 2022

Comparing human coding to two natural language processing algorithms in aspirations of people affected by Duchenne Muscular Dystrophy

Carolyn E. Schwartz
DeltaQuest Foundation
Tufts University Medical School

Roland B. Stark
DeltaQuest Foundation

Elijah Biletch
DeltaQuest Foundation

Richard B.B. Stuart
DeltaQuest Foundation

Yuelin Li
Memorial Sloan Kettering Cancer Center

Qualitative methods can enhance our understanding of constructs that have not been well portrayed and enable nuanced depiction of experience from study participants who have not been broadly studied. However, qualitative data require time and effort to train raters to achieve validity and reliability. This study compares recent advances in Natural Language Processing (NLP) models with human coding. This web-based study (N=1,253; 3,046 free-text entries, averaging 64 characters per entry) included people with Duchenne Muscular Dystrophy (DMD), their siblings, and a representative comparison group. Human raters (n=6) were trained over multiple sessions in content analysis as per a comprehensive codebook. Three prompts addressed distinct aspects of participants' aspirations. Unsupervised NLP was implemented using Latent Dirichlet Allocation (LDA), which extracts latent topics across all the free-text entries. Supervised NLP was done using a Bidirectional Encoder Representations from Transformers (BERT) model, which requires training the algorithm to recognize relevant human-coded themes across free-text entries. We compared the human-, LDA-, and BERT-coded themes. Study sample contained 286 people with DMD, 355 DMD siblings, and 997 comparison participants, age 8-69. Human coders generated 95 codes across the three prompts and had an average inter-rater reliability (Fleiss's kappa) of 0.77, with minimal rater-effect (pseudo R^2=4%). Compared to human coders, LDA does not yield easily interpretable themes. BERT correctly classified only 61-70% of the validation set. LDA and BERT required technical expertise to program and took approximately 1.15 minutes per open-text entry, compared to 1.18 minutes for human raters including training time. LDA and BERT provide potentially viable approaches to analyzing large-scale qualitative data, but both have limitations. When text entries are short, LDA yields latent topics that are hard to interpret. BERT accurately identified only about two thirds of new statements. Humans provided reliable and cost-effective coding in the web-based context. The upfront training enables BERT to process enormous quantities of text data in future work, which should examine NLP's predictive accuracy given different quantities of training data.

Key words: natural language processing, qualitative data, human, efficiency

While qualitative data collection is often used in the development of theory or conceptual models for new measures (Cappelleri et al., 2013; Ferrans, 2005), many qualitative studies utilize small sample sizes (Schwartz & Revicki, 2012), perhaps related to different logical, theoretical, and epistemological differences from quantitative research (Trotter II, 2012). There are, however, increasingly low-effort ways to collect qualitative data due to online survey engines, social media platforms, and other ways that people are asked to provide input in their clinical care. These developments dovetail with the growing interest in patient-centered measurement and care (Barry & Edgman-Levitan, 2012; Kebede, 2016), providing further motivation for expanding the feasibility and use of qualitative data in outcome research. Moreover, advances in the capability of Natural Language Processing (NLP) algorithms over the past decade have expanded their applications in medical and social science research (Agaronnik, Lindvall, El-Jawahri, He, & Iezzoni, 2020; Parker, 2020; Skaljic et al., 2019).

Early NLP algorithms extracted themes by tallying word frequencies across responses. One of the most widely used instances of this basic type of NLP software is Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001). LIWC compares words to a pre-determined dictionary file of various linguistic and psychological categories, allowing researchers to observe categorical associations between linguistic patterns and psychological state (Dönges, 2009; Pennebaker et al., 2001; Receptiviti, 2021). A 2011 study, for example, used LIWC to analyze transcripts of potential romantic partners on four-minute speed dates to measure how closely their speech matched in order to predict whether the couples would stay together after the first date (Ireland et al., 2011).

The major limitation of word-count algorithms such as LIWC is the requirement for investigators to predict words and categories relevant to the research topic for use in the algorithm's "dictionary" in order to be counted for analysis. Topic modelling algorithms from the mid-2000's such as Latent Dirichlet Analysis (LDA) and Hierarchical Latent Tree Analysis solved this problem by generating lists of abstract topics from text without the need for a "warmup" or "training" data set. Though researchers may use topics to extract information such as patient priorities and goals, the topics produced by LDA and Hierarchical Latent Tree Analysis are often unpolished and may not be relevant to the research question (Atkinson, 2019; Li, Rapkin, Atkinson, Schofield & Bochner, 2019).

Technology companies such as Facebook, Google, and OpenAI have recently developed deep learning neural network tools and made many of them open source and free to download. In this article, we applied a Bidirectional Encoder Representations from Transformers (BERT) model to classify free-text goal statements to themes (Devlin, Chang, Lee, & Toutanova, 2018). Transformers first appeared in 2017 (Vaswani et al.,

2017), when engineers at Google published a paper to address challenges in processing word sequences. For example, the two phrases "live to eat" and "eat to live" are semantic opposites due to how the words are ordered from left to right. Transformers use an attention-based structure to retain a memory of word sequences in hidden layers. This attention mechanism overcomes the limitations of LDA, which has no built-in mechanisms to distinguish word sequences (except in n-grams, sequences of words treated as unique entities, but it is a flawed approach). Word order is also ignored in machine-learning techniques such as naive bayes, Support Vector Machines (SVM), and random forest (Reyes, 2019).

In late 2018, Google released BERT (Devlin et al., 2018) which added enhancements to the attention mechanisms of Transformers. Given that generally the larger the quantities of data used to train a neural network, the more the predictive power, the immense network of data available to research groups at the technology companies have allowed for the development of these NLP techniques that more accurately assess nuanced contexts and motives in individuals' writing and speech (Mikolov, Chen, Corrado, & Dean, 2013; Tenney, Das, & Pavlick, 2019). Notably, this improved accuracy has allowed for the development of NLP systems capable of deriving clinical decisions based on automated electronic medical record analysis (Chen, Zafar, Galperin-Aizenberg, & Cook, 2018; Gonzalez-Hernandez, Sarker, O'Connor, & Savova, 2017). The primary use of NLP in social science and medical research, however, is to supersede the use of humans in assigning topic "codes" to open-text survey responses, interviews, and social media posts (Guetterman et al., 2018; Leeson, Resnick, Alexander, & Rovers, 2019).

Early articles comparing NLP to human coding were optimistic about its potential; Andrew Perrin postulated that NLP could expand the scope of qualitative studies by eliminating the need to pay and train coders and could potentially even eliminate issues regarding inter-rater reliability, though computer processing power at the time did not yet allow NLP to outpace human coders and thus limited its applications (Perrin, 2001). Accordingly, newer computing technologies have yielded promising results in certain fields; for instance, LDA topic modeling analysis of open-ended survey questions can allow for thematic information outside of a predefined coding rubric to be detected in survey responses, which serves to augment, rather than replace, the manual coding of data (Finch, Hernández Finch, McIntosh, & Braun, 2018).

Counseling psychology studies comparing NLP analysis to human coding of counselor-client conversations/motivational interviews have also found evidence that NLP techniques may be able to accurately apply a behavioral coding system on a large body of unstructured text. This may save significant time and money over a manual approach, which can range on average from 90 to 120 minutes per 20 minute interview segment, not

including the 40 or so hours required for coder training (Can et al., 2016; Moyers, Martin, Catley, Harris, & Ahluwalia, 2003). Some non-topic models lagged behind human reliability when coding certain highly contextual statements in motivational interviews; in one example, the Discrete Sentence Feature (DSF) and Recursive Neural Network (RNN) models struggled with coding isolated sentences discussing substance use. Those sentences could either be coded as favoring change in the client's habits or as the opposite (favoring maintenance of current habits), depending on subtle context clues from the preceding conversation, which human raters found easier to discern (Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016). A number of these studies express optimism about the potential speed advantage of NLP over human coders.

Baumer and colleagues compared LDA and human coding by grounded theory (Baumer, Mimno, Guha, Quan, & Gay, 2017). BERT was not yet available at the time of their application. They analyzed free-text data on reasons why individuals returned to social media after a brief (up to 99 days) and voluntary absence. Baumer et al. (2017) report good agreement, that LDA extracts themes that generally reflect the same content as the human-extracted themes.

In response to a dearth of literature (Raffel et al., 2019), the present study directly compared human coding to two NLP methods: the un-supervised LDA and the supervised BERT. One of the co-authors (YL) has previously applied LDA to summarize cancer patients' free-text goal statements as they undergo bladder cancer surgery (Landis & Koch, 1977). Thus, the main rationale for these two specific NLP methods is to go beyond LDA to capitalize on the latest NLP analytics.

Methods

Sample and Procedure

This secondary analysis utilized data from a study of Duchenne Muscular Dystrophy (DMD) patients, their siblings, and a comparison-group. The sample and methods are fully described in the two primary papers from this project (Schwartz, Biletch, Stuart, & Rapkin, 2022a, 2022b) and will only briefly summarized herein. These primary papers examine differences in aspirations for patients versus comparison participants, and siblings versus comparison participants. Accordingly, for the purpose of the present work, data were combined across groups, although demographic characteristics will be described by group. Eligible participants were age 8 or older and able to complete an online questionnaire.

The web-based survey was administered October through December 2020 through the Health Insurance Portability and Accountability Act of

1996 (HIPAA)-compliant, secure Alchemer engine (www.alchemer.com). Participants were paid honoraria to compensate them for their time. The protocol was reviewed and approved by the New England Independent Review Board (NEIRB #20203038), and all participants provided informed consent before beginning the survey.

Measures

Life aspirations was measured using the following open-ended prompts: (1) Three Wishes (Nereo & Hinton, 2003), in which participants were asked, "If you could make three wishes, any three wishes in the whole world, what would they be?"; (2) Goals: "What are the main things you want to accomplish?"; (3) Quality of Life (QOL) Definition: "In a sentence, what does the phrase "Quality of Life" mean to you at this time?" The latter two are part of the QOL Appraisal Profile$_{v1}$(QOLAP) (Rapkin & Schwartz, 2004).

Demographic Characteristics included year of birth, gender, and whether anyone in the household was or had been infected with the novel coronavirus-2019, and whether they received help completing the survey (all participants). Teens and adults were asked about comorbidities from a list selected on the basis of documented higher prevalence in people with DMD (Ciafaloni et al., 2009; Pane et al., 2012). Adult participants were asked about race, ethnicity, education, marital status, weight, height, with whom the person lives, difficulty paying bills, and employment status.

Statistical Analysis

Coding open-text data. The open-ended data were coded by six trained raters (EB, RBB, AD, JBL, EK, MCF), according to a standardized protocol and comprehensive codebook originally derived from an extensive sorting procedure (Li & Rapkin, 2009). [The interested reader can contact the corresponding author for the QOLAP coding manual which describes the theme definitions in detail.]

Themes were coded as "0" if they were not reflected in the individual's written text response, and "1" if they were reflected there. As the goal-delineation themes were originally developed with a Human Immunodeficiency Virus sample (Li & Rapkin, 2009), which generally has different sociodemographic characteristics than the current study sample, some themes were not as prevalent among the present sample. For example, themes related to drug and alcohol, immigration, and racism were prevalent among the Human Immunodeficiency Virus sample, but were not found at all in the current study sample. Themes were added as needed, resulting in a set of 40 themes for the Wishes and Goals prompts and 17 for the QOL Definition prompt. For each prompt, a theme of "no direct answer" was used if the respondent did not provide an answer or answering a different

question than was asked. For example, in response to the question "What are the main things you want to accomplish?" exemplary No-Direct-Answer responses "seems rather great" or "nothing idk lol."

Each text entry could be coded for as many themes as there. Thus, one goal could elicit one theme or more than one depending on how the individual worded it. For example, one individual's Accomplish goal was "My bills paid, my family healthy and happy, and family go to church". It was coded as reflecting family welfare, financial concerns, health issues, mental health/mood state, and religious/spiritual concerns. In contrast, another individual's Accomplish goal was "Move to a different state," Which was coded only as living situation. In this method of working with the aspirations data, we assumed that the relevant factor was the themes, not the different wishes, goals, or QOL definitions themselves.

Training took place in two multi-hour sessions to understand the protocol and to utilize fully the codebook where themes were described fully and exemplified. Raters coded an initial set of ten participants' data (all prompts), followed by a discussion of differences across raters. They then coded the next ten participants' data (all prompts), and comparison and discussion revealed almost no differences across raters. Raters then coded data from 40 more responses (all prompts), from which inter-rater reliability was computed in two ways on the 240 test responses (6 raters * 40 participant entries).

Inter-Rater Reliability. Fleiss's kappa (Fleiss, 1971) assessed degree of agreement over and above what would be expected by chance. This variant on the more familiar Cohen's kappa (Cohen, 1960) is used in cases of more than two raters. While there are no generally accepted rules of thumb for a desirable level of either form of kappa, some healthcare researchers have proposed values from 0.41-0.60 as "moderate," 0.61-0.80 as "good," and 0.81-1.00 as "very good."(Altman, 1999; Landis & Koch, 1977).

Logistic regression assessed level of agreement among raters, with each of 240 "0" or "1" values regressed on the Rater variable, with its six rater-categories. High inter-rater reliability (IRR) for any given theme would be indicated by a nonsignificant rater effect, and one that explained a low fraction of the variance in ratings (e.g., a pseudo-R-squared in the low single digits).

NLP Methods Tested

Two NLP methods were tested in this study: LDA and BERT. The main difference between these methods is that LDA is unsupervised, and BERT is supervised machine learning, in the sense that LDA is able to extract topics without human intervention while BERT (in text classification specifically) requires that topics be previously established. A crude but useful analogy may be that LDA behaves more like Exploratory Factor

Analysis, where the underlying factors are unknown, while BERT behaves more like Confirmatory Factor Analysis, where those factors are specified in advance.

LDA. The LDA analytic plan was similar to the one described in detail in a previous article on patients' free-text goal statements as they undergo bladder cancer surgery.(Li, et al., 2019) Separate LDA analyses were conducted for responses to each of three prompts: Wishes, QOL Definitions, and Goals. We followed the commonly-used steps in preprocessing (e.g., plotting 'word clouds', setting 'stop words' aside, and adding two-consecutive-word phrases as 'bigrams' for contextual information). We then determined the best number of topics as specified by LDA and fitted the final LDA model for each analysis. The LDA computation was primarily done by the scikit-learn tools written in the Python programming language (Pedregosa et al., 2011). The number of topics per analysis was evaluated by the R package ldatuning (Nikita, 2016) and the four supported metrics (Arun, Suresh, Veni Madhavan, & Murthy, 2010; Cao, Xia, Li, Zhang, & Tang, 2009; Deveaud, SanJuan, & Bellot, 2014; Griffiths & Steyvers, 2004), using all available text entries. The LDA analysis, unlike that for BERT, involved no evaluation of accuracy, as in use of a training set versus validation set.

Model selection was done using the four metrics provided in the ldatuning package (Nikita, 2016) to estimate the desired number of topics. Both the Arun et al. (2010) and Cao et al. (2009) metrics are akin to the scree plot in an exploratory factor analysis, where the location of the elbow indicates the desired number of topics. The Griffiths and Steyvers' (2004) and the Deveaud (2014) metric are based on the fit between words within topics, where the location of a plateau reflects the desired number of topics.

All subsequent analyses were fixed at this number of topics to make a consistent and streamlined presentation, including separate LDA models for patients, siblings, and comparison-group participants.

Bidirectional Encoder Representations from Transformers (BERT). BERT is widely viewed as a state-of-the-art, supervised deep-learning neural network. It was developed by scientists at Google (Devlin, Chang, Lee, & Toutanova, 2019) to address enduring challenges in NLP. Transformers such as BERT use an attention-based structure to retain a memory of word sequences in hidden layers of a neural network such that the network registers or "intuitively understands" their opposite semantic meaning. This property overcomes certain limitations of LDA, which has no built-in mechanisms to distinguish word sequences (except in n-grams, such as the bigram in the current LDA approach, an unsatisfactory workaround nevertheless).

To classify text using BERT. we used a publicly accessible, off-the-shelf machine-learning tool called the "huggingface transformers" (Hugging Face, 2021). The specific tool we used was the DistillBERT tool within the

huggingface transformers library. This a scaled-down version of the full BERT was designed to work more quickly due to fewer layers and hidden nodes. It is one of several alternative algorithms derived from the full BERT technology (see ("List of alternative algorithms derived from the full BERT technology,")). DistillBERT is what is known as a *pre-trained* model, in which technology companies have already trained it using the enormous amounts of unannotated text on the internet so that it learns a general-purpose language representation model (Devlin & Chang, 2018). After pre-training, an analyst can then fine-tune DistillBERT for specific tasks. Henceforth, for simplicity and readability we use the more generic term BERT to represent DistillBERT.

From a user's perspective, an application of BERT is divided into two components, known in the literature as *pre-training* and *fine-tuning*. This two-step approach is at the core of the concept of Transfer Learning(Vaswani et al., 2017). Once pre-trained, BERT and its variants can be reused for many downstream machine-learning tasks, including the current text classification. There are many pre-trained libraries available for download, for tasks such as next-sentence prediction (e.g., instant autocomplete suggestions in a search engine), named-entity recognition (e.g., a trained network knows that the Empire State Building is near Manhattan), and language translation (e.g., English to French).

The fine-tuning in this study proceeded as follows. Wishes, goals and definitions were analyzed separately. For example, the 1,613 entries of wishes were randomly divided into the training set (n=399), the validation set (n=76 for tuning configuration parameters), and a blinded test set (remaining n=1,214 that BERT had never encountered previously and blinded to the analyst who trained BERT).

Configuration Parameters for BERT. The training set entries were entered into BERT as the predictors and the corresponding human-coded categories were the target outcomes. Learning was achieved by optimizing network connections by the Adaptive Moment Estimation algorithm (Kingma & Ba, 2014). It is known that optimized network configurations are affected by hyperparameters such as the learning rate (the rate with which model weights are updated in response to the estimated error, where a learning rate too small may run slowly but a learning rate too large may lead to suboptimal weights), batch size (number of samples that are passed to the network at once, where smaller batch size facilitates learning but tends to run slower), and the number of epochs (one complete presentation of the entire training data to the network during the training process is called an epoch, an iteration, or one training cycle (Hakin, 1998)). We explored configuration settings by varying combinations of batch size (16 vs. 32), learning rate (5e-5, 3e-5, and 2e-5), and number of epochs (10 vs. 20) and used the validation set to tune the optimal hyperparameter settings that

yielded the best overall validation accuracy, which produced the final settings of a batch size of 16, a learning rate of 3e-5, and 10 epochs.

The trained model was then evaluated by the blinded test set (e.g., 1,214 blinded wishes) that the trained BERT model had never encountered before. BERT was analyzed using the Python programming language version 3.8.10 to call the transformers library version 4.11.3 and tensorflow 2.6.0 (details on software platform are available upon request).

Performance of BERT by Predictive Accuracy. Accuracy was evaluated using both *improper scoring* (the percentage of cases correctly classified) and *proper scoring* (the average point-biserial correlation [r_{pb}] between a given human-rated theme' binary value and the BERT-generated probability of a text entry fitting that theme). In the latter case, the average r_{pb} was obtained via Fisher's Z_r statistics (Harrell, 2010; "Scoring rule," 2021). For each prompt, there was a subset of themes with nonzero probabilities generated by BERT and that thus could be tested for their point-biserial correlations with the corresponding binary theme variable as rated by the human coders. Compared to Correct Classification Rate, this correlation constitutes a more finely grained method of evaluating BERT's performance. Even when correct classifications were not made, it would be evidence in BERT's favor if there were a systematic tendency for the probability of being rated with a given theme to be higher in the presence of that theme.

Results

Participant characteristics

The sample included 1253 participants: 285 patients, 349 of their siblings, and 619 in the comparison group (mean age 17, 18, and 19, respectively). The patients were all male, while males made up 48% and 47% of the other 2 groups. Participants resided in a broad cross-section of the United States. One percent of patients, 5% of the siblings and 23% of the comparison group were married. Percentages of Hispanics or Latinos were 9%, 8%, and 20%; percentages of Blacks, 8%, 6%, and 20%. Among patients, 5% were employed, and the rest were unemployed or disabled; in contrast, 42% of the siblings and 61% of the comparison group were employed. Educational levels were varied, with the comparison group having the highest fraction (37%) educated at the bachelor's level or higher. Only 1% of patients or of siblings, but 19% of the comparison group, reported that they or a family member had contracted COVID-19. Comparatively large numbers of participants in all groups reported having help completing the survey: by group, 49%, 26%, and 19%, respectively. Further information is available in the primary publications from this study (Schwartz et al., submitted for publication a, submitted for publication b)

Qualitative Coding Reliability

As reported in the primary papers from this project (Schwartz et al., under review a, b), the mean kappa was 0.77 (*SD*=0.17, range 0.51 to 0.98), reflecting a good level of agreement(Altman, 1999; Landis & Koch, 1977). The best estimated pseudo-R^2 for rater was 0.042 (p=0.24), suggesting that the rater effect in coded themes was negligible. Descriptive statistics on proportion of participants whose open-text data reflected various themes are provided in the primary papers from this project.

Examples of Participants' Wishes

Table 1
Illustrative Examples of Free-Text Entries

Role	Wishes	Definitions	Goals
Patient	Always have a dog, No disease in world, peace	Living life without pain	Going to every baseball stadium. And making a lot of friends
Sibling	I want to be an English teacher. I want to live in a big city. I want to find a partner who loves me very much	Money was plentiful	Doing a degree
Comparison	One wish would be go to Taylor swift next tour. The second, be a millionaire. And the third, have all of taylor swift's merch.	Quality of life to me means having lived your life in a way that you are proud and know that everything in it was worth it even though it did not seem like it. Also allow yourself to make mistakes and learn and always come back stronger than a 90's trend!	The main things i want to accomplish is get a bachelor's degree in science. Go to the next taylor swift tour, and finally travel the world and see as many of my favorite artists live.

Table 1 provides illustrative examples of free-text entries on what participants wished for, from the combined total of 1214 unique wishes, 480 goals, and 243 definitions randomized into the validation set.

Human Coding Results

Table 2 provides information about the prevalence across the whole sample of Wishes, Goals, and QOL Definition themes. Figure 1 shows the five most prevalent themes by prompt. Financial concerns were prominent across all three

Figure 1. Top five human-coded themes by prompt. The three prompts generated relatively distinct sets of themes, although financial concerns were prominent across all three prompts, and health across two of the three.

Table 2
Descriptive statistics for human coded themes from open-text prompts, listed from most to least prevalent

Theme	Proportion of sample
Wishes - Family Welfare	.29
Wishes - Health Issues	.29
Wishes - Financial Concerns	.28
Wishes - Material Acquisitions	.26
Wishes - Interpersonal Relationships	.21
Wishes - Travel	.18
Wishes - Work and Unemployment	.17
Wishes - Achievement	.16
Wishes - DMD-Related Goals	.16
Wishes - Societal and Altruistic Concerns	.16
Wishes - Fantasy	.13
Wishes - Leisure Activities	.13
Wishes - Self-Image and Personality	.09
Wishes - Mental Health and Mood State	.09
Wishes - Education	.09
Wishes - COVID-Specific	.08
Wishes - Living Situation, Housing, Neighborhood	.06
Wishes - Health Welfare (Societal)	.06
Wishes - Independent Functioning	.05
Wishes - No Direct Answer	.05
Wishes - Existential Concerns	.03
Wishes - Political Welfare	.02
Wishes - Religious and Spiritual Concerns	.02
Wishes - Accomplishing Chores and Tasks	.02
Wishes - Financial Welfare (Societal)	.02
Wishes - Provider- and Treatment- Related Concerns	.01
Wishes - Problem Resolution	.01
Wishes - Racism	.01
Wishes - Prevention	.01
Wishes - Environmental Welfare	.01
Wishes - Living Situation (Societal)	.01
Wishes - Community Involvement and Voluntarism	.01
Wishes - Disengagement	.00
Wishes - Maintenance	.00
Wishes - Legal and Crime / Safety Concerns	.00
Wishes - Legal and Crime (Societal)	.00
Wishes - Acceptance	.00
Wishes - Drug and Alcohol Use	.00
Wishes - Immigration and Citizenship	.00

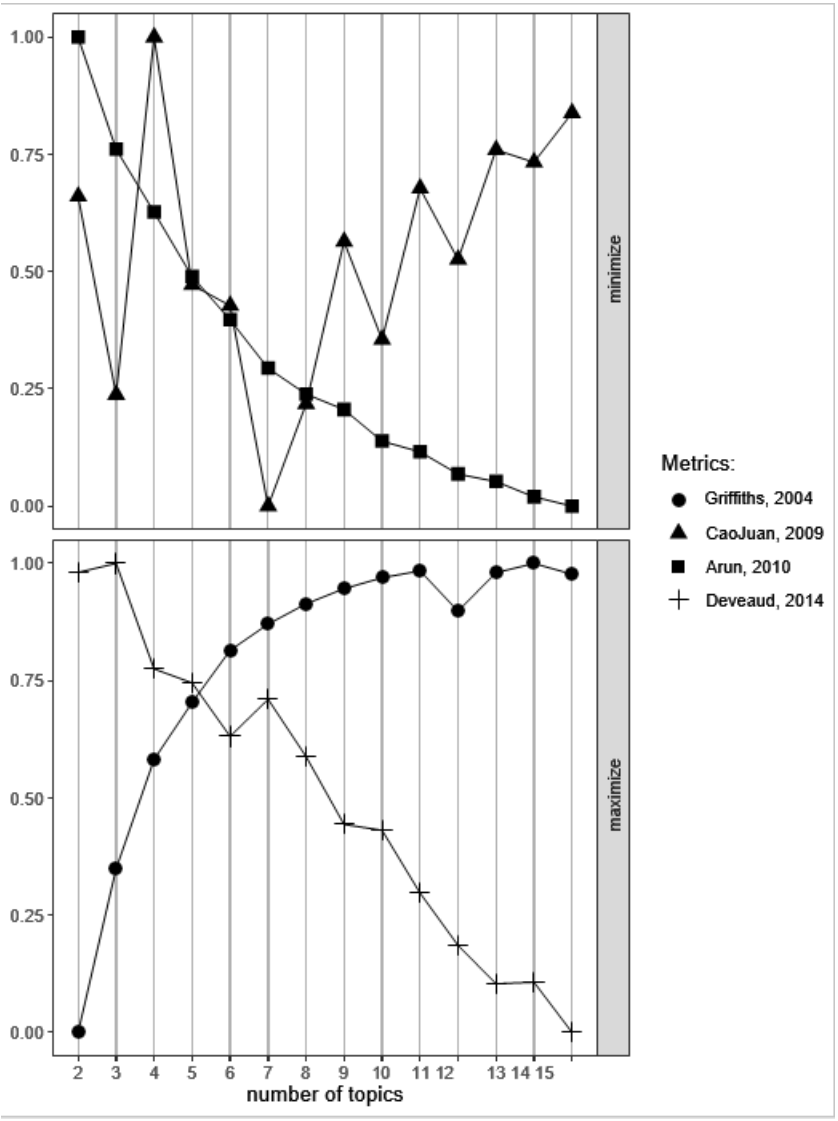
Theme	Proportion of sample
Wishes - Involvement in Community Outreach	.00
Goals - Education	.30
Goals - Work and Unemployment	.27
Goals - Achievement	.23
Goals - Financial Concerns	.23
Goals - Interpersonal Relationships	.20
Goals - No Direct Answer	.19
Goals - Family Welfare	.11
Goals - Mental Health and Mood State	.09
Goals - Living Situation, Housing, Neighborhood	.09
Goals - Health Issues	.07
Goals - Independent Functioning	.07
Goals - Material Acquisitions	.06
Goals - Self-Image and Personality	.06
Goals - Provider- and Treatment- Related Concerns	.04
Goals - Travel	.04
Goals - Societal and Altruistic Concerns	.03
Goals - Community Involvement and Voluntarism	.02
Goals - Accomplishing Chores and Tasks	.02
Goals - Leisure Activities	.02
Goals - Religious and Spiritual Concerns	.02
Goals - Existential Concerns	.02
Goals - DMD-Related Goals	.02
Goals - Acceptance	.01
Goals - Environmental Welfare	.01
Goals - Fantasy	.01
Goals - Health Welfare (Societal)	.01
Goals - Maintenance	.01
Goals - COVID-Specific	.00
Goals - Drug and Alcohol Use	.00
Goals - Prevention	.00
Goals - Disengagement	.00
Goals - Financial Welfare (Societal)	.00
Goals - Immigration and Citizenship	.00
Goals - Involvement in Community Outreach	.00
Goals - Legal and Crime (Societal)	.00
Goals - Legal and Crime / Safety Concerns	.00
Goals - Living Situation (Societal)	.00
Goals - Political Welfare	.00
Goals - Problem Resolution	.00
Goals - Racism	.00

Theme	Proportion of sample
Definition - Circumstances	.44
Definition - Contentment	.35
Definition - Positive Attitude (mental health)	.26
Definition - No Direct Answer	.22
Definition - Health	.20
Definition - Independence	.10
Definition - Personal Growth	.09
Definition - Family / friends	.08
Definition - Contribution	.02
Definition - Treatment-related	.02
Definition - Balance	.01
Definition - Survival	.01
Definition - Problems	.01
Definition - Provider-related	.01
Definition - Reminiscence	.00

LDA Results

How many topics in LDA?. Figure 2 plots the four model-selection metrics by the number of extracted topics. These four metrics provided limited guidance because of their inconsistency. Among the two metrics for which lower score indicates best fit, the Cao et al. (2009) metric suggested either 2 or 9 topics, and the Arun et al. (2010) suggested 10 to15. Among the metrics for which higher score indicates best fit, the Griffiths and Steyvers (2004) metric indicated a model with approximately 8 topics, and the Deveaud et al. (2014) clearly indicated two. We opted to retain 8 as a compromise between the 2 extremes.

Figure 2. Model-selection metrics by the number of extracted topics. Model selection metrics were used to estimate the desired number of topics, using the combined 1,253 statements of validation-set wishes. The top panel plots two metrics that theoretically should behave like a scree plot in an exploratory factor analysis, where the location of the elbow indicates the desired number of topics. Among the two metrics for which lower score indicates best fit, the Cao et al. metric suggested either 2 or 9 topics, and the Arun et al. suggested 10 to15. Among the metrics for which higher score indicates best fit, the Griffiths and Steyvers metric indicated a model with approximately 8 topics, and the Deveaud et al. clearly indicated two.



Topics. Table 3 summarizes the 3 words most strongly associated with each latent topic derived from LDA analysis of respondents' wishes, goals, and QOL definitions. These words inform the interpretation of those topics.

Table 3
Top 3 Words per Latent Topic Derived from LDA Analysis of Respondents' Wishes, Goals and QOL Definitions

Topic						
Wishes	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	like	.08	animals	.04	Walk	.04
2	live	.03	Life	.02	long	.02
3	health	.08	family	.04	love	.03
4	money	.04	DMD	.03	lots	.02
5	world	.07	peace	.04	end	.04
6	travel	.04	brother	.03	money	.02
7	new	.07	house	.05	buy	.03
8	money	.06	healthy	.06	happy	.05
Goals	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	job	.08	business	.04	married	.04
2	debt	.09	Pay	.04	live	.03
3	house	.06	Like	.04	school	.04
4	doctor	.04	Arrangement	.03	doctor arrangement	.03
5	financially	.06	things	.05	stable	.05
6	life	.07	good	.05	family	.03
7	money	.07	happy	.05	make	.05
8	work	.09	degree	.05	study	.04
QOL Definition	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	work	.07	balance	.06	know	.05
2	good	.18	quality	.03	good health	.03
3	happy	.12	Day	.06	healthy	.05
4	material	.11	spiritual	.07	satisfaction	.05
5	able	.09	want	.08	rich	.07
6	living	.13	Live	.11	fullest	.03
7	healthy	.12	happiness	.06	body	.06
8	quality	.08	things	.04	family	.04

For example, Wishes topic 1 includes “like”, “animals”, and “walk”. This topic does not lend itself to easy summary. Topic 2 seems to be related chiefly to living a long life; topic 3, to good health, family, and love; topic 4, to wealth and (finding a cure for) DMD; topic 5, to world peace; and topic 6, to travel, their brother, and money; topic 7, to worldly possessions; and topic 8 to money, health, and happiness. The fact that even these top three words by topic represent at most 8% of the corresponding text entries, and typically only 4%, makes most of these characterizations tenuous.

For respondents’ goals and QOL definition, the topics do lend themselves to more easy summary. For goals, the eight topics may be loosely characterized, respectively, as ‘finishing school and starting life’, ‘resolving financial debt’, ‘good housing and school’, ‘managing healthcare’, ‘financially stable’, ‘family happiness’, ‘career success’, and ‘college and job prospects’. These top three words by topic represent at most 9% of the corresponding text entries, and on average about 5%.

For QOL definitions, the eight topics may be summarized as ‘work-family balance’, ‘having good health’, ‘happiness & health’, ‘material & spiritual satisfaction’, ‘material wealth’, ‘living life to the fullest’, ‘healthy body’, and ‘provision for family’. These top three words by topic represent at most 18% of the corresponding text entries, and on average about 7%.

BERT Results

Improper Scoring: Correct Classification Rate. Table 4 provides example texts for the goals prompt and shows BERT probabilities for assigning the top five human-coded themes. Grey shading indicates that BERT correctly classified the statement as matching the indicated theme.

Table 4
Examples of BERT's Probabilities that a Given Statement Will Match a Given Theme

Example Text, Goals Prompt	Education	Work & Unemployment	Achievement	Financial Concerns	Interpersonal Relationships
Raising my children to be respectful.					
Spend as much time with family as possible.	.001	.002	.003	.002	.950 [a]
Live well make a lot of money and retire in asia	.005	.609 [a]	.007	.030	.009
Making a difference, leaving a mark, and achieving my goals	.009	.009	.047	.011	.091
Strive to learn new knowledge	.450	.050	.114	.008	.052

[a] Gray shading that BERT correctly classified the statement as matching the indicated theme.

While BERT correctly identified two themes, it missed others that would have been recognizable as related to one or more themes. For example, "achieving my goals" would have been coded as Achievement by humans but only had a 4.7% probability of such by BERT. Similarly, "strive to learn new knowledge" would have been coded as Education by humans but only had a 45% probability of such by BERT.

Table 5 summarizes the overall accuracy in BERT's predictions for text entries in the validation set, i.e., data that the model had never encountered previously. In the blinded validation set the theme identified by BERT was also identified by humans for 70% of Wishes, 68% of Goals, and 61% of QOL Definition entries, with an overall correct classification rate of 67%. This is despite the fact that BERT could be described as having in most cases "more than one chance." That is, the average statement was rated by human coders

as fitting 2.9 themes for Wishes, 2.2 for Goals, and 1.6 for Definitions. BERT thus typically had multiple ways, an average of 2.6, in which its classification could conceivably match some human-coded theme.

Table 5
DistillBERT's Predictive Accuracy

		Training Set		Blinded Validation Set	
Prompt	*n* codes	*n* entries	Accuracy	*n* entries	Accuracy
Wishes	40	399	100%	1214	70%
Goals	40	160	100%	480	68%
QOL Definition	15	139	100%	545	61%

Proper Scoring: Human-BERT Correlation. Table 6 shows the average correlations among themes coded by humans and BERT, separately by prompt. For themes within all three prompts, the algorithm's probabilities generally correlated only moderately with the binary theme variable, with average r_{pb} per prompt in the 0.3-0.4 range and an overall r_{pb} of 0.34 (Table 6). These correlations reflect a relatively low overall explained variance of 0.12, with more variance explained for Goals (R^2=0.14) than for Wishes (R^2=0.11) or QOL Definition (R^2=0.10).

Table 6
Correlations Among Themes Coded by Humans vs. BERT

Prompt	*n* Comparisons	Mean r_{pb}*	Minimum r_{pb}*	Maximum r_{pb}*
Wishes (*n* =1,207)	30	0.33	-0.01	0.85
Goals (*n* = 478)	21	0.37	-0.01	0.70
Definitions (*n* = 243)	11	0.32	0.06	0.57
Total (weighted by *n* Comparisons)	62	0.34	-0.01	0.85

Note. r_{pb} = point-biserial correlation coefficient

Relative Efficiency. Considering all of the time needed for training and scoring the open-text data, the three methods took very similar amounts of time. LDA and BERT took approximately 1.15 minutes per training sample (on a 64-bit workstation with a 6-core Intel Xeon CPU at 2.40 GHz and 32 GiB of memory running Ubuntu Linux version 20.04, no GPU was utilized). By comparison, human raters can code one entry at an average rate of 1.18 minutes. After removing time for training and programming, LDA took about 8 seconds per entry, and BERT took 4. After removing time for training, the human raters took about 52 seconds per entry.

Discussion

The present study is, to our knowledge, the first to compare human coding to two NLP methods - LDA and BERT - for analyzing large-scale qualitative data. Table 7 summarizes the features of the three methods. Compared to human coders, LDA in this study did not yield easily interpretable themes. LDA output is difficult to summarize in meaningful ways, partially because the same word, phrase, or theme can appear multiple times across latent topics. BERT has the potential to be more useful because it can be trained to recognize topics or themes already deemed meaningful by humans. Nonetheless, BERT accurately identified only about two thirds of statements that it had never encountered previously in training, despite having on average 2.6 themes that humans had coded for any given text entry. Moreover, the more sensitive point-biserial correlation showed an average explained variance of 12% per theme. Because LDA and BERT require specialized knowledge and software, their feasibility and accessibility may be limited for researchers without such access.

Table 7
Summary of Text-Analysis Methods

		Method	
Feature	Humans	LDA	BERT
Yields interpretable themes	√		
Training required	√		√
High hourly cost		√	√
Specialized knowledge required	(√)	√	√
Special Software required	(√)	√	√
Scalable to big data (n>100K)		√	√

Our findings on LDA are different from Baumer et al.'s (2017) results and from the impressive results found in the wider literature on LDA, in which LDA is able to extract coherent and meaningful themes. The seminal paper on LDA (Blei, Ng, & Jordan, 2003) showed that LDA extracted meaningful and unique topics from over 16 thousand newswire articles. LDA also successfully found themes from over 40 thousand entries of chapter-length reading materials for students (Steyvers & Griffiths, 2007) or scientific abstracts (Griffiths & Steyvers, 2004). Like any statistical procedure, LDA's performance depends on the contents in the input data. Our findings suggest that LDA does not perform well in the context of relatively brief open-text entries.

Other researchers may get different results if BERT is applied after a much larger training set (i.e., longer open-text entries, far fewer themes, and many more entries per theme). For example, Murarka, Radhakrishnan,

& Ravichandran (2020) analyzed 17,000 social-media posts and achieved 80% accuracy in classifying posts into one of five specific mental-health outcomes. In contrast, our data derived from three relatively broad prompts about wishes, QOL definition, and goals, and were human-coded into 95 themes. This is a more complex task that may draw on empathy and life experience. One other limitation of BERT in text classification is that it requires training of human coders to generate coded data that can be used to train BERT. Thus, the highest cost of human coders (i.e., the training and adjudication period) would need to be included in the overall cost of BERT.

It is worth noting that any successful implementation of BERT could be reused once trained. In our case, for example, if it had a greater accuracy (e.g., > 80% similar to (Murarka et al., 2020)), our BERT model could have been applied to classify the wishes and aspirations of people who post online about Muscular Dystrophy. Also, because our data include a comparison group, the neural-network weights devised might be applied to understand the aspirations of individuals from the general population. This reusability may offset the initial cost of training BERT.

Humans provided reliable, valid, and cost-effective coding in the web-based context with relatively short text entries. On average, they took only two seconds longer than LDA or BERT per open-text entry. Of note, the present study included coding of approximately 3,000 open-text entries. Thus, scaling up to larger data sets and longer text entries might be feasible for motivated and compensated human coders. We have not evaluated the three methods in processing other qualitative data such as interview transcripts. Future research might compare the three methods in the context of hour-long interview transcripts, where BERT's advantages may be more apparent.

This study has many advantages, including a robust sample with good quality data on multiple prompts. Nonetheless, the limitations of the study must be acknowledged. First, there is considerable uncertainty in the LDA results, as seen in the unexpected patterns in two of the four model-selection metrics. Also, the current BERT model only predicts one code at a time, even though it is capable of predicting multiple categories. This was a crude but reasonable and practical beginning of this line of inquiry. Future research should examine LDA's results and interpretability to provide guidance as to when the method is most appropriate. Future BERT modeling can go onto multi-class task. Another limitation of the present work is that the computed correlations between BERT and humans are likely attenuated by the continuous-binary pairing. BERT's average r_{pb} of 0.34 would thus likely be somewhat larger, and would translate to more than our documented 12% explained variance. However, even if that 12% were tripled, it would not seem enough to justify replacing human raters with this algorithm.

Conclusions

In summary, LDA and BERT provide potentially viable approaches to analyzing large-scale qualitative data, but both have limitations. When text entries are short, LDA yields latent topics that are hard to interpret. BERT accurately identified only about two thirds of new statements even given multiple opportunities. Moreover, the probabilities it assigned showed unsatisfactory correlations with the binary theme variables in question. Humans provided reliable and cost-effective coding in the web-based context. Future research should examine NLP's predictive accuracy given different contexts and quantities of training data.

Author Note. Corresponding author: Carolyn Schwartz, Sc.D., DeltaQuest Foundation, Inc., Concord, MA, email: carolyn.schwartz@deltaquest.org.
Acknowledgments. We are grateful to the participants themselves who provided data for this project; to Alexander Denby, Julia B. Lewis, Emily Kane, Matyas Csiki-Fejer, for their careful efforts coding the open-text data; and to Sarepta Therapeutics for support of data collection for this work.

References

Agaronnik, N., Lindvall, C., El-Jawahri, A., He, W., & Iezzoni, L. (2020). Use of natural language processing to assess frequency of functional status documentation for patients newly diagnosed with colorectal cancer. *JAMA Oncology, 6*, 1628-1630.

Altman, D. G. (1999). *Practical statistics for medical research*. New York: Chapman & Hall/CRC Press.

Arun, R., Suresh, V., Veni Madhavan, C. E., & Murthy, N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining*. (pp. 391 - 402). Heidelberg: Springer Berlin.

Atkinson, T. M. (2018). *Latent dirichlet allocation in discovering goals in patients undergoing bladder cancer surgery*. Paper presented at the Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics.

Barry, M. J., & Edgman-Levitan, S. (2012). Shared decision making—The pinnacle patient-centered care. *New England Journal of Medicine, 366*, 780-781. doi:10.1056/NEJMp1109283

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology, 68*, 1397-1410.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Can, D., Marín, R., Georgiou, P., Imel, Z., Atkins, D., & Narayanan, S. (2016). "It sounds like…": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology, 63,* 343-350.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing - 16th European Symposium on Artificial Neural Networks, 72*, 1775 - 1781. doi:10.1016/j.neucom.2008.06.011

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). Development of a patient-reported outcome. In *Patient-reported Outcomes: Measurement, Implementation and Interpretation* (pp. 21-29). Boca Raton, FL: CRC Press.

Chen, P.-H., Zafar, H., Galperin-Aizenberg, M., & Cook, T. (2018). Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of Digital Imaging, 31*, 178-184. doi:10.1007/s10278-017-0027-x

Ciafaloni, E., Fox, D. J., Pandya, S., Westfield, C. P., Puzhankara, S., Romitti, P. A., . . . Miller, L. A. (2009). Delayed diagnosis in Duchenne Muscular Dystrophy: Data from the Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STARnet). *The Journal of Pediatrics, 155*, 380-385.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Deveaud, R., SanJuan, É., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*, 61-84. doi:10.3166/dn.17.1.61-84

Devlin, J., & Chang, M.-W. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural language processing. *Google AI Blog, 2*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dönges, J. (2009). What Your Choice of Words Says about Your Personality. *Scientific American Mind 20*, 4, 14-15.

Ferrans, C. E. (2005). Definitions and conceptual models of quality of life. In J. Lipscomb, C. D. Gotay, & C. Snyder (Eds.), *Outcomes Assessment in Cancer: Measures, Methods, and Applications* (pp. 14-30). Cambridge, UK: Cambridge University Press.

Finch, W. H., Hernández Finch, M., McIntosh, C., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science, 4*, 403-424.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the patient's perspective: A review of advances in natural language processing of health-related text. *Yearbook of Medical Informatics, 26*, 214-227. doi:10.15265/IY-2017-029

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

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research, 20*, e9702.

Hakin, S., (1998). *Neural networks: A comprehensive foundation* (Second ed.). Hoboken, NJ: Prentice Hall PTR.

Harrell, F. E. (2010). *Regression modeling strategies*. New York: Springer.

Hugging Face. (2021). The AI community building the future. *URL* https://huggingface.co

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science, 22*, 39-44. doi:10.1177/0956797610392928

Kebede, S. (2016). Ask patients "What matters to you?" rather than "What's the matter?". *BMJ, 354*.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural language processing (NLP) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods, 18,* 1-9.

Li, Y., Rapkin, B., Atkinson, T. M., Schofield, E., & Bochner, B. H. (2019). Leveraging latent dirichlet allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. *Quality of Life Research, 28*, 1441-1455. doi:10.1007/s11136-019-02132-w

Li, Y., & Rapkin, B. D. (2009). Classification and regression tree analysis to identify complex cognitive paths underlying quality of life response shifts: A study of individuals living with HIV/AIDS. *Journal of Clinical Epidemiology, 62*, 1138-1147.

Hugging Face (2021). List of alternative algorithms derived from the full BERT technology.

URL https://huggingface.co/transformers/pretrained_models.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moyers, T., Martin, T., Catley, D., Harris, K. J., & Ahluwalia, J. S. (2003). Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy, 31*, 177.

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2020). Detection and classification of mental illnesses on social media using RoBERTa. *arXiv preprint arXiv:2011.11226*.

Nereo, N. E., & Hinton, V. J. (2003). Three wishes and psychological functioning in boys with Duchenne muscular dystrophy. *Journal of developmental and behavioral pediatrics: JDBP, 24*, 96.

Nikita, M. (2016). ldatuning: Tuning of the latent dirichlet allocation models parameters. In: R package version 0.2.0.

Pane, M., Lombardo, M. E., Alfieri, P., D'Amico, A., Bianco, F., Vasco, G., . . . Ricotti, V. (2012). Attention deficit hyperactivity disorder and cognitive function in Duchenne muscular dystrophy: Phenotype-genotype correlation. *The Journal of Pediatrics, 161*, 705-709. e701.

Parker, S. T. (2020). Estimating nonfatal gunshot injury locations with natural language processing and machine learning models. *JAMA Network Open, 3*, e2020664-e2020664.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825 - 2830.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*.

Perrin, A. (2001). The CodeRead System: Using natural language processing to automate coding of qualitative data. *Social Science Computer Review, 19*, 213-220.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*, 1-67.

Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health and Quality of Life Outcomes, 2*, 14.

Receptiviti. (2021). LIWC. *URL* https://www.receptiviti.com/liwc

Reyes, S. (2019). Multi-class text classification (TFIDF). *URL* https://www.kaggle.com/selener/multi-class-text-classification-tfidf

Schwartz, C. E., Biletch, E., Stuart, R. B. B., & Rapkin, B. D. Patient aspirations in the context of Duchenne Muscular Dystrophy: A mixed-methods case-control study. *Under review*.

Schwartz, C. E., Biletch, E., Stuart, R. B. B., & Rapkin, B. D. Sibling aspirations in the context of Duchenne Muscular Dystrophy: A mixed-methods case-control study. *Under review*.

Schwartz, C. E., & Revicki, D. A. (2012). Mixing methods and blending paradigms: Some considerations for future research. *Quality of Life Research, 21*, 375-376. doi:10.1007/s11136-012-0124-8

Scoring rule. (2021). In Wikipedia. https://en.wikipedia.org/wiki/Scoring_rule#Proper_scoring_rules

Skaljic, M., Patel, I. H., Pellegrini, A. M., Castro, V. M., Perlis, R. H., & Gordon, D. D. (2019). Prevalence of financial considerations documented in primary care encounters as identified by natural language processing methods. *JAMA Network Open, 2*, e1910399-e1910399.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. New York: Psychology Press, Taylor & Francis Group.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43-50.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Trotter II, R. T. (2012). Qualitative research sample design and sample size: Resolving and unresolved issues and inferential imperatives. *Preventive Medicine, 55*, 398-400.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008.