Generating Word Lists for Analyzing and Monitoring Social Good: The Case of Sustainability

Daniel E. O'Leary[1] and Yangin Ben Yoon[2]

[1]University of Southern California
[2]Seoul National University of Science and Technology
oleary@usc.edu ben.yangin.yoon@seoultech.ac.kr

Abstract

This paper examines issues associated with the development of bags of words that can be used to analyze the extent to which descriptors of a concept are related to some dependent variable and as an approach to continuously monitor the occurrence of those concepts in text. We focus on generating bags of words using Word2Vec, using two key sources of business text, Form 10-Ks and "earnings calls," to support issues of concern in social good. As an experiment, we drill down on building bags of words, that describe independent variables, descriptive of the concept of "sustainability," which could be related issues such as firm value measures (profitability), events (release of new products or mergers) or other dependent variables.

Introduction

One approach to analyzing and continuously monitoring information in the world for social good is using a bag of words approach. By analyzing text streams, we can gather the information that includes words in our bag of words of interest. As one example, O'Leary and Spangler (2016) examine a system used to find text information related to chocolate and chocolate companies for purposes of finding issues relating chocolate to health, such as choking incidents or concerns about diabetes or allergies. This approach has applications in many domains, including generating taxonomies and ontologies for information search about employment, individual histories or social media.

Another approach is to use bags of words to understand what variables in some text are related to a dependent variable. In that approach, statistical models are developed trying to map relationships between occurrences of particular types of text and some outcomes. In those settings, there is interest in relating a dependent variable (financial return, stock price, fraud, merger, etc.) to some text. As an example, Allen et al. (2021) built a word list to study the potential impact of information captured in text on effective tax rates. In order to structure and analyze some text, we generate a bag of words designed to measure the extent some concept (say about social good) is found in that text. We count occurrences of each of the words and perform a statistical analysis of some dependent variable, based on the number of occurrences of those words in the bag of words, as independent variables. This provides a statistical model of the dependent variable that can be used to better understand the relationship with the information in some text.

Still another approach is to use generative AI to help analyze information. Unfortunately, generative AI systems are periodically developed, leaving them periodically informationally out of date. One approach to ensuring that they have updated information is to provide them with additional information using so-called retrieval augmented generation (Lewis et al. 2020). However, in some cases chosing the information to include is not easy and also must be up-to-date. A bag of words could provide a targeted list of concepts, taxonomies or ontologies to facilitate search.

In these three and other settings, we use a bag of words, to find words that would signal or identify a concept, both individually and as a portfolio or group using contemporary information. This bag of words approach is well-known and used in psychology-based systems, such as LIWC – Language Inquiry and Word Count (Boyd et al. 2022) – where different psychology concepts, such as "power," are characterized by a set of words. However, we are interested in analyzing text for issues different than psychological concepts and concerned more with issues such as sustainability or environmental concerns or other issues of social good.

In order to determine a relevant bag of words, we identify a concept of interest, as captured in a "seed word," and we use Word2Vec, to analyze multiple corpora, generating words that are "similar" to the seed word. This approach provides an explicit specification of that seed word, characterized by a list of words that are "similar" to it from a specific corpus or set of corpora. The resulting list is referred to as a bag of words.

This Paper

Thus, the purpose of this paper is to examine an approach for generating a bag of words for use in the analysis of issues related to sustainability. This paper approaches that purpose in the following manner. The following section reviews our approach that is based on Word2Vec, while the subsequent section summarizes our findings. We then describe an alternative approach. In the final section, we summarize the paper and its contributions.

Background

This section provides a brief overview of some key concepts used in this paper.

Word2Vec

Word2Vec is an approach that allows representation of words as vectors. The approach allows development of relationships between words, measured in a vector space. This research uses Word2Vec, starting with the conceptual seed word of interest, to generate a set of "similar" words, based on a particular set of text corpora. The seed word(s) is chosen to capture a particular concept of interest, as related to "social good."

In the original uses of Word2Vec, Mikolov et al. (2013a), used a Google news corpus to find several types of semantic and syntactic "similarities" in their analysis. For example, they found country currencies (kwanza in Angola), cities in states (Chicago in Illinois), man-woman relationships (brother-sister), opposites (ethical and unethical) and several other relationships. This research extends that analysis from the Google News corpora to two different sets of text from a business environment and to the environment of social good, and in so doing finds several other similarities.

Ontology

Historically, an ontology was elegantly defined as "an explicit specification of a conceptualization" (Gruber 1993, p. 199). That definition later was extended to stress the importance of a "shared conceptualization," e.g., Guardino et al. (2009). In addition, as part of a discussion, Guardino et al. (2009) noted, "an ontology is a special kind of information object or computational artifact." We will be developing word lists that describe a concept, based on the particular seed word chosen, providing us with an artifact that has the ability to model the semantic meaning or structure of the concept of concern. Our word lists provide a set of signals that can be used to identify occurrences of a shared conceptualization of a seed word in a corpus.

Human in the Loop

To-date, artificial intelligence is well-known to provide incomplete, ambiguous or even incorrect information. As a result, it typically is necessary to have a human in the loop (HITL) to choose between potential answers, confirm answers and guide the development of a solution that a system can use, and that is sensible to other users. There has been substantial research on HITL (e.g. Holzinger 2016 and others) and experts (O'Leary 1993a), and we use a HITL as an important part of our approach. For example, the HITL provides expertise that allows us to ensure that the words are consistent with seed word and our intent to study sustainability.

Equivocality and Unigrams vs Bigrams

As noted by previous researchers, it takes uncertainty to destroy uncertainty. For example, Ashby's Law of requisite variety (1956) states, "it takes variety to destroy variety." Also, as noted by Karl Weick (1969), "it takes equivocality to remove equivocality." The implication is that when we build word lists, even if the seed word is a unigram, it may take an n-gram to effectively capture the ontological meaning of the seed word of concern. Our implementation of Word2Vec allows us to generate both unigrams and bigrams in our bags of words. As a result, we also take the bag of words approach beyond the typical unigram approach that is generally used (e.g., Boyd et al. 2022).

Social Good

In this paper we are concerned with using this approach in issues associated with "social good." However, as discussed in O'Leary (2025) it is not clear that there is a universal understanding of what is "AI for Good" or when something is for social good. For example, Arrow's (1970) well-known theorem indicates that for some decisions there is no solution that is preferred by all. Accordingly, there is not necessarily one choice that is good for all. This theorem has also been used to note that there is not likely to be a single ontology or list of words that is optimal for all applications (O'Leary 2003).

Although there is ambiguity, as to what is "good," there has been substantial research pursuing AI for Social Good. For the issue addressed in this paper, a Google search for "is sustainability an important issue" generated 1.33 billion Google results, suggesting its importance.

Implementation

Mikolov et al. (2013a, 2013b) use two different algorithmic approaches as part of their work on Word2Vec, CBOW (continuous bag of words) and Skip-gram. Those algorithms were used on different datasets, with both unigrams

and bigrams as the basis of the search for one of the datasets. We used multiple sets of text data: 10-K for 2020 and 2021, separately and combined and two sets of "earnings call" data.

Form 10-Ks contain text and numeric information about firms that are required by the United States government's Securities and Exchange Commission. Our dataset included more than 8000 different Form 10-Ks. Earnings calls are text data that derives from management's periodic discussions in public forums about different issues related to corporate earnings. Finally, for the combined set of 2020 and 2021 form 10-Ks we used two different approaches, one using unigrams and the other using bigrams. Taken together, this gave us twelve word sets, for which we gathered 30 words in each set. Those words were then edited for reasonableness and duplications.

Approach

Our research objective was to use an implementation of Word2Vec to generate a list of words based on some key "social good" concepts of concern, with a particular focus on business use and effects of social good on those firms. Because we are interested in tracking business related sustainability concerns, our analysis used text from corporate disclosure information that is publicly available, Form 10-K submissions to the Security and Exchange Commission, and Earnings Calls, where management discusses company earnings in an open forum.

In our research, we examine the seed word, "sustainability," using different financial corpora, we find additional types of relationships beyond the work of Mikolov et al. (2013a).

Equivocality and Unigrams vs Bigrams

We used a seed word in Word2Vec to generate words "similar" to that seed word. Unfortunately, not all words generated using Word2Vec are of general interest in monitoring and analyzing issues concerned with the seed word. We used multiple "filters" to choose the words for our resulting dictionary. First, we used a human in the loop to review the words and exclude words that they considered not similar to the concept. Typically, this meant that the word generated by Word2Vec were too general and not specifically related to our seed word, e.g., "improving" and "excellence." In addition, Word2Vec found some "inclusiveness" terms, as descriptive of sustainability. Our HITL choose not to include those terms. Second, for each word that the HITL thought was appropriate, we used Google's Search AI to ascertain "if x is related to sustainability." We included the word on our list if Google's AI indicated "yes." Third, the relative number of appearances of a word within the different approaches provides another way of determining if a word in question could be used on our list. For example, no words that appeared only a single time, among the twelve data sets, were included on the list.

Findings – Seed Word "Sustainability"

Using Word2Vec generated unigrams, bigrams and abbreviations. In this section we summarize our findings.

Word List

Using the approach outlined in this paper, Tables 1a and 1b summarize our sixteen unigrams and our seven bigrams generated by Word2Vec. The word list include word, the type (unigram or bigram) and the number of occurrences in our word lists.

Based on this dictionary, a holistic overview using this set of words provides a view that "sustainability" relates to the "environment," with a role of "stewardship" and "governance," a responsibility of "accountability," with goals of "survivability" and "safety."

The bigrams we chose includes one of the words, "environmental," "corporate," and "sustainability," and one of the words "goals," "initiatives," "social," "responsibility," "sustainability," and "stewardship." As part of "fleshing out" this word set we could add those other combinations not listed, such as "environmental goals," but we already capture the unigram sustainability.

Abbreviations Found

Interestingly, our use of Word2Vec found several abbreviations "similar" to sustainability, as seen in table 2. The most frequently appearing abbreviations were related to acronyms that included "environment," or "sustainable." In addition, they capture issues such as regulation, with the term "Global Reporting Initiative."

Number	Bigrams
2	Environmental Stewardship
2	Environmental Responsibility
2	Corporate Responsibility
2	Environmental Sustainability
2	Environmental Social
2	Sustainability Goals
2	Sustainability Initiatives

Table 1a: Bigrams in Word List.

Number	Unigrams
10	Stewardship
8	Resilience
7	Governance
6	Safety
5	Resiliency
5	Climate
5	Decarbonization
5	Sustainable
4	Accountability
2	Survivability
2	Cybersecurity
2	Survivability
2	Readiness
2	Vigilance
2	Environmental Sustainability

Table 1b: Unigrams in Word List.

ESG – "Environment, Social and Governance" was found by each of the 12 different corpora/approach combinations and was either the first or second rated word in 10 of the 12. After ESG, GRI (Global Reporting Initiative), EHS (Environment, Health and Safety) and HSE (Health, Safety, Environment) were the highest rank. Three of those four include "Environment," and two of those four include "Safety," illustrating the concepts that are appearing parallel to sustainability. In addition, using sustainability as a seed word also generated four different abbreviations related to diversity, suggesting some similarities between sustainability and diversity. It is interesting to speculate that the similarities may be driven by being areas of change or areas of corporate disclosure requirements.

In any case, the choice of which abbreviations to include would likely depend on the type of sustainability of concern. If we were concerned about "environment," it is likely that we would include the top seven (most frequently occurring) in our table 2.

Types of Semantic and Syntactic Relationships

Mikolov et al. (2013a) noted that Word2Vec generated several types of similarity, using their Google text database. As we examine relationships found using Word2Vec on our data we find several different types of semantic and syntactic relationships, extending those found in Google's word set, in our search using the seed word "sustainability," as summarized in Table 3.

No.	Abbreviation	Extended Meaning of Abbreviation
14	ESG	Environment, Social, Governance
6	GRI	Global Reporting Initiative
5	EHS	Environment, Health, Safety
5	HSE	Health, Safety, Environment
4	SDGS	Sustainable Development Goals
3	QHSE	Quality, Health, Safety, Environment
3	TCFD	Task Force Climate-related Financial Disclosures
3	DEI	Diversity, Equity and Inclusion
2	LGBTQ	Lesbian, Gay, Bi, Trans and Queer
1	DIB	Diversity, Inclusion, Belonging
1	IWD	International Women's Day
1	HDARS	Healthcare Data and Relationship Set

Table 2: Abbreviations.

Types of Semantic and Syntactic Relationships

Mikolov et al. (2013a) noted that Word2Vec generated several types of similarity, using their Google text database. As we examine relationships found using Word2Vec on our data we find several different types of semantic and syntactic relationships, extending those found in Google's word set, in our search using the seed word "sustainability," as summarized in Table 3.

Relationship	Example
Types of Sustainability	Social Responsibility
Acronyms	SDGS – Sustainable Development Goal
Word Variations	Sustainable
Things Created to Implement	Sustainable Initiatives
Bigram "and" Relationship	And governance
Bigram Type of Safety	Workplace Safety
Bigram Linking Word and Acronym	Governance ESG
Descriptors to Environmental	Environmental Stewardship

Table 3: Relationships.

Each of these types of relationships is different than those in Mikolov et al. (2013a), suggesting the importance of the particular corpora (data) driving the types of relationships that are found with Word2Vec.

Alternative Approach

Because Word2Vec uses a specific corpus or corpora, an analysis of these results suggest that it is possible to omit some related concepts that can be found in other settings or corpora. For example, in our analysis of our text of the seed word "sustainability" we did not find any relationships to renewable energy, green or green energy or intergenerational equity. If we were to use our resulting dictionary in other settings, that might lead to limitations. However, if we were using our dictionary for analysis of these corpora that we did use, then it is unlikely that those omissions would affect analysis, because Word2Vec did not find the concepts adjacent to our seed word.

Google Search List

An alternative approach to generating word lists is to use one of the available generative AI tools, including ChatGPT, Claude or others. We used "Google Search," we asked, "can you give me a list of words that indicate that something is sustainable?" That list is summarized in Table 4. As seen in the list, there are four bigrams and four unigrams.

It is interesting that other than the terms "sustainable" and "sustainability" there are no overlaps with the word lists generated from Word2Vec and our corpora in Tables 1-3. Accordingly, such a word list, generated outside of the business 10-K disclosures and earnings calls would likely not be helpful at finding information related to sustainability in those documents. As a result, this finding suggests building dictionaries for the corpora of interest.

A large language model (LLM), such as ChatGPT, also can be used to further validate the words generated by this approach. In addition, a LLM can be asked if a word is "similar" to the a seed word to confirm the results, from a different perspective.

Bigrams	**Unigrams**
Carbon Neutral	Sustainable
Carbon Negative	Sustainability
Circular Economy	Eco-friendly
Environmentally Conscious	Green

Table 4: Google Search Response List.

A Big Beautiful List

These findings also suggest that perhaps there are multiple uses of what the word sustainability even means (e.g., Leamon 2018). Accordingly, if we wish to build an all-encompassing dictionary, with multiple uses and purposes, then we would need to include multiple sources beyond those here. Unfortunately, that also could lead to mis-identification of sustainability text, with both type I and II errors. As a result, perhaps an approach such as using Word2Vec is likely to provide the greatest domain specific and corpora-specific accuracy.

Summary, Contributions and Extensions

This paper illustrates the development of a bags of word for a concept designed for social good, with a particular focus on sustainability. The paper noted two rationales for finding word lists include using the counts of the words as independent variables used to estimate some dependent variable (Allen et al. 2021) and as a basis for monitoring a stream of text, say for a real time organization (O'Leary and Spangler 2016, O'Leary 2008).

The approach used a seed word concept, "sustainability" in Word2Vec, to generate a set of similar words from Form 10-Ks and earnings calls. Our implementation employed both unigrams and bigrams. Although we focused on Word2Vec, we also considered one of the many emerging generative AI approaches currently available and found limited overlap with our approach. Our findings suggest the importance of using an appropriate corpus for development of the word list.

Appendix: Word2Vec Implementation

This paper employs the Word2Vec method to investigate the meanings and relationships between words in a business context. In order to reflect the usage of words in this context, we use Form 10-K filings and earnings call transcripts. Form 10-K is the official and statutory filing that public companies are required to disclose annually. It provides both financial and non-financial information, including management discussions and risk factors. Earnings calls, on the other hand, are voluntary communications through which companies share their past performance and future plans with investors and stock market analysts. We believe these documents are among the most widely used and carefully prepared by companies. Form 10-K filings are retrieved from the SEC's EDGAR website, while earnings call transcripts are collected from the Seeking Alpha website.

Based on these corpora, each word was converted into a 100-dimensional vector using the Word2Vec method. This paper employed the Gensim Python package to implement Word2Vec, as proposed by Mikolov et al. (2013a).

Word2Vec assumes that the meaning of a word can be identified from its neighboring words and co-occurrence patterns in sentences. To implement this assumption, Mikolov et al. (2013a) proposed models that either (1) predict a center word given its surrounding words (CBOW), or (2) predict surrounding words by using a given word (Skip-gram). This paper adopts the same modeling approaches. We parsed the corpora into sentences and trained the models accordingly. From the trained models, we obtained vector representations for each word. Once word vectors are obtained from the models, we can get a list of words based on a seed word by selecting those located near the seed word in the vector space.

References

Allen, E., O'Leary, D.E., Qu, H. and Swenson, C.W., 2021. Tax specific versus generic accounting-based textual analysis and the relationship with effective tax rates: Building context. *Journal of Information Systems*, *35*(2), pp.115-147.

Arrow, K. 1970. *Social choice and individual values*. Yale Univ. Press.

Ashby, W. R. 1956. *An Introduction to Cybernetics*. London, U.K.: Chapman & Hall

Boyd, R.L., Ashokkumar, A., Seraj, S. and Pennebaker, J.W., 2022. *The development and psychometric properties of LIWC-22.* Austin, TX: University of Texas at Austin, pp.1-47.

Guarino, N., Oberle, D. and Staab, S., 2009. What is an ontology? *Handbook on ontologies*, pp.1-17.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5 (2): 199–220. https://doi.org/10.1006/knac.1993.1008

Holzinger, A., Interactive Machine Learning for Health Informatics: When do we need Human-in-the-loop? *Brain Informatics,* 3(2), pp.119-131 (2016).

King, D. and O'Leary, D., 1996. Intelligent executive information systems. *IEEE Expert*, *11*(6), pp.30-35.

Leamon, J. 2018, Ask an Environmentalist: What do you mean when you say 'sustainable'? Input, Fort Wayne, https://www.inputfortwayne.com/features/what-does-sustainable-mean.aspx

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, pp.9459-9474.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. arXiv Working paper 1301.3781, Ithaca, NY: Cornell University Library.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119. Red Hook, NY: Curran Associates.

O'Leary, D.E., 1993a. Determining differences in expert judgment: Implications for knowledge acquisition and validation. *Decision Sciences*, *24*(2), pp.395-408.

O'Leary, D.E., 1993b. Verification and validation of case-based systems. *Expert systems with applications*, *6*(1), pp.57-66.

O'Leary, D.E., 2003. Different firms, different ontologies, and no one best ontology. *IEEE Intelligent Systems and Their Applications*, *15*(5), pp.72-78.

O'Leary, D., 2008. Supporting decisions in real-time enterprises: autonomic supply chain systems. *Information Systems & e-Business Management*, *6*(3).

O'Leary, D.E., 2016. Is knowledge management dead (or dying)?. *Journal of Decision Systems*, *25*(sup1), pp.512-526.

O'Leary, D.E., 2023. Digitization, digitalization, and digital transformation in accounting, electronic commerce, and supply chains. *Intelligent Systems in Accounting, Finance and Management*, *30*(2), pp.101-110.

O'Leary, D., 2025. AI for Good: History, Open Data and Some ESG-based Applications. *Journal of Decision Systems*, *34*(1), p.2443182.

O'Leary, D.E. and Spangler, S., 2016, December. Monitoring and mining digital media for brand and reputation information. In *International Conference on Information Systems*. Association for Information Systems.

Storey, V.C. and O'Leary, D.E., 2024. Text analysis of evolving emotions and sentiments in COVID-19 Twitter communication. *Cognitive Computation*, *16*(4), pp.1834-1857.

Studer, R., Benjamins, R. and Fensel, D. 1998. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1–2):161–198.

Weick, K. E. 1969. *The Social Psychology of Organizing*. Reading, MA: Addison-Wesley.