Top-Down Influence? Predicting CEO Personality and Risk Impact from Speech Transcripts

Anonymous ACL submission

Abstract

How much does a CEO's personality impact the performance of their company? Management theory posits a great influence, but it is difficult to show empirically—there is a lack of publicly available self-reported personality data of top managers. Instead, we propose a text-based personality regressor based on crowd-sourced Myers–Briggs Type Indicator (MBTI) assessments. The ratings have a high internal and external validity and can be predicted with moderate to strong correlations for three out of for dimensions. Providing evidence for the *upper echelons theory*, we demonstrate that the predicted CEO personalities have explanatory power of financial risk.

1 Introduction

How much influence does the personality of a chief executive officer (CEO) have on their company's performance? The personal news and antics of famous CEOs like Elon Musk, Jeff Bezos, or Bill Gates make headlines, and their personalities sometimes generate a cult-like following. But what measurable effect do they really have? The *upper echelons theory* (Hambrick and Mason, 1984) suggests that the personalities of CEOs also reflect in the organizational outcomes of their companies. However, presumably due to the lack of labeled data, no supervised models exist to detect CEOs' personalities from text and infer their effect on the financial performance of companies. In this paper, we close this research gap by presenting the first Transformer-based model to predict the impact of CEOs' Myers–Briggs Type Indicator (MBTI) personality on financial risk.

Ideally, personality is assessed with self-reported questionnaires. However, it is technically infeasible to request executives such as Elon Musk to fill out targeted pen and paper questionnaires. We were therefore motivated to explore crowd-sourced data. This approach is supported by past research showing that observer reports are an inexpensive and valid alternative to self-reports (Vazire, 2006), as they usually agree with them (Kim et al., 2019), and are particularly suitable for the assessment of top management personality (Connolly et al., 2007).

The dominant personality model is the Big 5, which presents personality on a continuum along the dimensions *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* (McCrae and John, 1992). The available data source we use lacks Big 5 ratings, so as proxy, we explore the MBTI (Briggs-Myers and Myers, 1995), which has been shown to correlate along the main dimensions with the Big 5 (McCrae and Costa, 1989; Furnham, 1996; Furnham et al., 2003). This model represents personality via the categories *extraversion–introversion*, *sensing–intuition*, *thinking–feeling*, and *judging–perceiving*. Addressing methodological criticism of the MBTI (McCrae and Costa, 1989), we

- explore an alternative MBTI representation as a vector of continuous values (§3.1);
- find a high internal and external validity of this measure (§3.1);
- show that it can be predicted from text (§3.3);
- and demonstrate that it is predictive of financial risk (§4.3).

Overall, our findings lend empirical support to the *upper echelons theory* of management.

2 Background and Related Work

Various personality measures exist in the literature. This section describes the personality model we explore (MBTI), the de-facto standard model (Big 5), and approaches to predict both representations of personality from text.

2.1 MBTI

The MBTI is named after Katherine Cook Briggs and Isabel Briggs Myers. They developed it based on the work of the analytical psychologist Carl Jung (Briggs-Myers and Myers, 1995). The MBTI classifies personalities binarily along the following axes:

- *extraversion* vs. *introversion* (E–I): describing an out- or inward-oriented social attention;
- *sensing* vs. *intuition* (S–N): information processing based on perceivable/known facts or conceptualization and imagination;
- *thinking* vs. *feeling* (T–F): decision-making based on logic and rationality or emotions and empathy;
- *judging* vs. *perceiving* (J–P): quick judgement and organized action or observation and improvisation on-the-go.

Combined, the four labels form one of 16 personality types (e.g., "ENTJ"). The MBTI is widely used in human resources management and by laypeople as a tool for self-exploration.

Psychological literature, however, has called assumptions of the MBTI into question. For example, McCrae and Costa (1989) find no evidence that personality can be binarized or distinguished into 16 different types. In addition, they find moderate to strong correlations between MTBI and Big 5 (McCrae et al., 2010), which is described in greater detail below (§2.2). We re-assess these correlations in our dataset and explore a continuous representation of the MBTI in line with the Big 5.

MBTI Prediction from Text In a literature study on text-based personality detection and a subsequent annotation study, Štajner and Yenikent (2020, 2021) conclude that predicting the MBTI from textual data is a difficult task. They hypothesize that this is due to the theoretical and qualitative origin of the index, which distinguishes it from the empirical and quantitative Big 5. In particular, the dimensions *sensing* vs. *intuition* (S–N) and *judging* vs. *perceiving* (J–P) depend on behavioral rather than linguistic signals (Štajner and Yenikent, 2020, p. 6291).

In a field survey of project managers, Cohen et al. (2013) show that managers are significantly more often of the *intuitive* (N) and *thinking* (T) type than the general population. We observe a similar pattern in our dataset (§3.1, Figure 2). Classifying the MBTI of Twitter users based on count-based features, gender, and tweet n-grams, Plank and Hovy (2015) outperform a majority class baseline for the E–I and the T–F dimensions. Gjurković and Šnajder (2018) predict the self-reported MBTI of Redditors with support vector machine (SVM) and multilayer perceptron (MLP) models based on linguistic and activity-level features. Their model outperforms a majority class baseline across all dimensions with the best results for E–I, followed by S–N, J–P, and T–F.

We compare the best-performing approaches identified by prior MBTI prediction studies (n-grams and Linguistic Inquiry and Word Counts (LIWC) dictionaries with SVMs and MLPs) to Transformer architectures. Furthermore, we consider a different domain (spoken financial disclosures) and perform a regression instead of a classification.

2.2 Big 5

The Big 5 are the established psychometric model. Here, personality is represented as a continuum along the five axes *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* (McCrae and John, 1992).

Big 5 Prediction from Text As part of the *myPersonality* project, Kosinski et al. (2015) find that liked Facebook pages predict Big 5, IQ, and other personal characteristics to varying degrees. Mairesse et al. (2007) create a text-based Big 5 prediction tool based on student essays and speech recordings.

Benischke et al. (2019) show that CEOs' Big 5 personalities moderate the relationship between CEO compensation and risk-taking. Hrazdil et al. (2020) use IBM WATSON PERSONALITY INSIGHT to predict the Big 5 of C-level executives in earnings calls and find that an executive's personality is associated with their risk tolerance and company audit fees. Harrison et al. (2020) find that CEO Big 5 are related to perceived firm risk and shareholder value. Another finding is that CEO *conscientiousness* moderates the effect of financial risk on returns positively, while the opposite holds for *extroversion* and *neuroticism*.

Different to these approaches, we focus on the MBTI rather than the Big 5. We create the first supervised model to predict CEOs' MBTI personality from text by collecting a new dataset of crowd-annotated MBTI profiles. This sets us apart from prior work using unsupervised approaches trained on out-of-domain corpora.

ELON MUSK (CEO): Thank you. So Q1 ended up being a strong quarter despite many challenges in the final few weeks. This is the first time we have achieved positive GAAP net income in a seasonally weak first quarter. Even with all the challenges, we achieved a 20% automotive gross margin, excluding regulatory credits, while ramping 2 major products. What we've learned from this is that—we've obviously learned a lot here.

Figure 1: Excerpt of Tesla's Q1 2020 earnings call.

3 Personality Prediction

Using transcribed speech data as an input, we predict the MBTI personality of CEOs via text regression. The following sheds light on the dataset collection and validation, methodology, and results.

3.1 Dataset Curation

For this task, we collect data from two sources: (1) text data and (2) crowd-sourced personality data.

Text Data We obtain 88K earnings call transcripts spanning years 2002–2020 from REFINITIV EIKON.[1] Earnings calls are quarterly teleconferences consisting of a scripted presentation and a spontaneous questions-and-answers (Q&A) session, in which CEOs such as Elon Musk answer open questions of banking analysts. Due to the improvised nature of these answers, earnings calls are particularly suitable for detecting personal style (Malhotra et al., 2018). Figure 1 shows an excerpt of Tesla's Q1 earnings call in 2020.

Given the dialogue nature of the calls, we need to map utterances to individual CEOs as we are not interested in the personality of the analysts. We identify CEO names with regular expressions and minimal preprocessing (e.g., stripping middle name initials or titles). Next, we require a match with the executive database COMPUSTAT EXECUCOMP for age and gender data (§4.2),[2] reducing our initial sample to 22K calls and 1.7K CEOs. For these, we retrieve all of their utterances in the presentation and the Q&A session of the calls.

Personality Data We obtain MBTI personality labels for the CEOs from PERSONALITY DATABASE,[3] which provides crowd-sourced personality profiles for celebrities, managers, and other noteworthy people. While each profile features vote results for the four dimensions of the MBTI, a minority also contains results for the Big 5. We find that 32 CEOs (e.g., Elon Musk and Steve Jobs) from our earnings call sample have at least three MBTI votes available. The minimum, maximum, and mean votes per CEO are 3, 1.8K, and 140, respectively. These CEOs participate in a total of 736 earnings calls. Table 2 gives the descriptive statistics of the merged text–personality data, and Table 1 contains example CEOs from our dataset across the MBTI.

Instead of representing each personality as one of 16 types, we represent each personality profile as a vector of 4 continuous variables ranging from 0 to 1, based on the crowd-sourced votes. We normalize the votes for the right-hand side of a scale s by the total votes:

$$\text{personality}_s = \frac{\text{votes}_{1,s}}{\text{votes}_{0,s} + \text{votes}_{1,s}}. \quad (1)$$

For example, for the E–I scale, we divide the votes for introversion (I) by the total votes for E and I. The resulting number is thus the likelihood of the CEO being intro- or extroverted. This representation is similar to the Big 5 model (excluding the *neuroticism* dimension) and allows for a more granular representation of personality than the usual operationalization of the MBTI. Figure 2 shows the distributions of the such obtained continuous labels. Most CEOs in our sample are rather *extroverted*, *intuitive*, *thinking*, and *judging* (Figure 2), which corresponds to the ENTJ "Decisive Strategist" MBTI type.[4]

Internal Validation To assess the validity of the crowd-sourced votes, we analyze the inter-annotator agreement between the MBTI raters of the 32 CEOs (Table 3). While p_a is high with values ranging between ca. 80 and 90%, Krippendorff's α (Krippendorff, 2013) yields only slight to moderate values between 0.14 and 0.43. Quarfoot and Levine (2016) call this phenomenon the "frequency distribution paradox," where highly skewed label distributions combined with high percentage agreements can lead to low values of α. As measures robust to this undesirable property, they suggest the Brennan–Prediger coefficient κ_{bp} (Brennan and Prediger, 1981) and Gwet's γ (Gwet, 2008), which in our case yield a high IAA between 0.60 to 0.88.

[1] https://eikon.thomsonreuters.com/index.html
[2] https://wrds-www.wharton.upenn.edu
[3] https://www.personality-database.com/
[4] https://eu.themyersbriggs.com/en/tools/MBTI/MBTI-personality-Types/ENTJ

MBTI	CEO Examples
Extraversion	Steve Jobs (Apple), Lisa Su (AMD), Mary Barra (General Motors)
Introversion	Rupert Murdoch (Fox), Mark Zuckerberg (Facebook), Sheldon Adelson (Las Vegas Sands)
Sensing	Jack Dorsey (Twitter), John Schnatter (Papa John's), Marcus Lemonis (Camping World)
Intuition	Marissa Mayer (Yahoo), Bob Iger (Disney), Evan Spiegel (Snap)
Thinking	Elon Musk (Tesla), Tim Cook (Apple), Steve Ballmer (Microsoft)
Feeling	Sundar Pichai (Google), Howard Schultz (Starbucks), Naveen Jain (Infospace)
Judging	Jeff Bezos (Amazon), Larry Ellison (Oracle), Martha Stewart (Martha Stewart Living)
Perceiving	Larry Page (Alphabet), Martin Shkreli (Retrophin), Donald Trump (Trump Entertainment)

Table 1: CEO examples for each MBTI dimension from our dataset.

Unit	Σ_x	$\bar{x}$	$\min_x$	$\max_x$
utterances	13,183	17.91	2	124
sentences	111,781	151.88	2	563
tokens	2,526,473	3432.71	22	9968

Table 2: Statistics of the CEO–call data considered for the personality prediction. Sums (Σ_x), averages ($\bar{x}$), minima ($\min_x$), and maxima ($\max_x$) are computed across all earnings calls ($n = 736$).

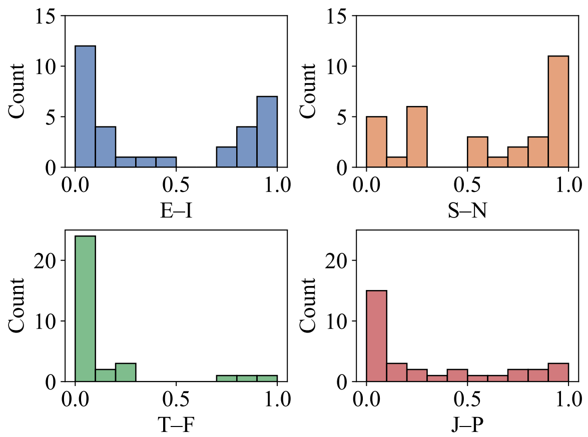


Figure 2: Label distributions for all CEOs considered in the personality prediction ($n = 32$) across the MBTI dimensions *extraversion–introversion* (E–I), *sensing–intuition* (S–N), *thinking–feeling* (T–F), and *judging–perceiving* (J–P).

MBTI	p_a	α	κ_{bp}	γ
E–I	87.45	0.40	0.75	0.76
S–N	80.20	0.43	0.60	0.62
T–F	83.33	0.14	0.67	0.71
J–P	90.62	0.17	0.81	0.88

Table 3: IAA per MBTI dimension in terms of percentage agreement (p_a), Krippendorff's α, Brennan–Prediger coefficient (κ_{bp}), and Gwet's γ.

External Validation To get a notion of external validity, we construct a correlation matrix between the crowd-based MBTI and Big 5 votes of

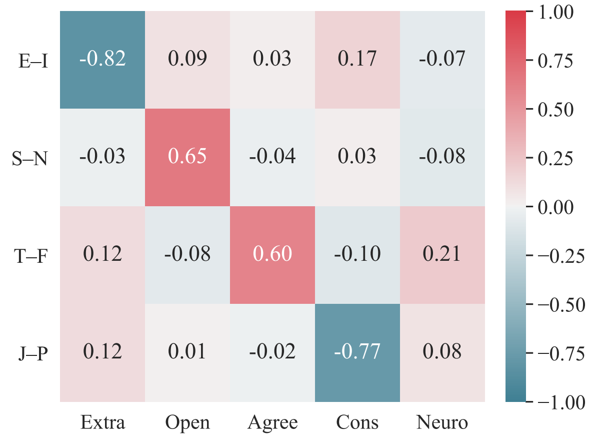


Figure 3: Correlation of MBTI (y-axis) and Big 5 (x-axis) scales for all profiles on the PERSONALITY DATABASE with at least three votes ($n = 2.2$K).

all 2.2K profiles with more than three votes available on PERSONALITY DATABASE (Figure 3). According to McCrae and Costa (1989) and subsequent work (Furnham, 1996; Furnham et al., 2003), strong correlations should exist between MBTI *introversion* and Big 5 *extraversion* ($r = -0.74$) as well as between MBTI *intuition* and Big 5 *openness* ($r = 0.72$). Furthermore, moderate correlations should exist between MBTI *feeling* and Big 5 *agreeableness* ($r = 0.44$) and between MBTI *perceiving* and Big 5 *conscientiousness* ($r = -0.49$). Our results confirm the findings of McCrae and Costa (1989) with similar correlations in the first two rows and stronger correlations in the third and fourth rows. This is most likely due to our increased sample size ($n = 2.2$K vs. $n = 267$).

3.2 Methodology

For each of the 32 CEOs appearing in 736 CEO–call instances, we compare sparse approaches suggested by past literature to Transformer architectures for a regression of MBTI personality.[5]

Data Split We apply an 80:10:10 split to our data to obtain separate training ($n = 568$), validation ($n = 84$), and test sets ($n = 84$). To avoid overfitting, we use sklearn's `GroupShuffleSplit` with the CEO names as group splitting criterion, i.e., we split the data such that no CEO present in the training data appears in the validation or test data.

Normalization Given the highly skewed distributions, after the train–validation–test split, we apply a Box-Cox transformation (Box and Cox, 1964) to y with the following formula:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln(y) & \text{for } \lambda = 0. \end{cases} \quad (2)$$

We obtain λ via maximum-likelihood estimation. The resulting transformation makes the four label distributions more Gaussian-like by stabilizing variance.

Transformers We explore cased-vocabulary $\text{BERT}_{\text{base}}$ (12-layer, 768-hidden, 12-heads, 109M parameters) (Devlin et al., 2019) and $\text{RoBERTa}_{\text{base}}$ (12-layer, 768-hidden, 12-heads, 125M parameters) (Liu et al., 2019) models with a linear regression head. The models are trained with a maximum sequence length of 512 and a sliding window approach. We determine the training batch size and learning rate by running a Bayesian optimization over the grid of batch sizes $b \in \{32, 64, 128, 256\}$ and learning rates $l \in [0, 5 \times 10^{-5}]$.[6] We train a model for up to 10 epochs and early stopping with a patience of one epoch. For each of the four MBTI dimensions, we evaluate 40 combinations of hyperparameters and select the model with minimal loss on the validation set. Different to the mean-squared error (MSE) loss, which is implemented per default in the 🤗 Transformers (Wolf et al., 2020) regressors, we minimize the L1 or alternatively called mean absolute error (MAE) loss, which is less sensitive to outliers.

Sparse Methods We also explore the sparse representations suggested by Plank and Hovy (2015) and Gjurković and Šnajder (2018). These include term frequency–inverse document frequency (tf–idf) vectors with n-grams of length $n \in \{1, 2, 3\}$ and dictionary features across all dimensions of LIWC 2015 (Pennebaker et al., 2015) fed into SVM and three-layer MLP regressors. We compare all possible feature–algorithm combinations with respect to their average MAE on the validation set and select the combination with the lowest error (SVM with trigram tf–idf).

Evaluation The final model performance is evaluated by inspecting the correlation and error between test set ground truth and prediction. As measures, we explore the linear correlation coefficient (i.e., Pearson's r) and the rank correlation coefficients Spearman's ρ and Kendall's τ. Instead of linear relationships, the latter two measure monotonic relationships and are more robust to outliers. In addition, we consider the error measure MAE, which is the minimized loss function of the Transformers. In case of a tie, we give precedence to τ, as this measure is least sensitive to outliers and particularly suited for small sample sizes.

3.3 Results and Discussion

The results of the personality prediction task are depicted in Table 4. An SVM performs competitive, especially for the dimensions E–I ($\tau = 0.44$) and S–N ($\tau = 0.20$). While the SVM outperforms BERT for all dimensions except for J–P, RoBERTa achieves the best results in most cases.

The largest correlations across all models are achieved for the *extraversion–introversion* (E–I) scale with strong linear and rank correlations for the RoBERTa regressor ($r = 0.70$, $\rho = 0.66$). This result is not surprising, as distinguishing between *extra-* and *introverted* CEOs based on linguistic style should be comparably easy. This is followed by the *sensing–intuition* (S–N) scale with moderate to strong correlations ($r = 0.45$, $\rho = 0.53$) and the *judging–perceiving* (J–P) scale with weak to moderate correlations ($r = 0.40$, $\rho = 0.36$). The worst results are obtained for the *thinking–feeling* (T–F) scale, with the SVM and RoBERTa obtaining correlations of around zero and BERT even obtaining weak to moderate negative correlations. There are several possible explanations for this: Conceptually, it could be the case that this dimension simply can not be captured by analyzing linguistic data. Furthermore, the predictive power could be low due to the comparably small sample size. Lastly, we hypothesize that the skewness of the label distribu-

[5]The supplementary material contains our implementation and the earnings call identifiers. Using those, our corpus can be re-assembled from REFINITIV EIKON, SEEKING ALPHA, or alternative sources.

[6]Final hyperparameter choices and results on our validation set can be found in Appendices A and B.

MBTI	Model	r	ρ	τ	MAE
E–I	SVM	0.57	0.58	0.44	0.38
	BERT	0.39	0.35	0.22	0.59
	RoBERTa	**0.70**	**0.66**	**0.52**	**0.34**
S–N	SVM	0.32	0.36	0.20	0.30
	BERT	0.08	0.23	0.16	0.46
	RoBERTa	**0.45**	**0.53**	**0.38**	**0.28**
T–F	SVM	**0.03**	−0.12	−0.08	**0.37**
	BERT	−0.47	−0.41	−0.27	0.41
	RoBERTa	0.01	**−0.10**	**−0.07**	0.39
J–P	SVM	−0.05	0.04	0.02	**0.35**
	BERT	0.39	**0.38**	**0.25**	0.52
	RoBERTa	**0.40**	0.36	0.21	0.36

Table 4: Correlation results of the personality regression task. CEO personality is predicted across the MBTI dimensions *extraversion–introversion* (E–I), *sensing–intuition* (S–I), *thinking–feeling* (T–F), and *judging–perceiving* (J–P). SVM is trained on trigram tf–idf vectors, $\text{BERT}_{\text{base}}$, and $\text{RoBERTa}_{\text{base}}$ on text. Best results in bold.

tion, which was the highest across all MBTI dimensions for the T–F scale (Figure 2), has contributed to the weak performance. This warrants further research exploring whether our findings hold for larger datasets with less skewed label distributions.

Štajner and Yenikent (2020) hypothesize that the S–N and J–P dimensions should theoretically make for the worst candidates in a text-based personality prediction task since they capture behavioral rather than linguistic dimensions of personality. Although our regressors perform worse on these dimensions than for the *extraversion–introversion* dimension, they still achieve moderate to strong correlations, showing that even the more latent dimensions of personality can be predicted from text.

Qualitative Analysis As a brief qualitative analysis, we use Shapley Additive Explanations (SHAP) developed by Lundberg and Lee (2017) to visualize the personality predictions for an exemplary text snippet across the four MBTI dimensions with heatmaps (Figure 4). The analyzed personality is Elon Musk, who, according to the crowd votes, scores high on E–I (*introversion*) and on S–N (*intuitive*), low on T–F (*thinking*), and medium on J–P (*judging/perceiving*). Particularly interesting are the results for T–F (Figure 4c), where statements related to factual content are related to increased T, and interpretative statements (e.g., "[e]ven with all the challenges") to increased F.

4 Risk Regression

According to *upper echelons theory* (Hambrick and Mason, 1984), strategic choices and performance measures of organizations can be predicted by characteristics of their top management. As a use case for our personality prediction task, we explore whether we can find empirical support for this theory. We hypothesize that having a different personality to most CEOs (i.e., ENTJ, see Figure 2 and Cohen et al. (2013)) should translate into increased financial risk.

4.1 Dataset Curation

As a basis for the risk regression task, we take the sample of 22K earnings calls and merge it with data obtained from the databases CRSP, IBES, and COMPUSTAT EXECUCOMP, which we access via WRDS.[7] To measure risk, we calculate the stock return volatility in the business week following each call as a label. We use the sample standard deviation of logarithmic stock returns for more robust measures. As features, we incorporate a comprehensive set of risk proxies suggested by Price et al. (2012) and Theil et al. (2019).[8] Furthermore, we include CEO age and gender to control for possible confounding effects (e.g., being introverted could have a different effect for male than for female CEOs). Definitions of all used controls are given in Table 5.

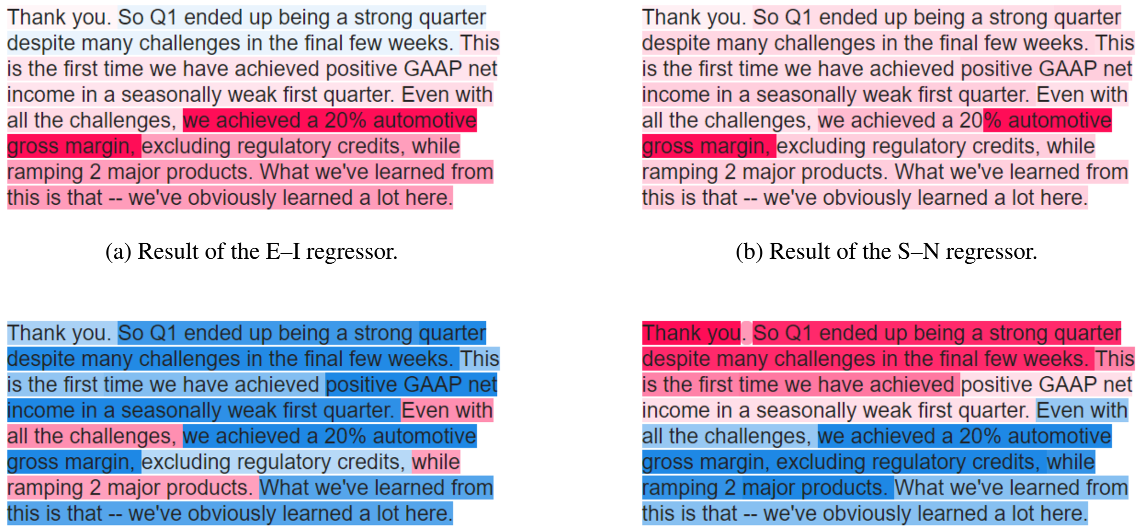
4.2 Methodology

We use the best-performing personality prediction model (RoBERTa) to infer the personality of the 1.7K unlabelled CEOs present in the 22K calls. Together with the financial covariates (see above), the predicted CEO MBTI is then used to explain short-term stock return volatility following the calls with multiple linear regression.[9] Volatility is the most common financial risk measure, and its prediction is an essential task for firm valuation and financial decision-making. Importantly, "risk" is a purely descriptive concept in finance, as it measures the fluctuation of stock returns.

[7] https://wrds-www.wharton.upenn.edu

[8] We initially also considered including a market volatility index (VIX), but decided against it as its low explanatory power and high variation inflation factor (VIF) indicated redundancy of this variable (Johnston et al., 2018).

[9] The supplementary material contains our dataset and implementation.



(a) Result of the E–I regressor.

(b) Result of the S–N regressor.

(c) Result of the T–F regressor.

(d) Result of the J–P regressor.

Figure 4: Example snippet from our dataset (uttered by Elon Musk in Tesla's Q1 2020 earnings call) with SHAP heatmap across the MBTI. Red indicates a positive and blue a negative influence on the prediction.

Feature	Definition
Age	CEO age on the call date
Gender	CEO gender
Past Vola	Standard deviation of logarithmic returns in the business quarter before the call
Size	Market value of the firm, i.e., the number of outstanding shares times stock price one day before the call
Volume	Stock trading volume on the call date
Leverage	Total liabilities divided by assets
Spread	Difference between the stock's bid and ask price on the call date
BTM	Book-to-Market = book value of the firm divided by market value
SUE	Mean absolute deviation of analysts' earnings-per-share forecasts from the actual value in the preceding quarter
ROA	Return on Assets, i.e., net income divided by assets
Industry	Fama–French 12 industry dummies
Time	Year–quarter dummies

Table 5: Controls used in the risk regression task. BTM is calculated following (Fama and French, 2001) and firms with a negative value are removed. Size, BTM, and volume are log1p-transformed.

4.3 Results and Discussion

The results of this risk regression task are shown in Table 6. We find that the first three MBTI dimensions are significantly associated with risk following the call. This significance is high ($p \leq 0.001$) for E–I and T–F. The direction of this association behaves as expected: a CEO communicating in an *introverted* and *feeling* manner is associated with increased risk ($\beta_i = 0.03$, $\beta_f = 0.10$, while an *intuitive* communication is associated with decreased risk ($\beta_s = -0.02$). Notably, these results are robust to age- and gender-fixed effects. Although seemingly small, the size of the personality effect (i.e., the coefficient height) is in line with that observed by related work (Harrison et al., 2020). It is expectable that fundamentals such as past risk or firm size have a stronger impact on future risk than, e.g., CEO extraversion. Remarkably, T–F has the third-largest impact ($\beta_f = 0.10$) out of all considered features. Though only weakly correlated with the ground truth (Table 4), the results suggest that the predictions for this scale contain strong economic signal for risk regression.

In sum, these results provide new empirical evidence to support the *upper echelons theory*. We show that situational aspects of CEO personality, predicted with our MBTI regressor, also reflect firm performance measured by stock return volatility, the most common financial risk measure.

5 Ethical Considerations

In the following, we discuss possible biases and environmental considerations.

Social Desirability Bias Past literature has shown that some Big 5 personalities are more socially desirable than others, which paves the way to discrimination: Overall, it is socially desirable to score low on *neuroticism* (an omitted scale in the MBTI) and high on *conscientiousness* and *agreeableness*. To a lesser extent, it is socially desirable

Feature	Fin	Fin + MBTI
E–I		0.03^{***}
		(5.01)
S–N		-0.02^{**}
		(-2.69)
T–F		0.10^{***}
		(13.67)
J–P		-0.00
		(-0.22)
Age		-0.01
		(0.38)
Gender		-0.02
		(-0.75)
Past Vola	0.44^{***}	0.43^{***}
	(45.80)	(44.72)
Size	-0.18^{***}	-0.19^{***}
	(-19.07)	(-19.83)
Volume	0.04^{***}	0.05^{***}
	(5.28)	(5.36)
Leverage	-0.06^{***}	-0.05^{***}
	(-8.68)	(-6.88)
Spread	0.03^{***}	0.03^{***}
	(4.30)	(4.10)
BTM	-0.04^{***}	-0.02^{***}
	(-6.22)	(-2.92)
SUE	-0.00	-0.00
	(-0.41)	(-0.73)
ROA	-0.00	0.00
	(-0.21)	(0.46)
n	21,787	21,787
Adj. R^2	33.40%	34.00%

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

Table 6: Results of the risk regression with z-standardized coefficients and t-statistics in parentheses. The sample consists of 22K earnings calls spanning 1.7K firms and years 2002–2020. Regressions include fixed effects for industry and time. Fin is a model with just the financial features (defined in §4.1) and Fin + MBTI is a joint model including the MBTI (E–I, S–N, T–F, and J–P) along with CEO age and gender.

to score high on *extraversion* and *openness* (Ones et al., 1996, Table 2). For the MBTI, in contrast, there exist no "bad" personality traits. As shown in §3.1, however, the Big 5 and the MBTI correlate. Therefore, the points raised about social desirability, albeit to a lesser extent, should apply here, too.

Sample Biases Critically, our gold standard consists of just 32 CEOs of large American (mostly tech) companies. While these companies (Alphabet, Facebook, Apple, etc.) constitute a large share of the American market, this renders the personality prediction model less applicable to non-American, small, or non-tech companies. Only four (i.e., 12.5%) of the 32 CEOs are female. While this gender ratio is twice as high as that of the S&P 500 (Catalyst, 2021), this highlights that the findings of this study might generalize poorly to non-male CEOs. In addition, as shown in §3.1, Figure 2, CEOs as a social cohort share a distinct distribution of personality traits, which is why we argue that the MBTI regressors should only be applied with caution, if at all, to non-executive samples.

Energy Consumption Training neural models can have substantial financial and environmental costs (Strubell et al., 2019), which motivates us to discuss the computational efficiency of the Transformers. Using an NVIDIA Tesla P100 GPU, we run a hyperparameter optimization over 40 configurations per MBTI dimension for both BERT and RoBERTa. The average power consumption is 200W and the optimization takes ca. 16 hours, i.e., 3.2 kilowatt hours (kWh) with an electricity cost of 40 cents per model.[10] Labeling the 22K earnings call instances with no available ground truth takes ca. 4.5 hours and 140W, i.e., 0.63 kWH of GPU time and 8 cents, respectively. Training time of the SVM with trigram tf–idf is negligible (ca. 2 minutes on a quad-core processor with 8GB RAM). Whether the performance increases of the Transformers over a sparse method justify the added computational costs should be considered carefully on a case-by-case basis.

6 Conclusion and Future Work

We present the first text regression approach for predicting the MBTI personality of CEOs. Although past research has contested the possibility of predicting MBTI from purely textual data, we observe moderate to strong correlations with the ground truth for three out of four dimensions. In a risk regression task, we demonstrate that—consistent with the *upper echelons theory*—the predicted CEO personality is significantly associated with financial risk in the form of stock return volatility. Qualitatively, extroverted, intuitive, and thinking CEOs seem to incur less financial risk.

In the future, we plan to model the personality prediction task as a multi-task learning problem, in which one single regressor is trained to predict all four MBTI dimensions at once. In addition, it would be interesting to incorporate speech signals of executives (e.g., voice modulation, tonality, and silence) into the personality predictions.

[10] Calculations assume the average U.S. electricity rate of 12.55 cents per 15 November 2021: https://www.electricchoice.com/electricity-prices-by-state

References

Mirko H. Benischke, Geoffrey P. Martin, and Lotte Glaser. 2019. CEO Equity Risk Bearing and Strategic Risk Taking: The Moderating Effect of CEO Personality. *Strategic Management Journal*, 40(1):153–177.

George E. P. Box and David R. Cox. 1964. An Analysis of Transformations. *Journal of the American Statistical Association*, 26(2):211–252.

Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41:687–699.

Isabel Briggs-Myers and Peter B. Myers. 1995. *Gifts Differing: Understanding Personality Type*. Davies-Black.

Catalyst. 2021. Women CEOs of the S&P 500.

Yuval Cohen, Hana Ornoy, and Baruch Keren. 2013. MBTI Personality Types of Project Managers and Their Success: A Field Survey. *Project Management Journal*, 44(3):78–87.

James J. Connolly, Erin J. Kavanagh, and Chockalingam Viswesvaran. 2007. The Convergent Validity Between Self and Observer Ratings of Personality: A Meta-Analytic Review. *International Journal of Selection and Assessment*, 15(1):110–117.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL*, pages 4171–4186.

Eugene F. Fama and Kenneth R. French. 2001. Disappearing Dividends: Changing Firm Characteristics or Lower Propensity to Pay? *Journal of Financial Economics*, 60(1):3–43.

Adrian Furnham. 1996. The Big Five Versus the Big Four: The Relationship Between the Myers–Briggs Type Indicator (MBTI) and NEO-PI Five Factor Model of Personality. *Personality and Individual Differences*, 21(2):303–307.

Adrian Furnham, Joanna Moutafi, and John Crump. 2003. The Relationship Between the Revised Neo-Personality Inventory and the Myers–Briggs Type Indicator. *Social Behavior and Personality*, 31(6):577–584.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A Gold Mine for Personality Prediction. In *Proceedings of the ACL Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematics and Statistical Psychology*, 61:29–48.

Donald C. Hambrick and Phyllis A. Mason. 1984. Upper Echelons: The Organization as a Reflection of Its Top Managers. *Academy of Management Review*, 9(2):193–206.

Joseph S. Harrison, Gary R. Thurgood, Steven Boivie, and Michael D. Pfarrer. 2020. Perception Is Reality: How CEOs' Observed Personality Influences Market Perceptions of Firm Risk and Shareholder Returns. *Academy of Management Journal*, 63(4):1166–1195.

Karel Hrazdil, Jiri Novak, Rafael Rogo, Christine Wiedman, and Ray Zhang. 2020. Measuring Executive Personality Using Machine-Learning Algorithms: A New Approach and Audit Fee-Based Validation Tests. *Journal of Business Finance and Accounting*, 47(3–4):519–544.

Ron Johnston, Kelvyn Jones, and David Manley. 2018. Confounding and Collinearity in Regression Analysis: A Cautionary Tale and an Alternative Procedure, Illustrated by Studies of British Voting Behaviour. *Quality & Quantity*, 52:1957–1976.

Hyunji Kim, Stefano I. Di Domenico, and Brian S. Connelly. 2019. Self–Other Agreement in Personality Reports: A Meta-Analytic Comparison of Self- and Informant-Report Means. *Psychological Science*, 30(1):129–138.

Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines. *American Psychologist*, 70(6):543–556.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition. Sage, Thousand Oaks (CA), USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint, arXiv: 1907.11692*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500.

Shavin Malhotra, Taco H. Reus, PengCheng Zhu, and Erik M. Roelofsen. 2018. The Acquisitive Nature of Extraverted CEOs. *Administrative Science Quarterly*, 63(2):370–408.

Robert R. McCrae and Paul T. Costa. 1989. Reinterpreting the Myers–Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality. *Journal of Personality*, 57(1):17–40.

Robert R. McCrae, Paul T. Costa, and Thomas A. Martin. 2010. The NEO-PI-3: A More Readable Revised NEO Personality Inventory. *Journal of Personality Assessment*, 84(3):261–270.

Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2):175–215.

Deniz S. Ones, Chockalingam Viswesvaran, and Angelika D. Reiss. 1996. Role of Social Desirability in Personality Testing for Personnel Selection: The Red Herring. *Journal of Applied Psychology*, 81(6):660–679.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. White paper, University of Texas at Austin.

Barbara Plank and Dirk Hovy. 2015. Personality Traits on Twitter—or—How to Get 1500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking and Finance*, 36(4):992–1011.

David Quarfoot and Richard A. Levine. 2016. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? *The American Statistician*, 70:373–384.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of ACL*, pages 3645–3650.

Christoph Kilian Theil, Samuel Broscheit, and Heiner Stuckenschmidt. 2019. PRoFET: Predicting the Risk of Firms from Event Transcripts. In *Proceedings of IJCAI*, pages 5211–5217.

Simine Vazire. 2006. Informant Reports: A Cheap, Fast, and Easy Method for Personality Assessment. *Journal of Research in Personality*, 40(5):472–481.

Sanja Štajner and Seren Yenikent. 2020. A Survey of Automatic Personality Detection from Texts. In *Proceedings of COLING*, pages 6284–6295.

Sanja Štajner and Seren Yenikent. 2021. Why Is MBTI Personality Detection from Texts a Difficult Task? In *Proceedings of EACL*, pages 3580–3589.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP*, pages 38–45.

A Hyperparameter Configurations

Using a Bayesian hyperparameter optimization as specified in §3.2, the following configurations led to minimal loss on the validation set. Table 7a summarizes the optimal configuration for BERT and Table 7b the one for RoBERTa.

MBTI	Batch Size	Learning Rate
E–I	128	4.8×10^{-5}
S–N	32	4.9×10^{-5}
T–F	32	1.0×10^{-6}
J–P	256	8.6×10^{-6}

(a) Hyperparameters for BERT.

MBTI	Batch Size	Learning Rate
E–I	256	4.3×10^{-5}
S–N	32	4.6×10^{-5}
T–F	128	9.4×10^{-8}
J–P	128	4.7×10^{-5}

(b) Hyperparameters for RoBERTa.

Table 7: Final hyperparameter configurations found by the Bayesian optimization searching over 40 configurations per MBTI dimension.

B Results on the Validation Set

The results of the MBTI regressors on the validation set are depicted in Table 8.

MBTI	Model	r	ρ	τ	MAE
E–I	SVM	0.70	0.69	0.55	0.38
	BERT	0.46	0.42	0.28	0.62
	RoBERTa	0.72	0.60	0.48	0.35
S–N	SVM	0.34	0.48	0.30	0.28
	BERT	0.20	0.35	0.24	0.53
	RoBERTa	0.43	0.61	0.43	0.27
T–F	SVM	0.13	−0.05	−0.03	0.33
	BERT	−0.43	−0.32	−0.22	0.38
	RoBERTa	0.11	−0.07	−0.03	0.36
J–P	SVM	−0.05	0.05	0.03	0.35
	BERT	0.32	0.28	0.19	0.53
	RoBERTa	0.25	0.14	0.06	0.40

Table 8: Results of the personality prediction task on the validation set.