INTERNET APPENDIX

Financing Sustainable Entrepreneurship: ESG Measurement, Valuation and Performance in Token Offerings

A Quantifying Startups' ESG Properties

A Machine-Learning Approach: Detailed Description

Textual analysis in economics, finance and accounting literature is mainly applied using a dictionary count approach, in which researchers rely on predefined word lists to extract information from textual data (Gentzkow et al., 2019). Should we rely on humans or machines to create these word lists?[4] The advantage of human wisdom in creating these word lists comes at the cost of subjectivity and requires substantial transparency. In the context of ESG measurement, subjectivity plays a crucial role as many of the available ESG scoring databases provide inconsistent ratings (Berg et al., 2020; Dimson et al., 2020).

In this paper, we choose a middle ground. On the one hand, the machine relies on itself in detecting the meaningful phrases in the context of startups' whitepapers (i.e., word embedding via word2vec), and on the other hand, we guide the machine to come up with the terminologies that are most relevant for our purpose, based on a set of seed words.

Our procedure to create the ESG-relevant lexicon is methodologically close to Li et al. (2020). First, we collect whitepaper documents and parse their textual contents. Second, we clean the text by performing standard prepossessing procedures and define the set of words and context-specific phrases. Third, we do word embedding using the word2vec method (Mikolov et al., 2013) to obtain vector representation of all the words and phrases that have appeared in the corpus of whitepapers. Fourth, we define a set of seed words that represent each of the three pillars of ESG. Fifth, we use our trained word2vec model and generate our word lists by finding the closest word and/or phrase to our seed words. In training our word2vec model, we accept all standard assumptions for the hyper-parameter tuning of the model. Specifically, we use the Python package provided by Li et al. (2020), which implements all the previous steps. Finally, we calculate the ESG intensity of each whitepaper using the generated word lists.

A.1 Text prepossessing

Before we feed the corpus (universe of texts) to our word2vec model, we apply standard text prepossessing procedures to ensure the efficiency of our ML training process.

First, we remove line breaks from the text and replace numbers/emails/URLs/phone numbers with the respective tags, i.e., "<num>", "<email>", "<url>", "<phone>".

[4] See (Loughran & McDonald, 2020b) for a comprehensive discussion

Second, we employ the Stanford CoreNLP pipeline (Manning et al., 2014) to generate a dependency representation of each sentence.[5] Figure I.A.1 shows the dependency representation for an example sentence: "The basis for the distribution of GNC in the domain of real economic activity is the loyalty system, this is the most important and central tool of the platform." This helps the machine to better understand the grammatical structure of the sentences, and enables it to form collocations, i.e., a collection of more than one word that tends to appear frequently together, like "initial_coin_offering". We treat these collocations as single words in the following steps.

Third, we remove the stop words, i.e., words that do not add much meaning to a sentence like 'the', 'as', 'of', etc., as well as punctuation marks. Note that this step must follow the creation of collocations, as they could consist of some stop words, such as in "as_well_as".

[Place Figure I.A.1 about here.]

A.2 Word Embedding

Word embedding is a way to mathematically represent words and enables the machine to compare the semantic similarity of the words. Our word embedding approach relies on the revolutionary word2vec method developed by Mikolov et al. (2013). The idea behind word2vec is to use a shallow (only one hidden layer) neural network, which is trained to predict words in the neighborhood of an input word, by exploring all the sentences in the corpus. In other words, during the training phase, an input word is translated to a vector in the hidden layer (a), and then this vector should predict the neighboring word (b). After the training, the trained weights of the neural network for (a) would be able to create a vector of real numbers for any input word of the corpus.[6] If trained on a vast corpus, the results of this seemingly simple algorithm would be very precise. A famous example of a trained word2vec model would be that one could find the vector closest to the vector of word 'Queen' by subtracting the vector of 'man' from the vector of the word 'King' and add the results to the vector of the word 'woman' (i.e., $King - Man + Woman = Queen$).[7]

[5]The CoreNLP pipeline incorporates several steps. The most important steps include 1) tokenization, i.e., breaking down the text to smaller language units like words, 2) lemmatization, i.e., converting a word to its base form (e.g., "coins" to "coin"), and 3) entity chunking, i.e., replacing the entities' names with a proper tag.

[6]The size of this vector is the same as the size of the hidden layer in the neural network. We use the same settings as in Li et al. (2020) and consider a vector of size 300 for the word representations.

[7]Like any other ML framework, word2vec has its limitations. See Nissim et al. (2020) for a discussion on interesting and humorous examples of word2vec predictions.

A.3 Seed Words

As the starting point for measuring ESG intensity of the startup whitepapers, we collect all the available Financial Times (FT) articles with the tag of "ESG Investing" or "Moral Money". We follow a standard bag-of-words approach and extract bi-grams and tri-grams[8] that appear most frequently in the FT corpus. We then manually go through these n-grams and decide if they belong to the E, S or G dimensions of the ESG. As the FT mostly covers the corporate world, it may not necessarily include the governance terms that are important for our context of ICO whitepapers. Therefore, we manually add terms like 'kyc', 'whitelist', 'blockchain', 'utility', 'security_token', etc. for the governance dimension. The full list of our seed words (available in Table I.A.1) consists of 70 Environmental, 38 Social, and 46 Governance related words/n-grams.

A.4 ESG wordlists

For any term t of the seed words in any of the ESG dimensions j, we obtain a vector representation with the size of 300 (the size of the hidden layer in our word2vec model) as $V^t_{j\in\{E,S,G\}} = [x^t_1, x^t_2, ..., x^t_{300}]$. We then calculate the average vector for each of the ESG dimensions as $\bar{V}^{j\in\{E,S,G\}} = \frac{1}{N}\sum_1^N [x^t_1, x^t_2, ..., x^t_{300}]$ where N is the size of seed words for the dimension j. This leaves us with three vectors of $\bar{V}^E$, $\bar{V}^S$, and $\bar{V}^G$.

Next, we perform a cosine similarity between $\bar{V}^j$ and the vector of all of the terms in our whitepaper corpus and select the 500 most similar terms for each dimension. If a term appears in more than one dimension, then it is only considered for the dimension that has a higher cosine similarity.[9] Furthermore, some of our seed words have never appeared in the corpus of whitepapers.[10] We did not remove them from our word lists, though not affecting our results at all, as these terms could be relevant for future out-of-sample studies. This leaves us with a total of 1,495 ESG-related terms consisting of 508, 463 and 524 terms in the respective ESG dimensions.

A.5 ESG Score

We quantify the E, S and G dimensions using a dictionary-based approach, by counting the number of distinct occurrences of our respective word list in the ICOs whitepapers, normalized to the size of the word list. Specifically, for ICO i we measure each dimension

[8]Please note that bi-grams and tri-grams are two and three-word combinations of the words that appear in a neighborhood, and are not necessarily a collocation.

[9]This is the reason why some dimensions could have a word list smaller than 500.

[10]This is the reason why some dimensions could have a word list greater than 500.

of the ESG as:

$$E[S\ or\ G]_i = \frac{\sum_t 1_{c(t)_i>0}}{c(n)}, \tag{8}$$

Where $c(t)_i$ is the count of term t in the whitepaper of ICO i and $c(n)$ is the size of the corresponding *word list*.

According to Loughran and McDonald (2020a), this approach slightly deviates from the norm in accounting and finance literature, where researchers count the total frequency of the words in a word list and normalize it to the total words in the document. In our context, however, this will lead to biases. Unlike corporate disclosures, ICO whitepapers are neither standardized nor regulated, and they vary substantially in length, format and content. Moreover, some ICOs have the words like 'green' or 'human' in their titles, which leads to bias in measuring the environmental or social score if a traditional frequency count method is applied.

Furthermore, we measure the total ESG score of the startup i by adding the three dimensions' intensity, i.e. $ESG_i = E_i + S_i + G_i$.

B Additional Controls

In this section, we check the robustness of our findings by including additional control variables to our baseline model. We control for the following additional controls: # investors, KYC, ICO duration, fiat accepted, % distributed in ICO, Twitter followers, LinkedIn, and crypto experience.

 # Investors. The logarithm of the number of institutional investors, as listed on the *CryptoFundResearch* list.

KYC. A dummy variable that equals one if the firm has a Know-Your-Customer (KYC) procedure, and zero otherwise.

ICO duration. The difference in days between the start and end of the ICO.

Fiat accepted. A dummy variable that equals one if the ICO accept fiat currencies.

Distributed in ICO. The percentage of tokens distributed in the token offering (i.e., 1 - "Distributed in ICO" is the token retention ratio).

Twitter followers. The logarithm of the number of the firm's Twitter followers.

Linkedin. A dummy variable that equals one if the ICO has a Linkedin page.

Crypto experience. The percentage of the team members who have experience in the crypto environments.

Table I.A.2 reports the results of this analysis. Adding the additional controls reduces our observations from 1043 in column (1) to 808 in column (5), which has the highest number of control variables. Our main results do not qualitatively change in these specifications. In all specifications, the coefficient on the normalized ESG score remains statistically significant at least at 5%.

[Place Table I.A.2 about here.]

C Other Seed Words

In this section, we address potential concerns that our results could be driven by our manual selection of the seed words. To this end, we repeat the steps in generating our word lists with the exception that we consider only two or three seed words for each dimension of the ESG. Specifically, we set the seed words to be ['environmental', 'climate'] for the E dimension, ['society', 'social_responsibility'] for the S dimension, and ['governance', 'white_paper', 'token'] for the G dimension. Figure I.A.2 illustrates the resulting word lists, and it shows that we are able to capture the most relevant terms needed to construct our ESG word lists with only two or three words.

[Place Figure I.A.2 about here.]

C.1 Other Seed Words and Funding

To test the validity of the word lists created with the small set of seed words, we repeat our baseline (OLS) regression with the log of the funding amount in $ million as the dependent variable, on the ESG score as well as its components derived from these word lists.

Table I.A.3 shows the results of this analysis. In column (1), the coefficient of the normalized ESG score is 0.34, with a p-value $<$ 1%, suggesting that a one standard deviation increase in the ESG score increases the average funding amount of $15.2 million by $6.1 million, or 40%. Columns (2), (3) and (4) report regression coefficients for the disaggregated and normalized E, S and G scores, respectively. All disaggregated scores are statistically significant at the 1% level in these models. The E score coefficient is 0.138 (p-value $<$ 0.01), the S score coefficient is 0.212 (p-value $<$ 0.01), and the G score coefficient is 0.321 (p-value $<$ 0.01). However, testing the effect of the three disaggregated scores simultaneously in column (5) shows that only the E (0.123) and the G (0.301) score are statistically significant at least at the 5% level. Thus, ceteris paribus increases by one standard deviation in E and G are associated with 13% and 35% increases in the average funding amount, respectively. These results are in line with the paper's analysis and strongly support the VPH that there is a sustainability-related valuation premium in token offerings.

[Place Table I.A.3 about here.]

References

Berg, F., Koelbel, J. F., & Rigobon, R. (2020). *Aggregate confusion: The divergence of esg ratings*. MIT Sloan School of Management.

Dimson, E., Marsh, P., & Staunton, M. (2020). Divergent esg ratings. *The Journal of Portfolio Management*, *47*(1), 75–87.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

Li, K., Mai, F., Shen, R., & Yan, X. (2020). Measuring corporate culture using machine learning. *The Review of Financial Studies*.

Loughran, T., & McDonald, B. (2020a). Measuring firm complexity. *Available at SSRN 3645372*.

Loughran, T., & McDonald, B. (2020b). Textual analysis in finance. *Annual Review of Financial Economics*, *12*(1), 357–375. https://doi.org/10.1146/annurev-financial-012820-032249

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(2), 487–497.

Internet Appendix — Exhibits

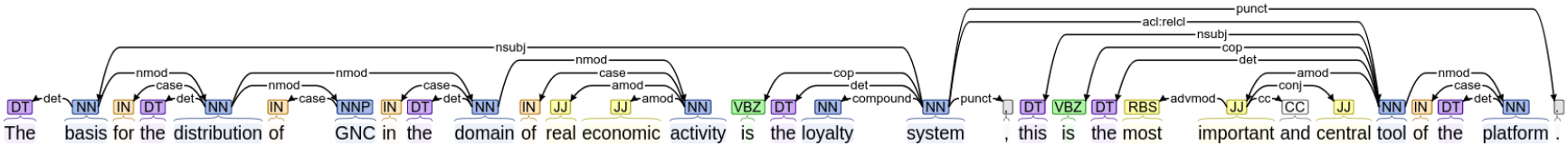

Figure I.A.1: Example of a dependency representation

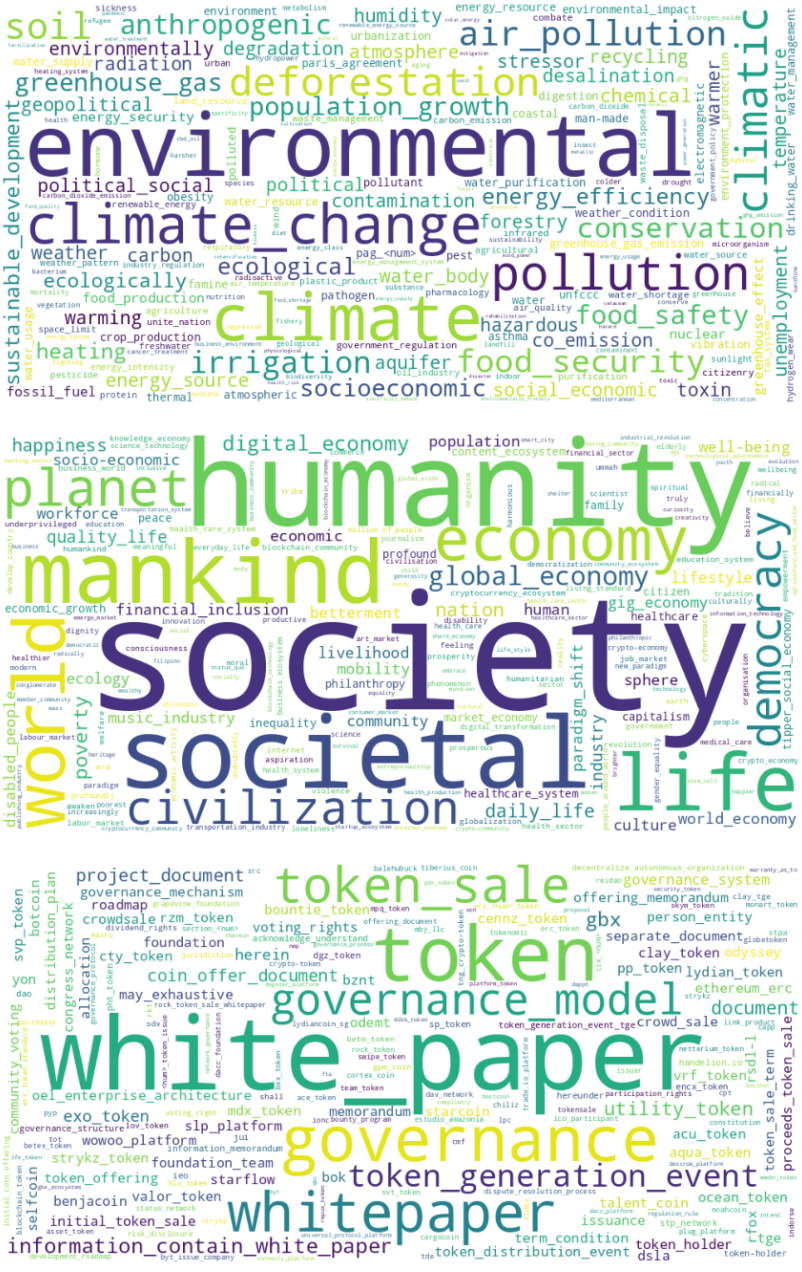

Figure I.A.2: Robustness Tests: ESG word lists with alternative seed words

Table I.A.1:
Seed Words

E	S	G
climate_change	moral_money	pension_funds
green_bonds	responsible_investing	investment_management
fossil_fuel	development_goals	supply_chain
green_bond	sustainable_development_goals	task_force
carbon_emission	impact_investment	investment_managers
carbon_footprint	social_issues	chief_investment_officer
renewable_energy	uns_sustainable	governance_issues
global_warming	social_impact	private_sector
greenhouse_gas	positive_impact	hedge_funds
climate_risk	essential_forward_thinking	managing_director
energy_source	gender_diversity	shareholder_proposals
green_finance	developing_countries	due_diligence
greenhouse_gas_emissions	decentralized	stakeholder_capitalism
carbon_footprint	defi	retail_investors
paris_climate	democratize	annual_meetings
climate_change_meets	democratization	esg_disclosure
paris_agreement	disintermediation	law_firm
fuel_companies	africa	global_advisors
fossil_fuel_companies	poor	board_members
climate_crisis	catching_up	investors_looking
natural_gas	india	passive_managers
environmental_impact	mobile	institutional_investors
thermal_coal	mobility	advisors
force_climaterelated_disclosures	cell_phone	bounty
green_bond_market	smart_phone	kyc
climaterelated_risks	access	whitelist

green_energy	geography	blockchain
low_carbon	dispersion	utility
oil_gas_companies	microfinance	security_token
environmental_issues	micro_finance	token_distribution
carbon_dioxide	impact_investing	intermediary
zero_emissions	equality	law
indispensable_energy	inequality	regulation
bn_green	care	policy
carbon_pricing	income	regulator
green_deal	responsible_investment	token_retention
carbon_neutral	impact_investing	airdrop
fight_climate_change	csr	founder
carbon_price		partner
coal_power		compliance
green_bonds		howey_test
fossil_fuel		sec
tackle_climate		equity
lowcarbon_economy		venture_capital
co_emissions		VC
risks_climate		incubator
zero_carbon		
green_investment		
risks_climate_change		
green_credentials		
reduce_carbon		
action_climate		
save_planet		
green_debt		
greenhouse_gases		
coal_projects		

away_fossil		
climate_accord		
carbon_credits		
first_green		
environmental_standards		
un_climate		
new_green		
netzero_carbon		
solar_wind		
renewable_energy		
global_warming		
sustainable_investing		
sustainable_investment		
sustainable_development		

Table I.A.2: Robustness Tests: Additional Controls

Column	(1)	(2)	(3)	(4)	(5)
Dependent variable: Valuation, Funding amount (log.)					
ESG	0.276***	0.234***	0.223**	0.198**	0.197**
	(0.085)	(0.088)	(0.097)	(0.099)	(0.100)
GR	-0.356	-0.151	0.192	0.121	0.161
	(0.755)	(0.761)	(0.849)	(0.872)	(0.872)
$Investors$		0.690***	0.605***	0.554***	0.547***
		(0.093)	(0.113)	(0.114)	(0.112)
KYC		0.186	0.257	0.256	0.261
		(0.165)	(0.180)	(0.184)	(0.183)
$ICO\ Duration$		0.002	0.002	0.002	0.002
		(0.001)	(0.002)	(0.002)	(0.002)
$Fiat\ Accepted$			0.427	0.509	0.505
			(0.304)	(0.318)	(0.317)
$Distributed\ in\ ICO$			-0.676*	-0.608	-0.633
			(0.406)	(0.412)	(0.411)
$TwitterFollowers$				0.103**	0.096**
				(0.042)	(0.042)
$Linkedin$					0.058
					(0.180)
$CryptoExperience$					0.507*
					(0.302)
Observations	1043	1039	835	808	808
R^2	0.314	0.345	0.368	0.369	0.372
Controls	✓	✓	✓	✓	✓
Quarter-year FEs	✓	✓	✓	✓	✓
Country FEs	✓	✓	✓	✓	✓

Table I.A.3: Robustness - Seed words

Column	(1)	(2)	(3)	(4)	(5)
Dependent variable: Valuation, Funding amount (log.)					
ESG	0.338***				
	(0.070)				
Environmental		0.138***			0.123**
		(0.052)			(0.059)
Social			0.212***		0.037
			(0.074)		(0.091)
Governance				0.321***	0.301***
				(0.077)	(0.088)
Observations	1043	1043	1043	1043	1043
R^2	0.318	0.310	0.311	0.318	0.322
Controls	✓	✓	✓	✓	✓
Quarter-year FEs	✓	✓	✓	✓	✓
Country FEs	✓	✓	✓	✓	✓

Table I.A.4: Whitepaper-related selectivity: Second stage from 2SLS

Column	(1)	(2)
Dependent variable: Valuation, Funding amount (log.)		
ESG	0.247***	0.247***
	(0.067)	(0.067)
IMR	✓	✗
GR	✗	✓
Observations	1043	1043
R^2	0.309	0.309
Controls	✓	✓
Quarter_FE	✓	✓
Country_FE	✓	✓

THE END OF THE INTERNET APPENDIX