This is an Author Accepted Manuscript of an article in press at *Motivation and Emotion*, DOI: 10.1007/s11031-020-09832-8

Automated Coding of Implicit Motives: A Machine-Learning Approach

Joyce S. Pang* and Hiram Ring

Nanyang Technological University, Singapore and University of Zürich, Switzerland

Author note:

*Corresponding author: Joyce S. Pang, School of Social Sciences, Nanyang Technological University, Singapore, Singapore. Email: joycepang@ntu.edu.sg.

Hiram Ring, Department of Comparative Language Science, University of Zürich, Zürich, Switzerland. Email: hiram.ring@uzh.ch.

The two authors are equal contributors, and order of names is alphabetical. Both authors contributed equally to conception and design. Dataset compilation, preparation, and statistical analyses were performed by Joyce S. Pang. Dataset processing and machine learning scripts were written by Hiram Ring. The first draft of the manuscript was written by Joyce S. Pang and both authors commented on and revised subsequent versions of the manuscript. Both authors read and approved the final manuscript.

Conflict of Interest: The authors declare that they have no conflict of interest.

Acknowledgments: We would like to thank Oliver C. Schultheiss, Jonathan E. Ramsay, Thuy-Anh Ngo, Rena Ng, and Tingxuan Leng for sharing their data with us, and Veronika Brandstätter for reviewing a previous version of this manuscript.

Abstract

Implicit motives are key drivers of individual differences but are time-consuming to assess, requiring many hours of work by trained human coders. In this paper we report on the use of machine learning to automate the coding of implicit motives. We assess the performance of three neural network models on three unseen datasets in order to establish baselines for convergent, divergent, causal, and criterion validity. Results suggest that this is a promising direction to pursue in developing an automatic procedure for coding implicit motives.

Keywords: Implicit motive assessment, Picture Story Exercise, machine-learning, natural language processing, automated content coding

Automated Coding of Implicit Motives: A Machine-Learning Approach

Implicit motives are non-conscious, affectively-based dispositional preferences that direct and energize behavior and serve as key drivers of individual differences, guiding short-term as well as long-term preferences and behavioral and emotional outcomes. They have been related to differences between individuals in such diverse life domains as organizational behavior and worker management (e.g., Thielgen, Krumm, & Hertel, 2015), sports (Furley, Schweizer, & Wegner, 2019), inter- and intra-group dynamics (Stoeckart, Strick, Bijleveld, & Aarts, 2018), and intimate relationships (Hagemeyer, Neberich, Asendorpf, & Neyer, 2013).

Implicit motives have been contrasted with explicit motives that are traditionally assessed with declarative measures such as the Personality Research Form (Jackson, 1984) and the Unified Motive Scales (Schönbrodt & Gerstenberg, 2012). Implicit and explicit motives are statistically as well as conceptually distinct, such that scores on implicit motive measures correlate weakly or not at all with scores on explicit motive measures (Schultheiss, Yankova, Dirilikvo, & Schad, 2009), and have been argued to predict different classes of behavior (McClelland, Koestner, & Weinberger, 1989; Schultheiss & Brunstein, 2010).

Since they are non-conscious and cannot be assessed directly via self-report, implicit motives have traditionally been assessed using non-declarative measures that rely on the analysis of thematic content in freely generated texts, such as elicited by the Picture Story Exercise (PSE; Schultheiss & Pang, 2007), in which participants write imaginative stories in response to a battery of pictures that depict characters in everyday situations. Human coders then read these stories and make determinations regarding the presence of particular motive imagery in the stories, based on coding categories in established empirically-derived coding systems.

Each motive coding system contains a set of guidelines for coding motive related imagery. As an example, in Winter's (1994) coding system for scoring the implicit motives

for achievement (*n*Ach), affiliation (*n*Aff) and power (*n*Pow) in running text, a sentence is coded for *n*Ach if it indicates a concern for competition against a standard of excellence, it is coded for *n*Aff if it indicates a concern for the establishment or maintenance of friendly relations with others, and it is coded for *n*Pow if there is a concern for having or maintaining social influence. Furthermore, the Winter (1994) coding system contains five sub-categories for *n*Ach, six sub-categories for *n*Pow, and four sub-categories for *n*Aff.

Besides learning and internalizing these categories and sub-categories, human coders are required to achieve at least 85% agreement with training materials contained in standardized coding systems such as Winter's (1994). It takes at least 20 hours of coding practice to achieve this standard (Weinberger, Cotler, & Fishman, 2010), and even after the human coder has fulfilled the minimum 85% agreement criterion of correspondence on training materials, all stories must be coded by at least two independent coders who are required to establish inter-rater reliability (again, the 85% standard applies) on a subset of the stories (typically 10%) before they proceed to code remaining text. Finally, researchers also need to factor into their research process the time required for a human coder to read and evaluate each story for motive imagery. This could be anywhere from 3-10 minutes per picture stimulus, depending on the length and density of the text produced for each picture by a particular respondent, as well as the coder's experience and ability to quickly apply the coding rules. When evaluating 100 participants on 5 pictures, with an average of 5 minutes to code a single picture response, this easily reaches upwards of 41 hours of coding, per coder.

Consequently, the assessment of implicit motives is time-consuming and effortful, which can serve as a significant deterrent to researchers. Nonetheless, given the wide range of outcomes and research areas in which implicit motives have contributed theoretically-interesting findings—from parent-child relationships (Safyer, Volling, Schultheiss, & Tolman, 2019) and income growth trajectories (Apers, Lang, & Derous, 2019) to racial and

ethnic relations (Ditlmann, Purdie-Vaughns, Dovidio, & Naft, 2017) and users' profile content on online social networks (Dufner, Arslan, & Denissen, 2018)—it would be a significant boon to researchers in multiple fields if the time burden for implicit motive assessment could be reduced. To this end, our research aims to evaluate the feasibility and the validity of a machine-learning derived automated coding process for implicit human motives.[1]

Previous Automation Research: The Marker Word Approach

Previous published research on automating implicit motive coding includes Smith (1968), Pennebaker and King (1999), Hogenraad (2005), and Schultheiss (2013), among others.[2] These works are all based on the marker word hypothesis, which asserts that there are specific words related to a single motive that can be identified beforehand and counted in order to generate an approximation of the human-coded motive score.

Schultheiss (2013) undertook the most recent and systematic published endeavor based on the marker word approach and used the popular and readily available Linguistic Inquiry and Word Count (or LIWC) software and dictionary (Pennebaker, Francis, & Booth, 2001; Wolf et al., 2008 for the German version) for his analysis. He showed that, although certain LIWC categories were associated with the human-coded implicit motive scores, only 89 out of 684 of the computed zero-order correlations passed the .05 significance threshold, and the absolute values of these statistically significant zero-order correlations ranged between .19 and .40, with most *r*s hovering around .25. He further developed regression models, which accounted for between 35% (*n*Pow) to 54% (*n*Aff) of the variance of the hand coded motive scores, and when he computed sample-specific linear regression equations based on the aforementioned LIWC categories, variance accounted-for increased to as high as 63% (*n*Aff for the US sample).

Schultheiss (2013) thus demonstrated that computer estimations of implicit motive scores using a marker word approach can achieve modest to moderately-high correspondence with human coders (with a custom-built linear regression estimation approach), and that computer-based implicit motive score estimations possess some degree of causal validity (study 2). However, lexicon-based approaches suffer from the limitations of oversimplification, reduced sensitivity to more subtle or ambiguous instances of motive expression, and a lack of verisimilitude or generalizability. Specifically, with respect to oversimplification, as Schultheiss (2013) suggests, a human coder is attuned to the unique combination of words in a sentence, rather than just the individual words of the sentence. While individual words such as "reporter", "skeleton", "closet", and "embarrass" would not be coded for any motive, the combination of these words in a sentence such as "The reporter sought to find a skeleton in the officer's closet that would embarrass him" and within the context of a passage of text about retaliative one-upmanship, would cause the human coder to make the judgment that the sentence should be coded for *n*Pow (a similar sentence appears on page 19 of the training materials for Winter's 1994 manual). This information is lost in a marker word approach.

Marker words are also likely to only be associated with human-coded motive scores when there is strong or unambiguous motive intent expressed in a piece of text. Schultheiss (2013) notes that "accomplished" is a strong marker word for *n*Ach that would be scored under Winter's (1994) coding sub-category of "unique accomplishment," but it is unlikely that there would be many marker words that are just as unambiguous. This problem is compounded by the fact that many words possess multiple meanings. Such cases of lexical ambiguity are often used in humor, such as in an advertisement now banned by the Australian government, in which a picture-framing shop claims "We can shoot your wife and frame your mother-in-law...we can hang them too" (Awford, 2016). Human coders are usually able to

cope with cases of lexical ambiguity by referring to the context of the text in which a sentence occurs, but this is difficult for automation research that uses a straightforward lexicon-based approach.

Additionally, the inferences that a human coder makes are partially dependent on the body of experience that he or she has cultivated from coding and from discussing texts with other experienced coders. This reference to prior experience and extra-linguistic real-world knowledge is related to the phenomenon of 'semantic prosody', which refers to the conceptual co-occurence of words in natural language that imbues otherwise neutral words with attitudinal overtones (Cheng, 2013). Classic examples are *happen* and *cause,* which have been shown to be collocated with negative words, and *bring about,* which is associated with positive words (Louw, 1993). A human coder would be familiar with the typical contexts in which a word is used, as well as whether these contexts have positive or negative connotations, which would color the implicit evaluations that the human coder makes about any new sentences in which they encounter the word (Hauser & Schwarz, 2016). Marker word approaches thus lack verisimilitude because they miss out on such common concordances between words and phrases that human coders are able to detect because of their wealth of experience with combinations of words used in many different contexts.

Furthermore, the marker word approach lacks generalizability as it relies solely on the dictionary and the lexical equivalences built into the dictionary by the researchers in question. As such, the marker word approach is highly dependent on the theoretical biases of said researchers when generating motive score predictions, as well as the particular set of stimuli that were used to generate dictionaries (Smith, 1968). The marker word approach thus suffers from a lack of generalizability because it is impossible for a single researcher or research group to account for the multitude of word combinations that might be produced in various natural language contexts (e.g., across different PSE picture sets or in non-PSE textual data).

Machine Learning and Natural Language Processing

In order to avoid some of the limitations inherent in the marker word approach, we attempt to use machine learning to approximate the holistic judgment of human coders, using the sentence as the unit of measurement, whereas most previous research on automating motive scoring used the word as the unit of measurement (e.g., Pennebaker & King, 1999; Schultheiss, 2013). Moreover, since an aim of machine learning is to approximate the process of human-based classification of objects, one underlying objective of our research is to derive a process that would even account for cases in which a motive score is assigned where there are no obviously strong marker words associated with the motive. Thus we focus on building a neural network model from the perspective of natural language processing (NLP; Goldberg, 2015). We take this approach since we share an underlying assumption of marker-word approaches, namely that it is likely that the individual semantics of words and their combination in clauses (as opposed to extra-linguistic world knowledge of human coders, which cannot be easily accounted for in any computational approach) contribute to some of the nuances of meaning that are classified by human coders as motive imagery.

The field of NLP has moved from using dictionaries to more complex semantic representations (see Khurana, Koli, Khatter, & Singh, 2017 for a review), following insights from the subfield of Distributional Semantics, whose central idea is that a difference of meaning between words correlates with a difference of distribution in word occurrence in natural speech or text (Harris, 1954). This idea was popularized by Firth's (1957, pg. 11) statement "You shall know a word by the company it keeps," and is the foundation of many computational approaches to understanding natural language. For common NLP tasks such as machine translation and sentiment analysis, a normal workflow includes the use of such distributional models or 'word embeddings' (e.g., Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2017; Nickel & Kiela, 2017; Pennington, Socher, & Manning, 2014; Vulic & Mrksic,

2017; Yu, Wang, Lai, & Zhang, 2017) to transform the text into computational features. These models allow words in a corpus of natural(istic) texts to be represented in continuous vector space, so that words with more similar meanings/distributions are numerically 'closer together' than words with less similar meanings/distributions. A commonly cited example from Mikolov, Chen, Corrado, and Dean (2013, pg. 2) shows that analogous relations, i.e. "king is to queen as man is to woman" are encoded in word embedding matrices, such that the vectors for the respective words have the approximate relation of equations, i.e. *king - man = queen - woman*.

As input for a machine-learning model, word embeddings allow for a relatively nuanced semantic representation of a word (at least theoretically), based on the training data that underpins the word embeddings. That is, a word embedding matrix of 300 dimensions trained on a corpus of billions of words in naturally-occurring sentences will represent each individual word as 300 numbers (a vector) that mathematically encode its occurrence in a sentence and its relationship with other words (Allen & Hospedales, 2019; Goldberg, 2015). These vector representations are a better measure of meaning than a simple dictionary which does not clearly capture relations between words. Since they are based on observations of actual sentences, word embeddings also help to mitigate researcher bias and account for more of the diversity in language than in a dictionary approach. A large number of word embeddings have been made available by various research groups, the most well-known of which are Google's Word2Vec, Stanford's GLoVe, and Facebook's FastText (see references above regarding 'word embeddings'). The use of such pre-trained word embeddings allows NLP researchers to benefit from distributional semantic models without access to the computational resources or the large amounts of data required for training such models.

In our research, we use word embeddings (or 'word vectors') to encode sentences as input to several neural network models, with the goal of classifying unseen data according to

implicit motives. There are many different kinds of neural network models available, with differing performance on various natural language processing tasks. The most commonly used architectures for text classification (see Bai et al., 2018 for a review) are the Long Short-Term Memory (LSTM) network and its Bidirectional variant (Bi-LSTM), both types of Recurrent Neural Network (RNN), as well as the Convolutional Neural Network (CNN) and the more recent Temporal Convolutional Network (TCN). LSTMs are very good at capturing dependencies such as might be found between words in a sentence but can be extremely time-consuming to train. CNNs and TCNs, with some modification, can also capture such dependencies and have shown excellent results on text classification tasks (see Kim, 2014, and Zhang & Wallace, 2016 on CNNs, and Kalchbrenner et al., 2016, Lea, Vidal, Reiter, & Hager, 2018, and Bai, Kolter, & Koltun, 2018 on TCNs). These neural network architectures have the added advantage of being much faster to train than RNNs. Newer architectures include Attention networks and the Transformer (Vaswani et al., 2017), but these have a higher bar for implementation in terms of computational resources, and our initial experimentation with these architectures yielded no significant benefits. Thus, in this paper we only report results from a one-dimensional CNN (which captures local dependencies between words in sentences), a two-dimensional CNN (which can capture non-local dependencies in word vectors), and a TCN model (which can capture non-local and longer-range historical dependencies than CNNs and LSTMs).

The Current Research

The aim of the current research program is to ascertain whether it would be possible to move the field of implicit motive research forward by automating the coding of implicit motive imagery in running text. Specifically, we evaluate whether a machine-learning approach aided by natural language processing, using three different neural network

architectures, could generate motive score predictions that correspond to those generated by human coders, or at least show higher correspondence than previous approaches.

The development of an automated motive coding process would allow researchers who are interested in implicit motives to assess them without the significant costs to their research timeline that this would normally entail. It would also help researchers in cases where it is physically impossible to administer the PSE, or other implicit motive measures that require the generation of imaginative material (e.g., the Operant Motive Test; Kuhl & Scheffer, 1999). Specifically, with an automated process, archived materials could be 'fed' into the machine, thus opening up an avenue for the assessment of implicit motives of individuals who are deceased, incarcerated, or otherwise physically unavailable for assessment.[3] The availability of an automated coding process could also allow for the processing of larger amounts of textual data than otherwise possible with human coding. This would open up the possibility of assessing motives in bigger and more diverse datasets. Finally, an automated coding process would hypothetically enable the 'real-time' assessment of implicit motives, which would make the research process more dynamic.

We evaluate for the first time whether motive score predictions from a machine-learning derived automated coding process using neural network model architectures and aided by natural language processing achieves correspondence with human-coded scores from several 'unseen' datasets that were not used to train the machine-learning models. We also examine whether automated motive score predictions possess convergent validity, in terms of showing moderately high and significant correlations with human-coded scores for conceptually similar implicit motives, as well as divergent validity in terms of showing low or no correlations with human-coded scores for conceptually dissimilar implicit motives. Finally, in order to examine whether the motive score predictions from the automated coding process possesses criterion and causal validity, we use data from two of the unseen datasets to

ask whether theoretically-consistent group differences observed with human-coded motive scores can also be observed with the automated motive score predictions.

While previous work on automating implicit motive scoring described some limited success, these researchers relied on the marker word approach, which, as argued, has some significant shortcomings. Moreover, they did not consistently demonstrate that their methods achieved the various forms of validity that Schultheiss (2013) gave as necessary for an automated coding process to be adopted by researchers. Thus, our main goal is to develop a process for automating motive scoring for all three motives of achievement, affiliation, and power, while a secondary goal is to apply our trained neural network models to unseen datasets in order to identify whether such models can achieve the level of validity required for adoption by implicit motive researchers. To be clear, in this paper, we are not proposing a tool that can replace human coders for implicit motive research. Instead, we wish to provide a benchmark for training and validating machine-learning models in the years-long task of automating implicit motive coding, with the ultimate goal of developing such a tool.

We use a "top down" approach (Mendez, Hinrichs, & Nacenta, 2017) in training machine-learning models, relying on human-coded scored sentences as input. We gathered English-language PSE datasets from various sources that had previously been coded by trained human coders for *n*Ach, *n*Aff, and *n*Pow using the Winter (1994) scoring system. We re-parsed all datasets and had independent coders (who were blind to study hypotheses) re-score the datasets for motives at a more granular, sentence level. We then used these human-coded data for training and validation of several machine-learning models.

Assessing Validity of the Neural Network Models

In order to assess the validity of this approach for use by motive researchers, we examined the correspondence between the human-coded motive scores and the machine predictions of the motive scores for three different unseen datasets (i.e. those that the

machine-learning model had not been trained on). The first unseen dataset (UD1) contains 1,357 sentences from practice sets in Winter's (1994) coding manual. The second unseen dataset (UD2) contains 28,686 sentences from the Enron Email Corpus (Klimt & Yang, 2004) from 18 individuals in upper management. The third unseen dataset (UD3) contains 475 sentences from an unpublished study on goal visualization. If the correspondence between the human-coded and the machine-predicted motive scores is high for all three unseen datasets, this would indicate that the machine-generated algorithm built from the training dataset is generalizable to a wide range of implicit motive data. The use of these datasets allowed us to conduct assessments of convergent, divergent, causal, and criterion validity for our machine-predicted motive scores.

Convergent and divergent validity. For convergent validity, we used human-coded scores for *n*Ach, *n*Aff, and *n*Pow in UD1 and UD2. We expected machine predictions of these motive scores to correlate positively with their corresponding human-coded scores. We used human-coded scores for implicit hope of success and fear of failure motivation in UD3 to examine convergent as well as divergent validity. Specifically, we computed correlations of these human-coded scores with the machine predictions for *n*Ach, *n*Aff, and *n*Pow. Insofar as the motives for hope of success and fear of failure are approach and avoidant orientations of the generalized implicit achievement motive (Pang, 2010), machine predictions for *n*Ach should correlate positively with human-coded scores for hope of success and negatively with human-coded scores for fear of failure, thus demonstrating convergent validity. Conversely, there should be minimal correlations for hope of success and fear of failure scores with the machine predictions for *n*Pow and *n*Aff, thus demonstrating divergent validity.

Causal validity. In addition, the participants in UD3 had been randomly assigned to visualize an achievement goal that was either positively framed (approach goal) or negatively framed (avoidant goal). We analyzed differences in the mean machine-coded achievement

motive scores for each experimental group. Insofar as scores on a measure of an attribute should show mean differences in scores for groups which are known to differ in that attribute or which have been experimentally induced to be at different levels of that attribute (e.g., McClelland, Atkinson, Clark, & Lowell, 1953), we expected that the machine-predicted *n*Ach would be higher for participants in the approach goal visualization condition than for those in the avoidant goal visualization condition, thus demonstrating causal validity (c.f. Borsboom, Mellenberg, & Heerden, 2004).

Criterion validity. Finally, we used UD2 to demonstrate criterion validity. Insofar as the machine-coded scores are a valid measure of implicit motives, they should predict outcome variables that have theoretically been associated with these motives. McClelland (1975) has theorized that achievement motivated individuals are spurred by the need to succeed so that they could be driven to obtain this success by any means possible—including unethical ones. While previous research has not investigated the direct relationship between implicit motives and unethical workplace behaviors, some have looked at the effects of the achievement motive on unethical behaviors in other settings. Johnson (1981) showed that individuals with a high achievement motive were more likely to cheat on college examinations than those with low achievement motive. Whitley's (1998) meta-analysis found that achievement motivation had a small positive effect (d=. 250) on cheating in examinations. However, Johnson's (1981) and Whitley's (1998) research relied on explicit measures of the achievement motive. Smith, Ryan, and Diggins (1972) showed that implicit *n*Ach is related to self-reported dishonesty, however one could argue that a self-reported measure of dishonesty is only marginally related to unethical behavior.

Winter (2010) also showed in several case studies of US presidents, that being highly achievement motivated could lead a president to take shortcuts within the political process—sometimes illegally (e.g., Nixon in the case of Watergate). As Winter (2010) has argued, the

reason that highly achievement motivated individuals tend to succeed in business but not in politics is because they are drawn to situations in which they are able to execute personal control and they become frustrated by the lack of control inherent in politics. Thus, the question of whether highly achievement-motivated individuals are more likely to engage in unethical behavior in the workplace—even though they are more likely to be able to maintain personal control of their actions—is an empirical one.

Although prior evidence for McClelland's (1975) theory is scant, we expected that machine predictions of *n*Ach should be related to unethical behavior in the workplace. Specifically, UD2—from the Enron Email Corpus (Klimt & Yang, 2004)—contains a subset of emails from 18 of roughly 150 employees of the Enron Corporation, some of whom were investigated for fraud resulting from insider trading (Leber, 2013). We examined whether machine-coded *n*Ach and/or *n*Pow scores differentiated between individuals in upper management who had been clearly implicated in the scandal (henceforth denoted as "misbehavers", $n = 9$) versus those who had not ("non-misbehavers", $n = 9$). Following McClelland (1975), we expected that machine-coded *n*Ach would differentiate between misbehaver and non-misbehaver status. However, given that the members of the misbehaver group were extremely influential employees in Enron, another possible hypothesis is that the misbehavers were driven by *n*Pow to conduct unethical behaviors in order to maintain their positions of influence. If this were the case, we might expect to see that machine-coded *n*Pow differentiates between misbehaver and non-misbehaver status.

Method

Method for Building the Machine-learning Model

Training datasets and data processing. To develop a set of data for training the neural networks we contacted several implicit motive researchers to request use of their anonymous, scored PSE data from studies conducted in English. A total of 11 datasets were

collected from six researchers in the USA and Singapore. All datasets were checked for grammatical, spelling and punctuation errors, and acronyms (e.g., "POW" = Prisoner of War) and abbreviations (e.g., "Prof." = "Professor") were expanded. Some contractions (e.g., "gimme" = "give me") that were judged to be colloquialisms were expanded, but proper contractions (e.g., "don't) went uncorrected. All data were split into single sentences which were re-scored by six different independent coders who were blind to study hypotheses, using Winter's (1994) scoring system. Additionally, the Winter (1994) manual recommends applying a "second-sentence-rule" for coding, such that, when the same motive imagery appears in two consecutive sentences, only one of the sentences is counted towards the raw motive score. However, our independent coders did not apply this second-sentence-rule as it would likely falsely increase the frequency of the null category and hence distort the diagnostic fidelity of the training dataset. In other words, each sentence was evaluated for motive imagery independently from its preceding sentence.

Each coder was responsible for coding a subset of the total PSE dataset. All six coders obtained at least 85% reliability with the expert coder in the training materials provided by Winter (1994). Since 85.4% of the PSE data were coded by three coders, these three coders coded a further 30 stories so that we were able to calculate the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) using a two-way, mixed-effects model (average rating, consistency measure). The ICC for the *n*Pow scores was .83, which could indicate good reliability (Koo & Li, 2016). However, the 95% CI had a large range from .69 to .91, indicating that the level of inter-rater reliability for *n*Pow scores range from moderate to excellent. The ICC for the *n*Ach scores was .90, and the 95% CI ranged from .82 to .95, indicating that the reliability for *n*Ach scores should be regarded as ranging from good to excellent. Finally, the ICC for the *n*Aff scores was .96, and the 95% CI ranged from .93 to .98, indicating that the reliability for *n*Aff scores should be regarded as excellent. The

average ICC for all three motives was .87, which is indicative of good reliability (Koo & Li, 2016).[4] There were altogether 65,681 sentences across 11 datasets containing PSE data—key characteristics of each of the datasets are given in Table 1. Additionally, since the linguistic style and content of emails are likely to differ substantially from PSE data, a portion of sentences from the Enron emails not included in UD2 were also scored and included in the training dataset, bringing the total number of sentences to 73,907. From this set we removed single-word sentences, giving a final training dataset of 73,003 sentences.

Building, training, and validation of the models. The final set of 73,003 sentences was used as the input for training using 5-fold cross-validation, meaning that in a given training session 80% of the sentences were used for training, while the remaining 20% were used to evaluate the model. We used a balanced split to ensure that the absence/presence of motive codes were proportional across the training and validation datasets. Additionally, given the imbalance of codes (positively-coded sentences were far outweighed by sentences without any code), before training we computed weight values for each motive based on the proportion of sentences in the dataset that were positively-coded for that motive. These values were used during training to penalize/reward the models for incorrect/correct predictions on the training data.

In pilot experiments we evaluated a variety of pre-trained word embeddings with a baseline CNN (modeled on Kim, 2014), finding that 300-dimensional embeddings performed better than lower dimensional embeddings, as did word embeddings trained on larger datasets. Based on these experiments we decided to use Facebook's FastText subword embeddings of 300 dimensions trained on Common Crawl (600 billion tokens).[5] This is the set of pre-trained vectors that we used to derive word features from sentences for all the experiments that we report below.

We trained three neural networks with different architectures: A one-dimensional convolutional network (CNN), a two-dimensional convolutional network (CNN2D), and a temporal convolutional neural network (TCN). Input to each of the neural networks was vectorized via the pre-trained FastText word embeddings, and then fed into the convolutional layers. We used grid search with 10-fold cross-validation (CV) to optimize the number of neurons and kernels for each architecture. The hyperparameters that gave the best average performance (high accuracy, low loss) during model-internal validation (on the held-out *k*th fold) were used to train each of the final models. Final CV accuracy/loss scores for the hyperparameter selection phase, by model, were 0.929/0.206 (CNN); 0.928/0.202 (CNN2D); 0.920/0.215 (TCN).

The final CNN and CNN2D models had four convolutional layers with kernel widths of 1, 2, 3, and 5 instantiated with 64 neurons. These layers were concatenated and passed through a max-pooling layer before being fed into the final dense layer. The TCN was implemented with 32 neurons, a kernel size of 4, 5 stacks, and dilations of [1, 2, 3, 5] and a final dense layer. For all three networks, the final layer consisted of 3 neurons (one for each motive) which each returned a probability between 0 and 1 (rounded using a threshold of 0.5) to provide the machine predictions regarding the presence of motive imagery in a given sentence. For each network, batch normalization was used after each convolutional layer to prevent overfitting (see Li, Chen, Hu, & Yang, 2018), and kernels were initialized using He normal weights (see He, Zhang, Ren, & Sun , 2015). All final training sessions were run in Python / Keras with Tensorflow 1.15 on a workstation equipped with an AMD 1700x CPU and NVIDIA RTX 2060 GPU running Ubuntu 18.04 LTS. Additionally we used the TensorFlow Determinism library (Riach, 2019) to ensure reproducible results.

The CNN trained for 100 epochs with a mini-batch size of 64, the CNN2D trained for 40 epochs with a mini-batch size of 32, and the TCN trained for 20 epochs with a mini-batch

size of 64. Each neural network model achieved the accuracy and loss figures as given in Table 2. It is important to keep in mind here that, in machine learning terms, the high accuracy and low loss values for the validation set in Table 2 only describe how well the neural network has learned from the dataset it is trained on, which in theory should also reflect how well it generalizes to unseen data. However, there is always the danger of overfitting, i.e. that the neural network will essentially memorize the data it is trained on, meaning that it will not be robust enough to generalize and make valid predictions on data that it did not 'see' in training.

In much machine learning research there is a focus on developing models that achieve the best accuracy and loss scores on a given dataset, to such an extent that reporting a 0.0001 increase in validation accuracy on a given dataset can be seen as justifying a particular model architecture. While it is useful to understand how model architecture affects learning performance, our focus here is to assess more generally whether the machine learning process might allow for automation of motive coding. Since models that achieve high accuracy and low loss do not necessarily generalize well to unseen data, we focus less on these measures and more on the assessment and evaluation of predictions generated by the trained models on the unseen datasets (UD1, UD2, UD3).

Preparation and pre-processing of unseen datasets 1, 2, 3

Each of the unseen datasets were processed in the same manner as the training dataset (e.g., split into sentences, spellchecked). Additionally, UD1 and UD2 were also coded by human coders using the Winter (1994) system in order to check the correlations between the human-coded and machine-coded scores. Any disagreements between human coders were resolved by discussion.

In the case of UD1, the 1,357 sentences were transcribed from practice sets A to F and the calibration set A of the Winter (1994) manual's training material. Next, UD1 was "coded"

by a different set of coders than the datasets from which the model was built. Since the human coders' task was simply to transcribe and check the expert coding from the Winter training materials for absence or presence of motive coding on a sentence-level basis, the two coders reached 100% inter-rater agreement in this task.

For UD2, we first identified individuals in upper management who were visible in the media during the Enron scandal and subsequent trials (see Appendix for details on the sources we consulted). This group of 18 individuals were then divided into "misbehavers" (convicted or involved in a plea bargain) and "non-misbehavers" (found not to be at fault). All the emails were manually coded according to Winter (1994) motives by two coders, each coder being responsible for approximately 50% of the emails. To establish inter-rater reliability for motive coding, the coders independently scored a common set of 393 PSE stories written by 65 research participants. The two-way mixed ICC (average rating, consistency measure) between the two coders was .84 for the power motive (.74 <95% CI < .90), .85 for the achievement motive (.76 <95% CI < .91), and .78 for the affiliation motive (.64 <95% CI < .87), so that on average, the ICC was .82 for all three motives, indicating that the reliability is good. The personally written (sent) email content for each of the 18 Enron employees for the period of fraud investigation (2000-2002) was extracted from the email corpus, resulting in 28,686 sentences for coding.

For UD3, 475 sentences were extracted from 38 participants (25 women, 12 men, mean age = 22 years, *SD* = 1.2) who took part in an unpublished study on achievement goal visualization. Participants in this study were asked to spend two minutes visualizing a previously attempted achievement goal and then to spend four or five minutes typing out the scenario and background during which the goal took place, as well as their thoughts and feelings that they encountered during the goal progress. Participants were randomly assigned to one of two conditions: In one condition, participants were asked to recount a previously

attempted approach achievement goal (i.e., one which was framed in a positively worded direction and focused on obtaining a desired outcome), while in the other condition participants were asked to recount a previously attempted avoidant achievement goal (i.e., one which was framed in a negatively worded direction and focused on preventing an undesired outcome). Since UD3 was intended to be used to assess convergent and divergent validity with respect to correlations of machine scores with human-coded scores of hope of success and fear of failure motives (Heckhausen, 1963, as translated by Schultheiss, 2001), these sentences were not re-coded for the motive scores in Winter's (1994) system.

Results

All subsequent analyses reported here with UD1, UD2, and UD3 were conducted with wordcount corrected scores for both human-coded as well as machine-predicted motive scores, using the procedure recommended by Winter (1994; i.e., [(motive score/wordcount) *1000 words]). For ease of readability, we have grouped our findings according to the criterion of validity that is being addressed (i.e., convergent, divergent, etc.). Finally, we also applied a Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) using a false discovery rate of .05 in order to minimize the risk of false positive findings.

Convergent and divergent validity. In order to examine convergent validity of our machine-predicted motive scores, we calculated bi-variate correlations between human-coded and machine-predicted motive scores for *n*Pow, *n*Aff, and *n*Ach for UD1 and UD2, and between machine-predicted *n*Ach scores and human-coded hope of success and fear of failure motive scores for UD3. In order to examine divergent validity, we calculated bi-variate correlations between machine-predicted *n*Pow and *n*Aff scores and human-coded hope of success and fear of failure motive scores for UD3.

For UD1, each neural network model predicted a raw *n*Pow, *n*Ach, and *n*Aff score for each of the 1,357 sentences. Table 3 presents the individual *r*s, *p*s, and corresponding 95% confidence intervals between human-coded and computer predicted scores for each motive

and each model (all correlations were significant, all *ps* < 0.001). The total averaged correlations between human-coded motive scores and their corresponding machine-predicted motive scores for the CNN, CNN2D and TCN models were 0.423, 0.457, and 0.370 respectively, indicating that on average, the computer predictions of motive scores were significantly and moderately positively correlated with the human-coded motive scores. The correlations in table 3 that originally achieved the $p < .05$ threshold remained significant even after the Benjamini-Hochberg correction.

For UD2, each neural network model also predicted a raw motive score for each sentence. Motive codes were then further aggregated by email sender. Table 4 presents the individual *r*s, *p*s, and corresponding 95% confidence intervals between human-coded and computer predicted scores for each motive and each model. The total averaged correlations between human-coded motive scores and their corresponding machine-predicted motive scores for the CNN, CNN2D and TCN models were .694, .683, and .540 respectively. The correlations in table 4 that originally achieved the $p < .05$ threshold remained significant even after the Benjamini-Hochberg correction.

For UD3 we again predicted per-sentence scores with each neural network model and aggregated the scores by participant. Table 5 depicts the correlations between computer-predicted scores for *n*Ach, *n*Aff, and *n*Pow, and the human-coded scores for HS and FF motives. As expected, the computer-predicted scores for *n*Ach from the CNN and the TCN models (but not for CNN2D) were moderately positively and statistically significantly correlated with the human-coded score for HS, even after we applied the Benjamini-Hochberg correction. Moreover, the correlations between the computer predictions for scores of *n*Pow or *n*Aff and the human-coded scores for HS and FF were low to modest (*r*s ranged from -.06 to .25), with absolute values of *r* across all model predictions with the human-coded motive scores averaging .15 and .18 for *n*Aff and *n*Pow respectively. The relative size

of these correlations compared to those for machine-predicted *n*Ach are suggestive that the human-coded HS and FF motive scores are more closely associated with the machine-coded motive score for *n*Ach than with the machine-coded motive scores for either *n*Pow or *n*Aff.

Overall, the pattern of findings across UD1, UD2, and UD3 suggests that the computer predictions from the machine-learning algorithm demonstrate some degree of convergent and divergent validity.

Causal validity. In order to examine causal validity of the machine-predicted motive scores, we conducted Mann-Whitney U tests to compare the mean achievement motive scores by experimental condition in UD3. As shown in figure 1a, participants who were asked to visualize and to recall an episode where they pursued an approach achievement goal demonstrated higher scores for human-coded hope of success (HS) motivation (M = 19.75, SD = 16.41) compared to participants who were asked to visualize and to recall an episode where they pursued an avoidant achievement goal (M = 2.26, SD = 5.12; Mann Whitney U = 29.50, p = .001, η^2 = .56). This result remained significant even after we applied the Benjamini-Hochberg correction. Conversely, participants who were asked to visualize and to recall an episode where they pursued an avoidant achievement goal demonstrated higher scores (M = 34.43, SD = 19.65) for human-coded fear of failure (FF) motivation compared to participants who were asked to visualize and to recall an episode where they pursued an approach achievement goal (M = 5.08, SD = 7.89; Mann-Whitney U = 16.00, p = .001, η^2 = .64). This result remained significant even after we applied the Benjamini-Hochberg correction. This finding is consistent with prevailing theory about the distinction between HS- and FF-motivated individuals, in terms of their differential sensitivity toward positive versus negative achievement goals (e.g., Pang, Villacorta, Chin, & Morrison, 2009).

Additionally, as shown in figures 1b-1d, machine predictions of the *n*Ach score from the CNN, CNN2D, and TCN models were also higher for the participants in the approach

achievement goal visualization condition (M_{CNN} = 6.27, SD_{CNN} = 8.48; M_{CNN2D} = 5.42, SD_{CNN2D} = 7.68; M_{TCN} = 7.18, SD_{TCN} = 9.67) than for the avoidant achievement goal visualization condition (M_{CNN} = 1.67, SD_{CNN} = 4.13; M_{CNN2D} = 2.55, SD_{CNN2D} = 4.64; M_{TCN} = 1.80, SD_{TCN} = 3.91), mirroring the results for human-coded HS scores. Moreover, the difference in mean computer-predicted *n*Ach scores between conditions was statistically significant for the CNN model (Mann-Whitney U = 122.50, p = .040, η^2 = .11) and the TCN model (Mann-Whitney U = 122.00, p = .043, η^2 = .11). These results remained significant even after we applied the Benjamini-Hochberg correction. These findings are consistent with the prevailing theory about the measures for generalized achievement motivation being more aligned with approach motivation rather than avoidance motivation (e.g., Kuhl, 1978). Furthermore, if one examines the sub-categories in Winter's (1994) coding manual for *n*Ach, it would become apparent that these focus on positively framed achievement goals and their accompanying goal-directed behaviors and affect. Taken together, the findings from UD3 provide partial support of the causal validity of the machine-predicted *n*Ach scores.

Criterion validity. In order to examine the criterion validity of the machine-predicted motive scores, we conducted Mann-Whitney U tests to compare the mean *n*Ach and *n*Pow scores between misbehaver and non-misbehaver groups in UD2. There were no statistically significant differences between misbehavers and non-misbehavers in terms of human-coded or machine-predicted scores for *n*Ach or for *n*Pow.

Discussion

The aim of this research was to examine the possibility of automating implicit motive coding using a machine-learning approach. Across three unseen datasets containing diverse textual content (PSE-type stories in UD1, corporate emails in UD2, and goal visualizations in UD3), machine predictions of motive scores for *n*Ach, *n*Aff, and *n*Pow demonstrated moderate to moderately high correspondence with their related human-coded motive counterparts (*r*s ranged from .30 for *n*Pow to .86 for *n*Ach). Convergent and

divergent validity was partially demonstrated in UD3 when machine predictions of *n*Ach for two out of three models were significantly and moderately correlated with human-coded motive scores for hope of success motivation, while machine predictions of *n*Aff and *n*Pow were weakly related to the human-coded hope of success scores.

Also in UD3, causal validity was demonstrated when, as expected, the mean scores of machine-predicted *n*Ach were higher for the experimental group who was asked to visualize an approach achievement goal, when compared to the mean scores for the group who was asked to visualize an avoidant achievement goal. These mean differences reached levels of statistical significance in two out of three models. Moreover, as was previously mentioned, machine predictions of *n*Ach were positively correlated with human-coded scores for hope of success motivation, the mean scores of which were also significantly higher in the approach visualization condition compared to the avoidant visualization condition. Taken together, the findings from UD1, UD2 and UD3 indicate that a machine-learning approach to automating implicit motive coding is able to produce motive score predictions that are not only moderately correlated with their related human-coded motives, but also with motive-relevant outcomes in theoretically consistent ways.

Moreover, the computer predictions seem to demonstrate convergent, divergent, and causal validity across a range of different types of textual data that are notably different from the primary training data that were used to construct the machine-learning models. Specifically, UD2 and UD3 contained emails from a multinational corporation and recall data from goal visualizations of Singaporean undergraduates respectively, while the UD1 stories were from Winter (1994) training materials which were collected from Thematic Apperception Tests (TATs; Morgan & Murray, 1935) that were administered to undergraduate students from elite universities in the USA. Thus, UD1 contained written text generated in response to different picture stimuli, administration conditions, time period, and

cultural contexts than the PSE stories that contributed to our training and validation dataset. This divergence of textual data from the unseen datasets versus the training datasets suggests that a machine-learning approach could generate motive score predictions that are externally valid and would generalize to data gathered from other diverse research and cultural contexts. Our findings bode well for future work in automating motive coding and could eventually lead to a decrease in the resource burden that implicit motive researchers currently face during the measurement process.

Our research also improves on previous automation attempts as it moves beyond a straightforward marker word approach, which as explained, suffers from the shortcomings of over-simplification, lack of generalizability and verisimilitude, as well as a tendency to overlook more subtle or ambiguous motive imagery expressions in text. To be clear, however, our claim is not that neural network models can actually capture lexical ambiguities of the sort we discuss (yet), but that the use of word embeddings has more promise toward this goal than marker word approaches. Given sufficiently large amounts of training data from diverse sources, deep learning NLP approaches can better account for the substantial complexity involved in natural language expression, and may lead to predictive models that are stable, reliable, and generalizable to unseen data. These advantages of machine learning seem to bear fruit, as correspondences between our machine-predicted motive scores and human-coded scores are markedly improved from those in Schultheiss (2013), where correlations between LIWC word categories and motive scores ranged mostly between 0.2 and 0.3. Thus, while previous attempts to automate PSE coding using marker-words (e.g., LIWC) typically reinforce the continued need to rely on human coders (e.g., Schultheiss, 2013), our findings suggest that machine-learning techniques have the potential to streamline the measurement process. Although Schultheiss (2013) did not use a machine learning approach to automation, we make explicit comparison of the performance from our approach

with the performance of Schultheiss (2013) because his was the most recent automation endeavor to date that also includes data on various forms of validity (convergent, causal, criterion).

There are some interesting theoretical implications which we would like to note. First, machine predictions of *n*Pow consistently demonstrated lower correspondence with human-coded scores than the machine predictions of the other two motives. This could be related to the fact that the majority of the datasets we used for training were intended to study *n*Ach, but another possibility has to do with the nature of power motive expression being more diverse than the other two motives in its representation, acceptance, and legitimation in societies, which in turn leads to a greater diversity of its expression in language. Specifically, power can be pursued either in socialized or in personalized ways, and societies differ in their provision of structures and cultural norms for socially acceptable ways to exert influence on others (Busch, 2018). In situations where a direct expression of dominance is either physically uneasy or socially unacceptable, individuals who have strong personalized power motivation would likely experience power stress (Hofer & Busch, 2019; Raihala & Hansen, 2019), or would need to internalize their need for power, resulting in physiological agitation and hostility on one hand (e.g., Fodor & Riordan, 1995) or repression, anxiety, and emotional displacement on the other (e.g., Fodor, Wick, & Conroy, 2012).

Due to the fact that direct expressions of power are generally incompatible with notions of social agreeableness, we believe that power imagery is likely to elicit more figurative and metaphorical language than for the other two motives. NLP techniques are easily misled by irony and sarcasm, whereby the words used in an utterance might convey a literal meaning that is in fact the opposite of what the speaker intended. Additionally, the pre-trained word embeddings we used assume that the meaning of individual words is relatively stable across sentences, such that the polysemy inherent in word forms is not taken into

account. Advances are being made in the field of computational linguistics to deal with classification problems in texts that use figurative language and problems of polysemy (see Weitzel, Prati, & Aguiar, 2016, for a review), and these developments should improve future attempts at automating implicit motive coding.

Second, contrary to predictions, we found that neither machine-predicted *n*Ach nor *n*Pow scores were significantly related to misbehaver status in UD2. Since the human-coded motives were also not found to be related to misbehaver status, we are at present unable to determine whether our machine-predicted motive scores for *n*Ach and *n*Pow possess criterion validity. The null findings could indicate a lack of support for the criterion validity of machine-predicted motive scores, or they could be a reflection of the true result of a lack of relationship between the implicit motives and unethical corporate behavior. Classical theory about *n*Pow states that it should be associated with occupation in positions of power and leadership (c.f., Winter, 1973). It is thus possible that *n*Pow might not have been discriminative between the two groups because all 18 individuals belonged to Enron's upper management and thus probably had high *n*Pow to begin with. As for *n*Ach, a possible reason that *n*Ach scores in the Enron emails were unrelated to misbehaver status could be due to the particular unethical behavior which these employees were accused of. Specifically, they were accused of insider trading, which is the use of confidential and privileged information to one's financial advantage. Perhaps the financial incentive was far too attractive and thus the major impetus for the unethical behavior, overriding the influence of any personality factor. Nonetheless, to our knowledge, our study is the first empirical test of McClelland's (1975) theory regarding achievement motivation and cheating behavior in a corporate setting. Considering this and given the idiosyncratic and selective nature of our sample, more research is required before ruling out McClelland's theory.

There are some limitations of our research that need to be addressed. A possible shortcoming of our approach is that we sacrifice some degree of theoretical grounding. Machine learning is essentially a data-driven approach, whereby a statistical model is given a large number of examples containing data correctly sorted into different groups/classes/labels and tasked to derive its own set of rules for classifying the examples. This process of model building is in contrast to a theoretically-driven approach such as was developed by Stone, Dunphy, Smith, and Ogilvie (1966) using the General Inquirer system, which is an algorithm that included marker words in a series of rules for each motive which were theoretically consistent with the motivational behavioral sequence for that motive. For instance, according to Smith (1968), in the General Inquirer, if a sentence contained specific types of words in a particular sequence such as *Authority role* + *Subservient role* + *Influence verb*, then the sentence would be scored for power motive imagery. Thus, sentences were coded for motive imagery only when a certain combination of features occurred together. Schultheiss (2013) followed a similar procedure by combining specific marker words in a regression approach, e.g., groups of words tagged as *positive feelings*, *friends*, and *sexuality* belonged to the regression equation that he computed for *n*Aff for his USA sample. Much has been written about the "black box" metaphor (e.g., Krause, Perer, & Ng, 2016) of machine-learning approaches whereby the algorithm has achieved high accuracy in predictions, but the individual components that the machine has used to build its model are not apparent to the user. This makes the task of explaining why the algorithm works a difficult one, and may even introduce other kinds of unpredictable bias or ethical concerns (Swinger, De-Arteaga, Heffernan, Leiserson, & Kalai, 2018; Yapo & Weiss, 2018).

Another limitation is the handling of parts of speech and contextual information in the language data that was used to build the machine-learning models. In principle, the pre-trained word embeddings that we used should take part-of-speech information as well as the

multiple contexts in which words derive their meanings (the problem of polysemy) into account. NLP deep learning experiments with word embeddings show that they do capture some syntactic information similar to parts of speech (Shen, Satta, & Joshi, 2007). However, as we previously noted, the current state of affairs is such that polysemy still presents a significant problem for most publicly available pre-trained word embeddings. While human coders are naturally able to process extra-linguistic information and rely on their wealth of experience to deal with polysemy, much research is needed before NLP machine-learning approaches can approximate the human in this respect.[6] Nonetheless, given the intensity of research interest on word embeddings in deep learning, it is likely that new techniques will soon become available to better account for polysemy and word order (e.g., Mrksic et al., 2017; Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018).

An additional limitation of our research stems from a challenge inherent in building representative training datasets for data such as ours—PSE stories—which notoriously have an imbalance in positively coded sentences relative to sentences where motive code is absent (see online supplementary material which presents a breakdown of motive codes by sentence in our training data)[7]. Since the vast majority of sentences in our datasets had no motive imagery present, we controlled for this imbalance by ensuring that training and validation sets had roughly the same number of motive codes for each motive, but this meant that we could not control for persons by ensuring that they occur completely in the training or completely in the validation set. It is thus possible that some of what the machine learns from the training data is the linguistic style of persons whose sentences occur in both the training and test datasets. This potential confound should be mitigated in future research with the use of larger training datasets that would allow blocking of persons to occur so that their data appear entirely either in the training or in the validation set.

Finally, the fact that the three models we used showed such great variation in learning mappings between words within sentences and appropriate motive codes highlights the difficulty of such research. For instance, in UD2, while correlations between CNN and CNN2D model predictions for *n*Ach and the human-coded scores were high (*r*s above .80) and statistically-significant, the correlation for the TCN model prediction was substantially lower at $r = .42$, and non-significantly so. This suggests that the TCN model was unable to represent *n*Ach scoring very well, compared to the other two models. As illustrated by the 'black box' metaphor, since we don't have a clear idea of what kind of mappings the computer is learning in order to predict unseen data, it is difficult to identify exactly what to change or address in the underlying dataset in order to improve how and what the computer learns. This is one of the biggest issues in machine-learning research and is inherent in using stochastic gradient descent algorithms, which include some degree of random activation. We have tried to partially mitigate this by using machine-learning libraries that allow for deterministic output, such that training a model with exactly the same data and settings on the same hardware leads to the same results (see Riach, 2019). This may allow us to 'reverse engineer' the neural mappings in the future, but as a small change in one of the parameters can easily change how and what the model learns, and given limited resources, it is as much an art as a science to make educated guesses that improve the results. This is particularly the case given the size of our training dataset (less than 100,000 training samples), which in machine learning terms is tiny.

Another possible reason for the uneven performance of our models on different motives is the imbalance in source data for the PSEs in the training dataset: There were more contributing studies that focused on *n*Ach as the motive of interest, compared to the other two motives. We requested for PSE training data from close collaborators in our immediate social network, and since one of the authors of the current research focuses on achievement

motivation, this may have resulted in more contributions from other researchers who also examine *n*Ach more closely. While we were initially unaware of the need to include a balanced representation of the three motive categories in building the training data, we know now that there is most likely an advantage for future automation researchers to be more intentional about including datasets from a more diverse range of studies that examine multiple motives.[8]

Our research presents a first look at whether it is possible to build a machine-learning based algorithm to approximate human-coding of implicit motives. Over three unseen datasets, it seems that the computer is able to predict motive scores at a moderately high level of concordance with human-coded motive scores. Moreover, there is evidence of convergent, divergent, and causal validity for machine-predicted scores. Given that coding of implicit motives is a difficult task for humans, the results we have achieved with this research suggests that there is merit in pursuing machine-learning-based automation of implicit motive coding. Several directions already come to mind for future researchers to undertake, such as: Developing larger training datasets with more diverse sources of data, including training data from studies which investigated all three motives in a balanced manner, using deeper machine learning model architectures, and experimenting with more complex contextual embedding models of words and sentences. Should the automation enterprise be successful in the near future, it would benefit motive researchers in terms of reducing the research burden for manpower and time, as well as appeal to interested researchers who are otherwise unfamiliar with implicit motive measurement.

References

Apers, C., Lang, J. W. B., & Derous, E. (2019). Who earns more? Explicit traits, implicit motives and income growth trajectories. *Journal of Vocational Behavior*, *110*(Part A), 214–228. https://doi.org/10.1016/j.jvb.2018.12.004

Awford, J. (2016, March 22). 'We can shoot your wife and frame your mother-in-law': The ad by a small picture framing business BANNED for being 'violent'… or is it just a bad joke? *Daily Mail UK Online*. Retrieved from https://www.dailymail.co.uk/news/article-3504544/Joke-slogan-shooting-wife-banned.html

Allen, C. & Hospedales, T.M. (2019). Analogies explained: Towards understanding word embeddings. *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, PMLR 97, June 2019. abs/1901.09813. URL http://arxiv.org/abs/1901.09813. 1901.09813.

Bai, S., Kolter, J.Z., & Koltun, V. (2018). An empirical evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv, abs/1803.01271*. URL http://arxiv.org/abs/1803.01271. 1803.01271.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B (Mehodological)*, 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Borsboom, D., Mellenbergh, G., & Heerden, J. (2004). The concept of validity. *Psychological review*, 111. 1061-71. 10.1037/0033-295X.111.4.1061.

Brunstein, J. C., & Maier, G. W. (2005). Implicit and self-Attributed motives to achieve: Two separate but interacting needs. *Journal of Personality and Social Psychology*, *89*(2), 205–222. https://doi.org/10.1037/0022-3514.89.2.205

Busch, H. (2018). Power motivation. In Heckhausen & H. Heckhausen (Eds.), *Motivation and Action* (3rd. ed., pp. 335-368). J. Springer International Publishing.

Cheng, W. (2013). Semantic prosody. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics,* pp. 1-7. Wiley-Blackwell.

Dufner, M., Arslan, R.C., & Denissen, J.J.A. (2018). The unconscious side of Facebook: Do online social network profiles leak cues to users' implicit motive dispositions? *Motivation and Emotion*, 42**,** 79–89. https://doi.org/10.1007/s11031-017-9663-1

Firth, John R. (1957). A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis*, pp. 1–32. The Philological Society, Oxford. Reprinted in F. R. Palmer (Ed.) (1968), pp. 168–205. Basil-Blackwell.

Fodor, E. M., & Riordin, J. M. (1995). Leader power motive and group conflict as influences on leader behavior and group member self-affect. *Journal of Research in Personality*, 29, 418–431. https://doi.org/10.1006/jrpe.1995.1024

Fodor, E. M., Wick, D. P., & Conroy, N. E. (2012). Power motivation as an influence on reaction to an imagined feminist dating partner. *Motivation and Emotion, 36*, 301–310. https://doi.org/10.1007/s11031-011-9254-5

Furley, P., Schweizer, G., & Wegner, M. (2019). The power motive as a predictor of receptiveness to nonverbal behavior in sport. *Motivation and Emotion*, 43**,** 917–928. https://doi.org/10.1007/s11031-019-09788-4

Goldberg, Y. (2015, October 2). A Primer on neural network models for natural language processing. Retrieved October 22, 2019, from arXiv:1510.00726v1 [cs.CL].

Hagemeyer, B., Dufner, M., & Denissen, J. (2016). Double dissociation between implicit and explicit affiliative motives: A closer look at socializing behavior in dyadic interactions. *Journal of Research in Personality.* 65. 89-93. 10.1016/j.jrp.2016.08.003.

Hagemeyer, B., Neberich, W., Asendorpf, J. B., & Neyer, F. J. (2013). (In)congruence of implicit and explicit communal motives predicts the quality and stability of couple relationships. *Journal of Personality, 81*(4), 390-402. http://dx.doi.org/10.1111/jopy.12016

Harris, Z.S. (1954). Distributional structure. *Word*, 10(23), 146–162. doi: 10.1080/00437956.1954.11659520.

Hauser, D., & Schwarz, N. (2016). Semantic prosody and judgment. *Journal of Experimental Psychology General*, 145, 882-896. doi:10.1037/xge0000178.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852. http://arxiv.org/abs/1502.01852. 1502.01852.

Heckhausen, H. (1963). *Hoffnung und Furcht in der Leistungsmotivation [Hope and fear components of achievement motivation]*. Anton Hain.

Hofer, J. & Busch, H. (2019). Women in power-themed tasks: Need for power predicts task enjoyment and power stress. *Motivation and Emotion*, *43*, 740-757. doi: 10.1037/a0020053.

Hogenraad, R. (2005). What the words of war can tell us about the risk of war. *Peace and Conflict: Journal of Peace Psychology*, 11, 137-151. doi: 10.1207/s15327949pac1102_2.

Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Sigma Assessment Systems.

Johnson, P. B. (1981). Achievement motivation and success: Does the end justify the means? *Journal of Personality and Social Psychology, 40*(2), 374-375. http://dx.doi.org/10.1037/0022-3514.40.2.374

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., & Kavukcuoglu, K. (2016). Neural machine translation in linear time. *CoRR*, abs/1610.10099. http://arxiv.org/abs/1610.10099. 1610.10099.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural language processing: State of the art, current trends and challenges. *CoRR*, abs/1708.05148. http://arxiv.org/abs/1708.05148. 1708.05148.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746– 1751. Association for Computational Linguistics. doi:10.3115/v1/D14-1181.

Klimt, B., & Yang, Y. (2004). The Enron Corpus: A new dataset for email classification research. *In* J.F. Boulicault, F. Esposito, F. Giannotti, & D. Pedreschi. (eds). *Machine Learning: ECML 2004. Lecture Notes in Computer Science*, vol. 3201. Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163. doi:10.1016/j.jcm.2016.02.012

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5686–5697. doi: 10.1145/2858036.2858529

Kuhl, J. (1978). Situations-, reaction-, and person-related consistency of achievement motive according to Heckhausen TAT. *Archiv Für Psychologie*, 130(1), 37–52.

Kuhl, J., & Scheffer, D. (1999). Der operante multi-motive-test (OMT): Manual [The operant multi-motive-test (OMT): Manual]. University of Osnabrück.

Lea, C., Vidal, R., Reiter, A., & Hager, G.D. (2016). Temporal convolutional networks: A unified approach to action segmentation. *CoRR*, abs/1608.08242. http://arxiv.org/abs/1608.08242. 1608.08242.

Leber, J. (2013, July 2). The immortal life of the Enron e-mails. *Technology Review*. Retrieved from https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/

Li, X., Chen, S., Hu, X., & Yang, J. (2018). Understanding the disharmony between dropout and batch normalization by variance shift. *CoRR*, abs/1801.05134. http://arxiv.org/abs/1801.05134. 1801.05134.

Louw, B. (1993). Irony in the Text or Insincerity in the Writer? — The Diagnostic Potential of Semantic Prosodies. In G. Francis, E. Tognini-Bonelli (Eds.),*Text and technology: In honour of John Sinclair*, pp. 157-176. John Benjamins Publishing.

McClelland, D. C. (1965). Toward a theory of motive acquisition. *American Psychologist, 20*(5), 321-333. http://dx.doi.org/10.1037/h0022225

McClelland, D. C. (1975). Power: The inner experience. Irvington.

McClelland, D. C. (1987). Human motivation. Cambridge University Press.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). The Achievement Motive. Appleton-Century-Crofts.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96(4),* 690-702. http://dx.doi.org/10.1037/0033-295X.96.4.690

Méndez, G., Hinrichs, U., & Nacenta, M. (2017). Bottom-up vs. Top-down: Trade-offs in efficiency, understanding, freedom and creativity with InfoVis tools. *Proceedings of the 2017 CHI conference on human factors in computing systems,* 841-852. 10.1145/3025453.3025942.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781. http://arxiv.org/abs/1301.3781. 1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *CoRR*, abs/1712.09405. http://arxiv.org/abs/1712.09405. 1712.09405.

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: the thematic apperception test. *Archives of Neurology and Psychiatry*, *34*, 289-306.

Mrksic, N., Vulić, I., Séaghdha, D., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., & Young, S. (2017). Semantic specialisation of distributional word vector spaces using monolingual and cross-Lingual constraints. Retrieved from http://arxiv.org/abs/1706.00374 .

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. *CoRR*, abs/1705.08039. URL http://arxiv.org/abs/1705.08039. 1705.08039.

Oxford University Press. (1989). *Oxford English Dictionary*. Oxford University Press.

Pang, J. S., Villacorta, M. A., Chin, Y. S., & Morrison, F. J. (2009). Achievement motivation in the social context: Implicit and explicit hope of success and fear of failure predict memory for and liking of successful and unsuccessful peers. *Journal of Research in Personality*, *43(6),* 1040-1052. http://dx.doi.org/10.1016/j.jrp.2009.08.003

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count (LIWC): LIWC2001. Lawrence Erlbaum Associates.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296–1312. doi: 10.1037/0022-3514.77.6.1296.

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations.

Polysemy. Oxford Reference. Retrieved 17 Oct. 2019, from https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100336222.

Raihala, C., & Hansen, G. (2019). Power stress in primary school children. *Motivation and Emotion*, *43*, 82–92. https://doi.org/10.1007/s11031-018-9724-0

Riach, D. (2019). Determinism in deep learning. *Gpu Technology Conference 2019*. http://bit.ly/dl-determinism-slides-v2 (updated 2019-05-17). Retrieved 17 Oct. 2019, from https://developer.nvidia.com/gtc/2019/video/S9911.

Schönbrodt, F. D., & Gerstenberg, F. X. R. (2012). An IRT analysis of motive questionnaires: The Unified Motive Scales. *Journal of Research in Personality, 46*, 725–742. doi:10.1016/j.jrp.2012.08.010

Safyer, P., Volling, B. L., Schultheiss, O. C., & Tolman, R. M. (2019). Adult attachment, implicit motives, and mothers' and fathers' parenting behaviors. *Motivation Science*, *5*(3), 220–234. https://doi.org/10.1037/mot0000112

Schultheiss, O. C. (2001). *Manual for the assessment of hope of success and fear of failure (English translation of Heckhausen' s need Achievement measure).* Department of Psychology, University of Michigan, Ann Arbor: Unpublished manuscript.

Schultheiss, O. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2013.00748

Schultheiss, O., & Brunstein, J. (Eds.) (2010). *Implicit motives.* Oxford University Press.

Schultheiss, O. C., & Pang, J. S. (2007). Measuring implicit motives. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 322-344). The Guilford Press.

Schultheiss, O. C., Yankova, D., Dirilikvo, B., & Schad, D. J. (2009). Are implicit and explicit motive measures statistically Independent? A fair and balanced test using the Picture Story Exercise and a cue- and response-matched questionnaire measure. *Journal of Personality Assessment, 91*(1), 72-81. http://dx.doi.org/10.1080/00223890802484456

Shen, L., Satta, G., & Joshi, A. (2007). Guided learning for bidirectional sequence classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Retrieved 5 Mar. 2020 from https://www.aclweb.org/anthology/P07-1096

Shrout P.E., & Fleiss J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Pyschological Bulletin*, *86*, 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Smith, M. S. (1968). The computer and the TAT. *Journal of School Psychology*, *6*, 206–214. doi: 10.1016/0022-4405(68)90017-4

Smith, N.A. (2019). Contextual word representations: A contextual introduction. *CoRR*, abs/1902.06006. http://arxiv.org/abs/1902.06006. 1902.06006.

Smith, C. P., Ryan, E. R., & Diggins, D. R. (1972). Moral decision making: Cheating on examinations. *Journal of Personality*, *40*, 640–660

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112*(1), 140-154. http://dx.doi.org/10.1037/0033-2909.112.1.140

Sridharan, M., & Swapna, T.R. (2019). Manipulating attention with Temporal Convolutional Neural Network for offense identification and classification. *Proceedings of the 13th*

International Workshop on Semantic Evaluation, 540–546. Association for Computational Linguistics: Minneapolis, Minnesota, USA. doi:10.18653/v1/S19-2097.

Stoeckart, P.F., Strick, M., Bijleveld, E., & Aarts. (2018). The implicit power motive predicts decisions in line with perceived instrumentality. *Motivation and Emotion*, 42**,** 309–320. https://doi.org/10.1007/s11031-018-9687-1

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis. MIT Press: Cambridge, MA, USA.

Swinger, N., De-Arteaga, M., Heffernan, N. T., Leiserson, M. D. M., & Kalai, A. T. (2018). What are the biases in my word embedding? https://arxiv.org/abs/1812.08769.

Thielgen, M. M., Krumm, S., & Hertel, G. (2015). When being old pays off: Age mitigates adverse effects of low implicit–explicit motive congruency on work motivation. *Journal of Career Assessment, 23(3),* 459–480. https://doi.org/10.1177/1069072714547613

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762. http://arxiv.org/abs/1706.03762. 1706.03762.

Vulic, I., & Mrksic, N. (2017). Specialising word vectors for lexical entailment. *CoRR*, abs/1710.06371. http://arxiv.org/abs/1710.06371. 1710.06371.

Wegner, M., Bohnacker, V., Mempel, G., Teubel, T., & Schüler, J. (2014). Explicit and implicit affiliation motives predict verbal and nonverbal social behavior in sports competition. *Psychology of Sport and Exercise*, *15*, 588-595. 10.1016/j.psychsport.2014.06.001.

Weinberger, J., Cotler, T., & Fishman, D. (2010). The duality of affiliative motivation. In O. C. Schultheiss & J. C. Brunstein (Eds.), *Implicit motives* (pp. 71–88). Oxford University Press: Oxford, England. 10.1093/acprof:oso/9780195335156.003.0003.

Weitzel, L., & Prati, R., & Aguiar, R. (2016). The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques? *In* W. Pedrycz & S-M. Chen (Eds). *Sentiment Analysis and Ontology Engineering*. doi: 10.1007/978-3-319-30319-2_3.

Whitley, B. E. J. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235–274.

Winter, D. G. (1973). The power motive. Free Press.

Winter, D. G. (1994). Manual for scoring motive imagery in running text. Unpublished instrument, University of Michigan, Ann Arbor.

Winter, D. G. (2010). Why achievement motivation predicts success in business but failure in politics: The importance of personal control. *Journal of Personality*, *78(6),* 1637–1667. https://doi.org/10.1111/j.1467-6494.2010.00665.x

Wolf, M., Horn, A., Mehl, M., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count [Computer-aided quantitative text analysis: Equivalence and reliability of the German adaptation of the Linguistic Inquiry and Word Count]. *Diagnostica, 2*, 85–98. doi: 10.1026/0012-1924.54.2.85

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 5365–5372. http://hdl.handle.net/10125/50557.

Yu, L-C, Wang, J., Lai, K.R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural*

Language Processing, 545–550. Association for Computational Linguistics: Copenhagen, Denmark.

Zhang, Y., & Wallace, B.C. (2016). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv: Computation and Language (cs.CL)*. URL arXiv:1510.03820v4.

Table 1. Characteristics of individual datasets used in the machine learning training and validation dataset.

Dataset	N	Sentences	Words	Country	Sample characteristics	# PSE pictures	Motive of interest	Other major variable of interest	Source
1	205	8101	111299	SG	undergraduate students, 69.7% female, age (*M* = 21.55 years, *SD* = 1.73)	7	*n*Ach	hope of success and fear of failure	Ramsay, J. E. (2014). Refining the picture story exercise : towards a better understanding of hope, fear, and the achievement motive. Doctoral thesis, Nanyang Technological University, Singapore.
2	67	2975	40667	SG	undergraduate students, 54.7 % female, age (*M* = 21.52 years, *SD* = 1.83)	8	*n*Ach	hope of success and fear of failure	Ramsay, J. E., & Pang, J. S. (2013). Set ambiguity: A key determinant of reliability and validity in the picture story exercise. Motivation and Emotion, 37(4), 661-674. http://dx.doi.org/10.1007/s11031-012-9339-9
3	70	3358	45545	SG	undergraduate students, 54.7 % female, age (*M* = 21.52 years, *SD* = 1.83)	8	*n*Ach	hope of success and fear of failure	Ramsay & Pang, 2013
4	67	3252	42496	SG	undergraduate students, 54.7 % female, age (*M* = 21.52 years, *SD* = 1.83)	8	*n*Ach	hope of success and fear of failure	Ramsay & Pang, 2013
5	124	6111	91766	SG	undergraduate	8	*n*Ach	hope of	Ramsay, 2014

					students, 81.4% female, age (*M* = 20.86 years, *SD* = 1.80)			success and fear of failure	
6	131	7399	116744	USA	undergraduate students, 42 female, mean age = 19 years	8	*n*Pow, *n*Ach, *n*Aff		Schultheiss, O. C. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. Frontiers in Psychology, 4. doi: 10.3389/fpsyg.2013.00748 [Study 1]
7	32	1413	19841	USA	undergraduate students, 17 female, mean age = 20 years	6	*n*Pow, *n*Ach, *n*Aff		Schultheiss, 2013 [Study 2]
8	49	1296	20684	SG	undergraduate students, 53% female, age (*M* = 20.5, *SD* = 1.23)	6	*n*Pow, *n*Ach, *n*Aff	mindfulness	Pang, J.S. & Kang, N.Q. (2016). Effects of a Brief Short-Term Self-Administered Mindfulness-Based Intervention Program on Young Adults' Life Satisfaction, Psychological Need Satisfaction, and Attentional Control. The Stockholm Criminology Symposium, Stockholm, Sweden.
9	147	2928	42618	SG	undergraduate students, 43% females, age (*M = 21.7, SD* =1.67)	4	NA	music appreciation	unpublished dataset
10	514	13334	197551	USA	Mturk workers,	6	*n*Pow,	volunteering	Ngo, T.A. (2019). Developing a

					70% female, 80% university students, age range 18 - 25 years		*n*Aff		hierarchical model of personality and motivation to predict youth volunteerism: A cross-culture study. Doctoral thesis, Nanyang Technological University, Singapore.
11	478	15514	239159	SG	undergraduate students, 72% female, age range 18 - 25 years	6	*n*Pow, *n*Aff	volunteering	Ngo, 2019
Total	1884	65681	968370						

Notes: USA = United States of America; SG = Singapore. 8,449 sentences (107,401 words) were also extracted from the Enron Email Corpus and added to the training data. These sentences originally appeared in email thread conversations used in UD2 but were replies written by other Enron employees and not by the target individuals in UD2.

Table 2. Training and Validation Accuracy/Loss of Models

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Epochs
CNN	0.9961	0.1430	0.9940	0.0241	100
CNN2D	0.9922	0.2305	0.9904	0.0331	40
TCN	0.9479	1.2586	0.9452	0.1463	20

Note: CNN = one-dimensional convolutional neural network; CNN2D = two-dimensional convolutional neural network; TCN = temporal convolutional network.

Table 3. Correlations between human-coded achievement, affiliation, and power motive scores and their respective corresponding computer-predicted motive scores for UD1 for three machine-learning architectures.

Model/ Motive	*N*Achievement	*N*Affiliation	*N*Power
CNN	r = .50, p = .001, .46 <95% CI< .54	r = .46, p = .001, .42 <95% CI< .50	r = .31, p = .001, .26 <95% CI< .36
CNN2D	r = .53, p = .001, .49 <95% CI< .56	r = .48, p = .001, .44 <95% CI< .52	r = .40, p = .001, .35 <95% CI< .44
TCN	r = .50, p = .001, .46 <95% CI< .54	r = .43, p = .001, .39 <95% CI< .48	r = .40, p = .001, .36 <95% CI< .45

Note: CNN = one-dimensional convolutional neural network; CNN2D = two-dimensional convolutional neural network; TCN = temporal convolutional network.

Table 4. Correlations between human-coded achievement, affiliation, and power motive scores and their respective corresponding computer-predicted motive scores for UD2 for three machine-learning architectures.

Model/ Motive	*N*Achievement	*N*Affiliation	*N*Power
CNN	r = .84, p = .001, .61 <95% CI< .94	r = .56, p = .015, .13 <95% CI< .82	r = .68, p = .002, .32 <95% CI< .87
CNN2D	r = .86, p = .001, .66 <95% CI< .95	r = .63, p = .006, .22 <95% CI< .85	r = .56, p = .015, .13 <95% CI< .82
TCN	r = .42, p = .084, -.06 <95% CI< .74	r = .75, p = .001, .43 <95% CI< .90	r = .45, p = .059, -.02 <95% CI< .76

Note: CNN = one-dimensional convolutional neural network; CNN2D = two-dimensional convolutional neural network; TCN = temporal convolutional network.

Table 5. Correlations between human-coded scores for hope of success and fear of failure motivation and computer-predicted scores for achievement, affiliation, and power motivation for UD3.

Model/motive	Computer predicted *N*Achievement	Computer predicted NAffiliation	Computer predicted *N*Power
CNN			
Human-coded HS	r = .50, p = .002, .22 <95% CI< .71	r = -.20, p = .239, -.49 <95% CI< .13	r = -.19, p = .268, -.48 <95% CI< .14
Human-coded FF	r = -.31, p = .059, -.58 <95% CI< .01	r = .25, p = .129, -.08 <95% CI< .53	r = .10, p = .569, -.23 <95% CI< .41
CNN2D			
Human-coded HS	r = .14, p = .422, -.19 <95% CI< .44	r = -.10, p = .550, -.41 <95% CI< .23	r = -.21, p = .220, -.50 <95% CI< .12
Human-coded FF	r = -.26, p = .126, -.53 <95% CI< .07	r = -.06, p = .718, -.37 <95% CI< .26	r = .14, p = .422, -.19 <95% CI< .44
TCN			
Human-coded HS	r = .49, p = .002, .21 <95% CI< .70	r = -.11, p = .530, -.42 <95% CI< .22	r = -.22, p = .184, -.50<95% CI< .11
Human-coded FF	r = -.24, p = .15, -.52 <95% CI< .09	r = .16, p = .333, -.17 <95% CI< .46	r = .22, p = .200, -.11 <95% CI< .50

Note: HS = Hope of success motivation. FF = Fear of failure motivation.

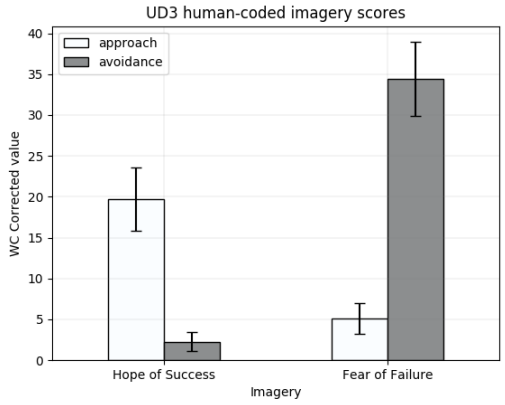

Figure 1a: Human-coded motive imagery after achievement goal visualization exercise

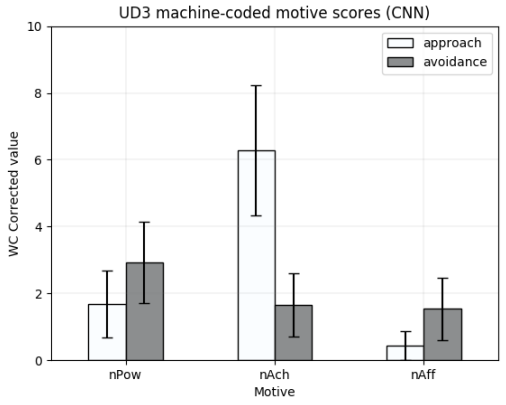

Figure 1b: CNN predictions of motives after achievement goal visualization exercise

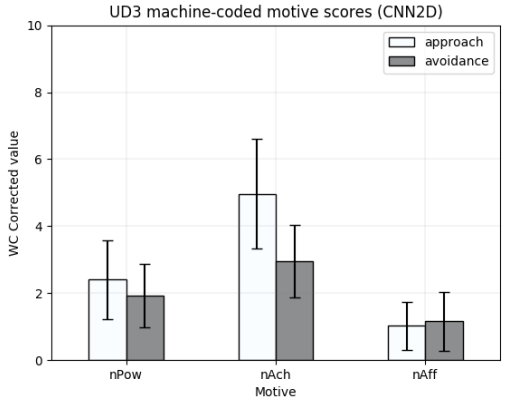

Figure 1c: CNN2D predictions of motives after achievement goal visualization exercise

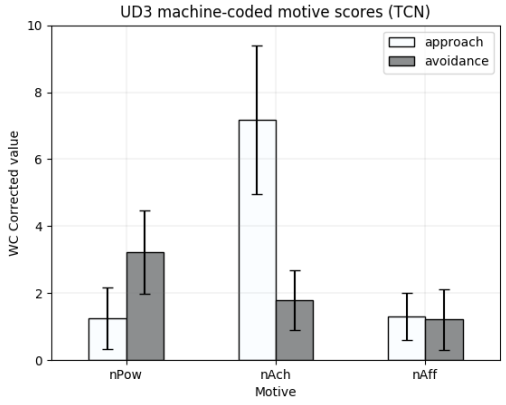

Figure 1d: TCN predictions of motives after achievement goal visualization exercise

Appendix

Sources consulted for list of "misbehavers" and "non-misbehavers" in Enron Email Corpus

Note: A time span of two calendar years was used when consulting sources to determine the role of players that were involved during this time period. Specifically, 1999 was used as a starting point when Andy Fastow created the first of two partnerships that bought poorly performing Enron assets and to hedge risky investments. The end date was when the actual collapse took place when Enron Corporation filed for bankruptcy on December 2, 2001.

Book sources:

Brewer, L., & Hansen, M. S. (2002). House of cards: Confessions of an Enron executive. Virtualbookworm.com.

Bryce, R. (2002). Pipe dreams: Greed, ego, jealousy and the death of Enron. Public Affairs.

Eichenwald, K. (2005). Conspiracy of fools: A true story. Broadway Books.

Fox, L. (2004). Enron: The Rise and Fall. John Wiley and Sons.

McLean, B., & Elkind, P. (2003). The smartest guys in the room: the amazing rise and scandalous fall of Enron. Portfolio.

Olson, C.K. (2008). The Whole Truth ... so Help Me God: An Enlightened Testimony from Inside Enron's Executive Offices. Tate Pub and Enterprises LLC.

Appendix - Continued

Smith, R., & Emshwiller, J. R. (2003). 24 days: How two Wall Street Journal reporters discovered the lies that destroyed faith in corporate America. Harper Business.

Swartz, M., & Watkins, S. (2013). Power failure: The inside story of the collapse of Enron. Crown Business.

News sources:

Barrionuevo, A. (2006, January 29). 10 Enron Players: Where They Landed After the Fall. The New York Times. Retrieved July 11, 2017, from http://www.nytimes.com/2006/01/29/business/businessspecial3/10-enron-players-where-they-landed-after-the-fall.html

Enron: Key players. (2002, January 12). The Guardian. Retrieved July 11, 2017, from https://www.theguardian.com/business/2002/jan/13/corporatefraud.enron

Fast Facts: Key Enron Players. (2004, July 8). Retrieved from http://www.foxnews.com/story/2004/07/08/fast-facts-key-enron-players.html

Major Players in the Enron Trial. (2006). The Wall Street Journal. Retrieved July 11, 2017, from http://online.wsj.com/public/resources/documents/info-enrontrial-0602.html

Rosten, E. (2002, February 4). The Enron Players. Time. Retrieved July 11, 2017, from http://content.time.com/time/magazine/article/0,9171,1001767,00.html

The Enron cast: Where are they now? (n.d.). Financial News. Retrieved July 11, 2017, from https://www.fnlondon.com/articles/enron-ten-years-on-where-they-are-now-20111201

[1] An example of a potential use case for automated coding can be found at www.implicitmotives.com.

[2] To our knowledge, there are only two available investigations on automating implicit motive coding using a machine-learning approach. Additionally, there has been some work by Felix Schönbrodt which was not available to us at the time we submitted this manuscript. Both the available investigations are student theses and were unpublished at the time of this writing. In his Ph.D. thesis, Halusic (2015) attempted to develop an automated coding system for *n*Ach by creating a set of synonym-based word vectors. His approach can be classified as a variant of the bag-of-words approach, which is in contrast to our more data-driven approach, since the features of our models were not built around particular sets of target words. While Halusic was able to obtain *r*s in the .5 region between machine-coded and human-coded scores, he was unable to demonstrate either predictive or causal validity for model-predicted scores. In a masters thesis on automating PSE coding for the Winter (1994) motives, Adler (2017)'s main goal was to arrive at a binary classification system, namely, to predict whether a text was generated by someone in the motive-arousal condition or in the non-aroused ("control") condition in an experiment. Thus, his research objective was different from ours, in that he did not seek to develop a machine-learning algorithm that could predict the motive score of an individual in an unseen dataset.

[3] Regarding this second advantage, while it would still be possible for human coders to code archival material, the financial cost and time burden of relying on human researchers to code unstructured texts (i.e., material that was not generated in the context of a standardized motive measure such as the PSE) would be much higher than feeding such texts into a machine.

[4] Due to conflicts with study timing and coder availability, we were unable to obtain motive scores for the same set of 30 stories from the other 3 coders, however each of these three coders individually accounted for approximately 5% of the dataset.

[5] https://fasttext.cc/docs/en/english-vectors.html

[6] Ring and Pang (2017) presented preliminary findings from machine-learning experiments at a talk at Friedrich-Alexander University in Erlangen, Germany. They inserted part-of-speech information using WordNet (2010), employed a bag-of-words approach, and tested 11 machine-learning algorithms using 10-fold nested cross-validation. Their findings indicated that part-of-speech information only improved validation accuracy minimally, on average by 4%.

[7] https://osf.io/52ntj/?view_only=68b37d6ddb87445e8144aa3f1ce89db3

[8] To aid future automation research, the structure of the three models we built, the unseen datasets, and a Python script to replicate the results we report are available here: https://osf.io/563xn/?view_only=6870e14b364743a688ff17fea80f2c59