1 FINANCE 751 Technical Note

This technical note informs software installation, transcript extraction, implementation and comparison methodologies to ascertain measures of corporate culture in NZX50 companies, and six Australian commercial banks, using 258 earnings call transcripts. This note describes the steps taken to implement Option-2, replicating the corporate culture results. Additionally, this technical note is for a MacOS operating system and assumes basic proficiency in package management and Python programming.

1.1 Installation

This section informs the installation of required software to facilitate analysis.

1. Install several software packages to run the StanfordCoreNLP to measure corporate culture from text files and develop transcript processing code. Anaconda is a distribution of the Python programming language, simplifying package management to develop code in the Python language. Microsoft Visual Studio Code is an integrated development environment, suitable for application building. The combination of both Anaconda and Microsoft Visual Studio Code enable programming environments to process the transcripts.

2. Secondly, clone the remote repository implementing the method described by Li et al., (2021) to your local directory. Read the instructions carefully for correct installation. In particular, change the **os.environ ["CORENLP_HOME"]** variable in **global_options.py** file to the installation location of the **stanford-corenlp-full-2018-10-05** directory on your local device. Install the required python packages excluded from Anaconda using the pip package and requirements.txt with the **pip install requirement.txt** command in the terminal. Some packages require specific versions. If you need to revert to a previous version, use the terminal command **pip install PackageName==Version** to revert to a previous version.

3. Test the correct installation of the StanfordCoreNLP using the document text files from remote repository by following the ReadMe instructions. Progress to transcript extraction after the successful execution of the StanfordCoreNLP. Otherwise, review the above installation process before progressing.

1.2 Transcript Extraction

This section informs extracting earnings call transcripts from Capital IQ.

1. Review the firm_id column in the 1.firm_score.xlsx sheet from Option-1 to identify the companies related to the 258 earnings call transcripts.

2. Navigate to Capital IQ, selecting the companies tab, followed by the transcripts link.

3. In the search criteria company search bar, type in and select each of the unique companies listed in the firm_id column mentioned above. The selected entities will update to list sixteen companies as Telecom Corp of New Zealand Ltd changed rebranded to Spark new Zealand Limited.

4. Change the time frame from 01/01/2009 to 01/10/2021 to ensure you include all 258 transcripts listed in the 1.firm_score.xlsx spreadsheet and select search in middle-right of the webpage.

5. Select all transcripts on the page by ticking the top tick box middle-left of the webpage.

6. Click the options dropbox middle-left of the webpage, and select Download in .Zip file to download all selected documents into a .Zip file.

7. Scroll to the bottom of the page to select the next subset of transcripts.

8. Repeat steps five through seven to downloaded all transcripts in .Zip files.

9. Create a new local directory titled 'transcripts', unzip all .Zip files, moving all transcripts to this newly created directory.

10. Review the transcripts in the 'transcripts' directory. The filenames align with the filename column in the 1.firm_score.xlsx spreadsheet. There will be multiple transcripts with the same name e.g., Air New Zealand Limited - ShareholderAnalyst Call.pdf. Consult filename and calltime columns in 1.firm_score.xlsx to identify the correct transcripts according to date e.g., 201510 is October 2015 , deleting the incorrect duplicates. After, the subset of 258 transcripts will exist amongst the full set in the transcript directory.

1.3 Implementation

This section highlights the code to process earnings call transcripts, execute the StanfordCoreNLP and compare the results. The implementation was partitioned into three Python functions within the finance-751-cmcd398.py script (1.5.3). This section provides a high level overview of the code with further details described in the code comments. Transcripts have a common structure. The first three pages are front-matter. The last page is the legal disclaimer. Some transcripts don't have Q&A sections while others have multiple. The transcripts without Q&A sections are isolated and excluded during processing. Transcripts with multiple Q&A sections are manually condensed prior to processing but excluded during comparison.

1.3.1 Variables

The definition of several variables and arrays take place prior to implementation.

1. Set strings describing the relative paths for the 1.firm_score.xlsx file, transcript directory, selected transcript directory to move 258 transcripts of interest, transcript directory for processed transcripts after removing Q&A sections, documents.txt file, documents_ids.txt, and processed text directory.
2. Review each transcript in the 1.firm_score.xlsx filename column to record the page number for the first page of the Q&A section, appending each value to the end of an array. If no Q&A section exists, record a value of 4. **The preservation of order is imperative with the position of the page number matching the position of the filename in the filename list from 1.firm_score.xlsx**.
3. Set an array listing the set of company ids from the 1.firm_score.xlsx spreadsheet aligning with an array listing the cumulative position of the final transcript corresponding to the company id. For example, Air New Zealand (ANZ) has 11 transcripts. Auckland International Airport (AIA) has 13 transcripts. Therefore, ANZ and AIA have values of 11 and 24, respectively, in the cumulative position array.
4. Set strings describing the relative paths for output files, results spreadsheet, and firm scores outputs from the StanfordCoreNLP.
5. Set binary variables (TRUE or FALSE) to control the execution of the below functions.

1.3.2 Prepare_documents.py

This function isolates the Q&A sections of each transcript, converts each transcript to a line in a text file, and returns the document text file and identification. The following sequence of functions are nested within, called on in the order below.

1. **get_transcripts** extracts a list of filenames from the 1.firm_score.xlsx spreadsheet, transferring transcripts of interest to the transcripts selected directory.
2. **remove_transcript_metadata** deploys the pdfrw package to extract each page of the Q&A section per transcript, using the array denoting the starting page number for the Q&A section, creating a processed transcript stored in the transcripts processed directory.
3. **create_ids** creates various forms of identification in data frames for comparison while excluding transcripts without Q&A sections.
4. **create_documents_text** deploys the pdfminer package to convert each processed transcript into a single line of text, appending each line to the document.txt file to use as an input for the StanfordCoreNLP.

1.3.3 Perform_stanford_nlp.py

This function executes each one of the five Python functions integral to StanfordCoreNLP in the following order. The provision of two separate dictionaries (NZD/AUS and US) informs analysis.

1. **parse.py** to parse the raw documents.
2. **clean_and_train.py** to clean, remove stopwords, and named entities in the parsed documents text file.
3. **create_dict.py** to create the expanded dictionary.
4. **score.py** to score the document. This implementation uses the TF-IDF weights used in the article.
5. **aggregate_firms.py** to aggregate the scores to the firm-time level.

Complete steps one, two and three. Next, replace the expanded_dictionary.csv in the dict directory with the AUS/NZD trained dictionary. It is possible to manually edit these dictionaries in attempts to improve scores. However, the provided dictionaries trained to ascertain the original scores. Therefore, the provided dictionaries were left unchanged in replicating scores. Next, Run score.py and aggregate_firms.py, saving the scores_TFIDF.csv as an xlsx file to the comparisons directory. Repeat steps four and five with the US dictionary.

1.3.4 Compare_results.py

This function combines a formatted 1.firm_scores.xlsx document with the TF-IDF output scores from perform_stanford_nlp.py by merging data frames on document identification in order to make comparisons. Compare_results.py must be repeated for both dictionaries. After, combine both comparison spreadsheets to compare results from both sets of dictionaries, deleting duplicate values.

1.4 Comparison

This section compares our replication of the measures for corporate culture across the five values (Innovation, Integrity, Quality, Respect, Teamwork) using NZ/AUS and US dictionaries. We acknowledge the provided scores have slightly shorter document lengths, likely from different pdf to text conversion methodologies. Our analysis detected a few abnormalities in the aforementioned subset of transcripts, omitting the majority of Q&A sections (1.5.1), in addition to a subset of transcripts not including Q&A sections but trained on presentation sections. The author's emphasize the presentation sections in transcripts are likely not a true reflection of company culture as edited by corporate lawyers and PR personal. Subsequently, we exclude these transactions during processing.

1.4.1 Accuracy Measures

Absolute and percentage differences between our replication and the provided results are displayed in the 751-comparison.xlsx workbook. However, we utilize the following equations to measure the accuracy of our replication across companies, values, and total results.

$$\text{Individual} = 1 - \frac{\sum_i |\text{New}_{i,j,k} - \text{Old}_{i,j,k}|}{\sum_i \text{Old}_{i,j,k}} \forall j,k \quad (1)$$

$$\text{Total} = 1 - \frac{\sum_i \sum_j \sum_k |\text{New}_{i,j,k} - \text{Old}_{i,j,k}|}{\sum_i \sum_j \sum_k \text{Old}_{i,j,k}} \quad (2)$$

$$\text{Company} = 1 - \frac{\sum_i \sum_k |\text{New}_{i,j,k} - \text{Old}_{i,j,k}|}{\sum_i \sum_k \text{Old}_{i,j,k}} \forall j \quad (3)$$

$$\text{Value} = 1 - \frac{\sum_i \sum_j |\text{New}_{i,j,k} - \text{Old}_{i,j,k}|}{\sum_i \sum_j \text{Old}_{i,j,k}} \forall k \quad (4)$$

$$i \epsilon \{1, ..., N\} \quad (5)$$

$$j \epsilon \{\text{Air New Zealand,...,Westpac Banking Corporation}\} \quad (6)$$

$$k \epsilon \{\text{Innovation,Integrity, Quality, Respect, Teamwork}\} \quad (7)$$

1.4.2 Results

Individual, **Company**, **Value**, and **Total** measure the accuracy of our replication for a specific company and value, company across all values, value across all companies, and across all values and companies respectively. The accuracy results are displayed in a matrix (1.5.2). There are a few abnormalities. The value Teamwork for Goodman Property Trust is NA as both values in the original results are zero. Our replication for Teamwork using the NZD/AUS dictionary, and Respect using the US dictionary, deviate relatively from provided figures in our replication. The later driven by material differences in Infratil's replication (-89%) and 80% accuracy for Westpac Banking Corporation. The remaining results from the Respect value using the US dictionary are above $\tilde{8}0\%$. However, all Teamwork results using the NZD/AUS dictionary are above 83%, not raising cause for concern. The **Company** level of accuracy is above 90% for all Company IDs. Each **Value** level of accuracy is above 90% except for the Respect value measured by the US dictionary ($\tilde{8}0\%$). Finally, the **Total** level of accuracy is 93%. Discrepancies may be caused by small differences in documents lengths, or abnormalities when parsing documents using StanfordCoreNLP. In summary, our results are highly accurate and satisfactory across Company IDs and Values, providing supporting evidence our replication is successful.

References

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*(7), 3265–3315.

1.5 Appendix

1.5.1 Transcripts with Multiple Q&A Sections

The following transcripts have multiple Q&A sections. The sections are consolidated into one section by deleting the presentation material in between the Q&A sections. However, they are excluded from comparison calculation as Helen only uses the last Q&A section. We took the perspective the last section alone does not proxy for the entire Q&A sections in the transcript. Therefore, not suitable for measuring corporate culture given the document lengths.

- Australia and New Zealand Banking Group Limited - ShareholderAnalyst Call.pdf,
- Bank of Queensland Ltd. - ShareholderAnalyst Call.pdf
- Commonwealth Bank of Australia - ShareholderAnalyst Call.pdf
- Infratil Limited - AnalystInvestor Day.pdf
- Infratil Ltd. - AnalystInvestor Day.pdf
- National Australia Bank Limited - ShareholderAnalyst Call.pdf

1.5.2 Result Matrix

Firm	Document Length	Innovation (ANZ)	Integrity (ANZ)	Quality (ANZ)	Respect (ANZ)	Teamwork (ANZ)	Innovation (US)	Integrity (US)	Quality (US)	Respect (US)	Teamwork (US)	**Company**
Air New Zealand Limited	96%	95%	91%	97%	96%	92%	93%	94%	95%	90%	93%	**97%**
Auckland International Airport Limited	96%	95%	94%	95%	96%	87%	94%	91%	93%	92%	93%	**98%**
Australia New Zealand Banking Group Limited	96%	93%	94%	93%	94%	93%	94%	90%	94%	84%	95%	**96%**
Bank of Queensland Limited	96%	94%	93%	95%	93%	83%	95%	91%	93%	93%	89%	**97%**
Bendigo and Adelaide Bank Limited	96%	93%	95%	94%	94%	94%	94%	94%	94%	89%	95%	**97%**
Commonwealth Bank of Australia	96%	92%	92%	94%	92%	93%	93%	92%	92%	88%	93%	**96%**
Contact Energy Ltd	94%	90%	93%	93%	94%	93%	91%	93%	91%	88%	91%	**94%**
Fisher Paykel Healthcare Corporation Limited	95%	93%	92%	94%	92%	84%	94%	92%	93%	84%	93%	**97%**
Fletcher Building Ltd	96%	93%	94%	95%	94%	96%	92%	93%	94%	95%	95%	**96%**
Goodman Property Trust	96%	93%	91%	93%	97%	95%	94%	90%	90%	89%	#N/A	**97%**
Infratil Limited	95%	92%	94%	93%	94%	95%	94%	90%	92%	-89%	93%	**94%**
Kiwi Income Property Trust	95%	96%	95%	95%	81%	95%	96%	87%	93%	96%	95%	**98%**
National Australia Bank Limited	96%	94%	94%	95%	94%	94%	94%	93%	94%	89%	94%	**97%**
Spark New Zealand Limited	95%	94%	96%	93%	96%	88%	94%	96%	92%	89%	95%	**96%**
Telecom Corp of New Zealand Ltd	97%	95%	92%	94%	96%	84%	93%	88%	94%	85%	95%	**97%**
Vector Limited	94%	92%	93%	92%	92%	90%	91%	92%	92%	87%	92%	**95%**
Westpac Banking Corporation	96%	96%	94%	94%	94%	95%	94%	91%	94%	79%	93%	**98%**
Value	**96%**	**94%**	**93%**	**94%**	**94%**	**91%**	**94%**	**92%**	**93%**	**80%**	**93%**	93%

Figure 1: Results Matrix

1.5.3 Python

```python
# Descriptions
# This script/function implements the StanfordNLP to score corporate culture,
    replicating the production of inputs in Option 1 as outputs
# Inputs for Option 1 include:
# 1.  Firm_score.xlsx contains five scores estimated with two different dictionaries
    for all calls. Scores ended with 1 (for example, integrity1) are estimated with
    the dictionary trained on the 258 call transcripts included in this sample.
    Scores ended with 2 (for example, integrity2) are estimated with the dictionary
    from the original paper (Table IA3 in the Internet Appendix). Other variables
    include document_id (used in your coding), filename (file name used by CapitalIQ)
    , firm_id (firm name) and call time (year and month of the call).
# 2.  Expanded_dict1.csv is the culture dictionary trained with the 258 call
    transcripts (the new dictionary).
# 3.  Expanded_dict2.csv is the culture dictionary from the original paper (the
    original dictionary).
# 4.  Word_contributin_TFIDF1.csv (Word_contributin_TFIDF2.csv) contains word
    contribution based on TFIDF score estimated with the new dictionary (the original
     dictionary).
# 5.  The Li, Mai, Shen and Yan (2021) paper and the Internet Appendix of this paper.
#
# Author: Connor McDowall
# Date: 25th August 2021

# Imports
# Transcript Processing Modules
import pandas as pd
from pathlib import Path
import shutil as sh
from pdfrw import PdfReader, PdfWriter
import pdfminer as pdfm
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser
import io
import datefinder as dtf
# General Python Modules
import datetime
import functools
import logging
import sys
import math
import os
import pickle
import gensim
import itertools
from pprint import pprint
from collections import Counter, defaultdict, OrderedDict
from tqdm.auto import tqdm
from typing import Dict, List, Optional, Set
from multiprocessing import Pool
from operator import itemgetter
from tqdm import tqdm as tqdm

# StanfordNLP Specific Functions
from culture import culture_models, file_util, preprocess, culture_dictionary,
    preprocess_parallel
from stanfordnlp.server import CoreNLPClient
import global_options
import parse
import clean_and_train
import create_dict
import score
import aggregate_firms

# Functions
def get_transcipts(firm_score_xlsx, transcript_directory,transcript_selected):
    """Locates and isolates transcripts for processing

    Args:
        firm_score_xlsx (xlsx): Excel file containing the initial list of transcripts
        transcript_directory (str): Source of all transcripts
        transcript_selected (str): Destination for transcripts of interest

    Returns:
        transcript_list (list): List of transcript filenames
        calltimes (list): List of calltimes
    """
    # Get list of filenames
    firms_df = pd.read_excel(firm_score_xlsx)
    firms_df=firms_df.dropna()
```

```
    firms_df.columns = firms_df.iloc[0]
    firms_df = firms_df.drop(2)
    firms_df = firms_df.reset_index(drop=True)
    transcript_list = firms_df['filename'].tolist()
    # Get list of calltimes for the firm ID
    calltimes = firms_df['calltime'].tolist()
    # Copy file into selection if exists
    files_found = 0
    files_to_find = len(transcript_list)
    missing_files_list = []
    for filename in transcript_list:
        transcipt_x = Path(transcript_directory +'/'+filename)
        if transcipt_x.is_file():
            transcipt_y = Path(transcript_selected +'/'+filename)
            sh.copy(transcipt_x,transcipt_y)
            files_found = files_found + 1
        else:
            missing_files_list.append(filename)
    missing_files = files_to_find - files_found
    if missing_files > 0:
        print('You are missing the following transcripts...')
        print(missing_files_list)
    else:
        print('All transcripts found')
    return transcript_list,calltimes

def create_ids(transcript_list,qa_num, company_ids_set, company_ids_order, documents_ids_text, calltimes):
    """Creates document identification, updates transcript list to only include transcript lists
    with Question and Answer Sections, and creates dataframe to compare results.

    Args:
        transcript_list (list): List of transcript filenames
        qa_num (list): List of page numbers denoting the start of question and answer sections
        company_ids_set (list): List of company names
        company_ids_order (list): List of numbers referencing number of file relating to one company
        documents_ids_text (str): Directory to store document id list as a text file
        calltimes (list): List of calltimes

    Returns:
        updated_transcript_list (list): List of updated filenames
        updated_document_ids (list): List of updated document ids
        updated_firm_id (list): List of updated firm ids
        output_df (dataframe): Dataframe with document information
    """
    # Initial lists
    document_ids = []
    firm_id = []
    # Updated lists
    updated_document_ids = []
    updated_firm_id = []
    updated_transcript_list = []
    updated_calltimes = []
    # Assigns document id
    idx = 0
    for i in range(len(transcript_list)):
        document_ids.append(str(i + 1)+'.F')
        if i < company_ids_order[idx]:
            firm_id.append(company_ids_set[idx])
        else:
            idx = idx +1
            firm_id.append(company_ids_set[idx])
    # Updates lists to remove entries with no question and answer sections
    for j in range(len(qa_num)):
        if qa_num[j] != 4:
            updated_document_ids.append(document_ids[j])
            updated_firm_id.append(firm_id[j])
            updated_transcript_list.append(transcript_list[j])
            updated_calltimes.append(calltimes[j])
    # Creates document_id text file
    with open(documents_ids_text, "w") as file:
        # Clear the file
        file.truncate(0)
        for element in updated_document_ids:
             file.write(element + "\n")
        file.close()
    # Creates a dataframe with updated transcript list
    output_df = pd.DataFrame(list(zip(updated_document_ids, updated_transcript_list, updated_firm_id)),
               columns =['document_id', 'filename', 'firm_id'])
```

```python
    # Creates id2firsm csv
    for i in range(len(updated_calltimes)):
        val = updated_calltimes[i]
        new_val = int(str(val)[:4])
        updated_calltimes[i] =  new_val
    id2firms_df = pd.DataFrame(list(zip(updated_document_ids,updated_firm_id, updated_calltimes)),
                   columns =['document_id', 'firm_id','time'])
    print(id2firms_df.head())
    id2firms_df.to_csv('data/input/id2firms.csv')
    return updated_transcript_list, updated_document_ids, updated_firm_id, output_df

def remove_transcript_metadata(transcript_list,qa_num,transcript_selected, transcript_processed):
    """Removes front matter, table of contents, call participants, and copyright disclaimer
    to process transcripts to a format suitable for combination. This is possible as the
    format is consistent for all earnings call transcripts.

    Args:
        transcript_list (list): List of transcript filenames
        qa_num (list): List of page numbers denoting the start of question and answer sections
        transcript_selected (str): String of selected transcipt directory
        transcript_processed (str): String of processed transcript directory
    """
    # Count for
    i = 0
    # Create copy, remove pages, and move to processed directory
    for filename in transcript_list:
        # Defines pdfs
        input_pdf = Path(transcript_selected +'/'+filename)
        output_pdf = Path(transcript_processed +'/'+filename)
        # Defines objects
        reader_input = PdfReader(input_pdf)
        writer_output = PdfWriter()
        for page_x in range(len(reader_input.pages)):
            # Adds pages excluding sections prior to Q&A section and legal disclaimer
            if page_x >= qa_num[i]-1 and page_x < (len(reader_input.pages)-1):
                writer_output.addpage(reader_input.pages[page_x])
        writer_output.write(output_pdf)
        i = i + 1
    return

def create_documents_text(transcript_list,transcript_processed, text_processed, documents_text):
    """Creates documents.txt file for the Stanford NLP

    Args:
        transcript_list(str): List of processed transcipts
        transcript_processed (str): String of processed transcript directory
        text_processed (str): Directory to store text file
        documents_text (str): Directory for documents.txt file

    Returns:
        documents_test_list (list): Returns a list of processed transcript document strings
    """
    # Adapted from https://towardsdatascience.com/pdf-text-extraction-in-python-5b6ab9e92dd
    # Erase object contents to reset the textfile
    with open(documents_text, "r+") as file:
        file.truncate(0)
        file.close()
    # Creates empty list
    documents_test_list = []
    # Begin extracting files
    for file_name in transcript_list:
        file_pdf = Path(transcript_processed +'/'+file_name)
        file_text = io.StringIO()
        with open(file_pdf, 'rb') as in_file:
            parser = PDFParser(in_file)
            doc = PDFDocument(parser)
            rsrcmgr = PDFResourceManager()
            device = TextConverter(rsrcmgr, file_text, laparams=LAParams())
            interpreter = PDFPageInterpreter(rsrcmgr, device)
            for page in PDFPage.create_pages(doc):
                interpreter.process_page(page)
        # Extract text to and remove characters
        textname = Path(text_processed +'/output.txt')
        with open(textname, "w") as file:
            file.write(file_text.getvalue())
```

```python
            file.close()
        # Print the lines
        with open(textname, "r+") as file:
            line = file.read().replace("\n", " ")
            file.truncate(0)
            file.close()
        # Write line to the documents file
        with open(documents_text, "a") as file:
            file.write(line)
            if file_name != transcript_list[-1]:
                file.write("\n")
            file.close()
        # Create list of texts and dates
        documents_test_list.append(line)
    return documents_test_list

def prepare_documents(firm_score_xlsx, transcript_directory, transcript_selected, transcript_processed, text_processed, documents_text, documents_ids_text, qa_num, company_ids_set,company_ids_order):
    """ Isolate transcripts of interest, process Q&A sections, and create document files

    Args:
        firm_score_xlsx (xlsx): Excel file containing the initial list of transcripts
        transcript_directory (str): Source of all transcripts
        transcript_selected (str): Destination for transcripts of interest
        transcript_processed (str): Directory for processed transcripts
        text_processed (str): Directory to store text file
        documents_text (str): Directory for documents.txt file
        documents_ids_text (str): Directory to store document id list as a text file
        qa_num (list): List of page numbers denoting the start of question and answer sections
        company_ids_set (list): List of company names
        company_ids_order (list): List of numbers referencing number of file relating to one company

    Returns:
        documents_test_list (list): Returns a list of processed transcript document strings
        document_ids (list): List of document ids
        firm_id (list): List of firm ids
        output_df (df): Dataframe with document information
    """
    # Prepares the documentation
    # Get list of transcripts
    transcript_list, calltimes = get_transcipts(firm_score_xlsx, transcript_directory, transcript_selected)
    # Isolates Q&A sections while removing legal disclaimers
    remove_transcript_metadata(transcript_list,qa_num,transcript_selected, transcript_processed)
    # Creates supplementary identification (Changed here to remove files without text files)
    transcript_list, document_ids, firm_id, output_df = create_ids(transcript_list, qa_num, company_ids_set, company_ids_order, documents_ids_text, calltimes)
    # Creates the documents.txt file, documents ids, firm_ids, and dataframe of outputs
    documents_test_list = create_documents_text(transcript_list,transcript_processed, text_processed, documents_text)
    # Saves csv for comparison
    dataframe_file = Path('data/input/results.csv')
    output_df.to_csv(dataframe_file)
    return documents_test_list, document_ids, firm_id, output_df

def perform_stanford_nlp():
    """Executes Stanford NLP algorithm on processed documentation via
    """
    print("Implementing Stanford NLP...")
    # Creates variables and directories in global options
    exec(open("global_options.py").read())
    # Step 1: Use `python parse.py` to use Stanford CoreNLP to parse the raw documents.
    exec(open("parse.py").read())
    # Step 2: Use `python clean_and_train.py` to clean, remove stopwords, and named entities in parsed `documents.txt`
    exec(open("clean_and_train.py").read())
    # Step 3: Use `python create_dict.py` to create the expanded dictionary.
    exec(open("create_dict.py").read())
    # Step 4: Use `python score.py` to score the documents.
    exec(open("score.py").read())
    # Step 5: Use `python aggregate_firms.py` to aggregate the scores to the firm-time level.
    exec(open("aggregate_firms.py").read())
    return
```

```python

def compare_results(results,output_scores):
    """Creates comparison excel sheets with helens results

    Args:
        results (str): Directory to the document id files
        output_scores (str): Directory for scoring sheets
    """
    # Load in the results
    output_df = pd.read_csv(results)
    # Set directories
    tf = 'firm_scores_TF.csv'
    tfidf = 'firm_scores_TFIDF.csv'
    wfidf = 'firm_scores_WFIDF.csv'
    helen_results = 'outputs/scores/firm_score_helen.xlsx'
    firm_scores_tf = Path(output_scores+'/'+tf)
    firm_scores_tfidf = Path(output_scores+'/'+tfidf)
    firm_scores_wfidf = Path(output_scores+'/'+wfidf)
    helen_results = Path(helen_results)
    # Read csv and excel files
    firm_scores_tf_df = pd.read_csv(firm_scores_tf)
    firm_scores_tfidf_df = pd.read_csv(firm_scores_tfidf)
    firm_scores_wfidf_df = pd.read_csv(firm_scores_wfidf)
    helen_results = pd.read_excel(helen_results)
    # Merge results with dataframes for comparison
    target_df = firm_scores_tfidf_df
    user_results_df = pd.merge(output_df,  target_df,how = 'left',on = output_df.index)
    comparison_df = pd.merge(user_results_df,helen_results,how = 'left',on = ['document_id'])
    print('Please enter a filename')
    filename = input()
    # Save comparison csv
    file_string = 'outputs/comparisons'+'/'+filename+'.xlsx'
    comparison_df.to_excel(file_string)
    return

# Inputs - established all the directories for the locations
# Inputs for processing
firm_score_xlsx = 'data/input/option-1/1.firm_score.xlsx'
transcript_directory = 'data/input/transcripts'
transcript_selected = 'data/raw/selected_transcripts'
transcript_processed = 'data/processed/processed_transcripts'
text_processed = 'data/processed/processed_text'
documents_text = 'data/input/documents.txt'
documents_ids_text = 'data/input/document_ids.txt'
# Creates array of pages numbers indicating the start of the Q&A section for each PDF
# Note: This is labourous but necessary. Values of 4 indicate no Q&A section in the document,
# starting at the presentation section
air_nz_num=[8,10,10,7,8,10,8,11,8,8,8]
aia_num = [4,4,12,12,12,9,10,15,11,10,10,10,10] # Changed to 4
anz_num = [14,6,10,11,11,13,11,13,11,10,11,13,13,8,7,8,7,10,11,12,13,11,12,10,11,11,13,12,11,12,8]
bql_num = [24,12,10,11,11,12,11,11,14,11,9,16,12,13,16,14,13,12,13,10,10,11,12,12,12,13,8]
bab_num = [4,10,10,12,11,10,10,10,14,15,10,10,10,12,10,10,12,12,15,15,9,10]
cba_num = [5,11,11,12,12,11,12,11,10,10,4,11,11,10,11,11,11,10,12,12,11,12,12,12,6,6,6,7,8] # Changed to 4 (29)
ce_num = [8,4] # Changed to 4
fph_num = [9,8,9,8,9,8,8,7,8,8,8]
fbu_num = [12,10,11,10,9,10,10,9,10,11,9]
gpt_num = [10,9]
il_num = [15,15,15,15,13,14,13,12,13,15,15,16,13]
kip_num = [11,10]
nab_num = [12,4,4,13,12,14,10,18,15,9,10,11,11,12,13,13,10,12,15,10,9,10,10,12,11,9,8,15,7,7,6] # Changed to 4 (31)
spk_num = [16,12,12]
tnz_num = [15,14,11,16,14,12,13,9,16]
vec_num = [9,12,9,9,10,9,9,8,12,11,10,9]
wpc_num = [12,19,12,14,14,13,13,11,12,11,11,12,11,11,16,14,12,12,12,11,11,12,10,11,7,8,7,7,7]
# Combines the arrays
qa_num = [air_nz_num,
            aia_num,
            anz_num,
            bql_num,
            bab_num,
            cba_num,
            ce_num,
```

```python
            fph_num,
            fbu_num,
            gpt_num,
            il_num,
            kip_num,
            nab_num,
            spk_num,
            tnz_num,
            vec_num,
            wpc_num]

qa_num = air_nz_num+aia_num+anz_num+bql_num+bab_num+cba_num+ce_num+fph_num+fbu_num+gpt_num+il_num+kip_num+nab_num+spk_num+tnz_num+ vec_num + wpc_num
# Sets list for company ids
company_ids_set = ['Air New Zealand Limited','Auckland International Airport Limited','Australia New Zealand Banking Group Limited', 'Bank of Queensland Limited','Bendigo and Adelaide Bank Limited','Commonwealth Bank of Australia', 'Contact Energy Ltd','Fisher Paykel Healthcare Corporation Limited','Fletcher Building Ltd','Goodman Property Trust','Infratil Limited','Kiwi Income Property Trust','National Australia Bank Limited','Spark New Zealand Limited','Telecom Corp of New Zealand Ltd','Vector Limited','Westpac Banking Corporation']
company_ids_order = [11,24,55,82,104,133,135,146,157,159,172,174,205,208,217,229,258]
# Inputs for comparison
output_scores = 'outputs/scores'
results = 'data/input/results.csv'
output_word_contributions = 'outputs/scores/word_contributions'
firm_scores_tf = 'outputs/scores/firm_scores_TF.csv'
firm_scores_tfidf = 'outputs/scores/firm_scores_TFIDF.csv'
firm_scores_wfidf = 'outputs/scores/firm_scores_WFIDF.csv'
####################################################
# Function Calls
# Set binary variables to control function calls
transcript_preparation = False
stanford_nlp_implementation = False
results_comparison = True
# Executes functions based on binary variables
if transcript_preparation == True:
    # Prepare the documents
    print("Preparing documents...")
    documents_test_list, document_ids, firm_id, output_df = prepare_documents(firm_score_xlsx, transcript_directory, transcript_selected, transcript_processed, text_processed, documents_text, documents_ids_text, qa_num, company_ids_set, company_ids_order)
if stanford_nlp_implementation == True:
    # Implements Stanford NLP
    perform_stanford_nlp()
if results_comparison == True:
    print('Comparing results...')
    compare_results(results,output_scores)
# Note: Australia and New Zealand Banking Group Limited - ShareholderAnalyst Call.pdf, Bank of Queensland Ltd. - ShareholderAnalyst Call.pdf
# Commonwealth Bank of Australia - ShareholderAnalyst Call.pdf, Infratil Limited - AnalystInvestor Day.pdf, Infratil Ltd. - AnalystInvestor Day.pdf
# National Australia Bank Limited - ShareholderAnalyst Call.pdf
```

Measuring Corporate Culture Using Machine Learning

Authors: Kai Li, Feng Mai, Rui Shen & Xinyan Yan, 2020

Connor McDowall

Evolution of Corporate Culture

Top and bottom-ranked S&P500 firms by corporate cultural values

A. Top- and bottom-ranked S&P 500 firms, 2001–2006

Innovation	Integrity
Procter & Gamble Co	Fannie Mae
Nvidia Corp	Franklin Resources Inc
Gap Inc	Kate Spade & Co
Lauder (Estee) Cos Inc	Encompass Health Corp
PTC Inc	Synovus Financial Corp
Penney (JC) Co	Northwest Airlines Corp
Harman International Inds	EMCOR Group Inc
Home Depot Inc	Exelon Corp
Kate Spade & Co	Service Corp International
BroadVision Inc	Compuware Corp
Luby's Inc	VF Corp
Genuine Parts Co	Luby's Inc
Univision Communications	M & T Bank Corp
Patterson Cos Inc	Amazon.Com Inc
Archer-Daniels-Midland Co	TECO Energy Inc
Tyson Foods Inc	Bristol-Myers Squibb Co
Automatic Data Processing	Bausch & Lomb Hldgs
Texas Instruments Inc	Regions Financial Corp
Tribune Media Co	Citigroup Inc
CenturyLink Inc	Equity Residential

B. Top- and bottom-ranked S&P 500 firms, 2007–2012

Innovation	Integrity
Nvidia Corp	Tribune Media Co
Procter & Gamble Co	Wynn Resorts Ltd
Adobe Inc	Beam Inc
Discovery Inc	Ambac Financial Group Inc
Lauder (Estee) Cos Inc	Intercontinental Exchange
Netflix Inc	Lockheed Martin Corp
Salesforce.com Inc	Exelon Corp
VF Corp	American Electric Power Co
Fossil Group Inc	Kate Spade & Co
Kate Spade & Co	Lorillard Inc
Genuine Parts Co	Bausch & Lomb Hldgs
CVS Health Corp	Public Storage
Univision Communications	Sigma-Aldrich Corp
Archer-Daniels-Midland Co	Wyndham Destinations Inc
American Greetings	VF Corp
Texas Instruments Inc	Equity Residential
Ryerson Holding Corp	Winn-Dixie Stores Inc
DXC Technology Co	Host Hotels & Resorts Inc
Patterson Cos Inc	Spire Inc
Cintas Corp	Luby's Inc

C. Top- and bottom-ranked S&P 500 firms, 2013–2018

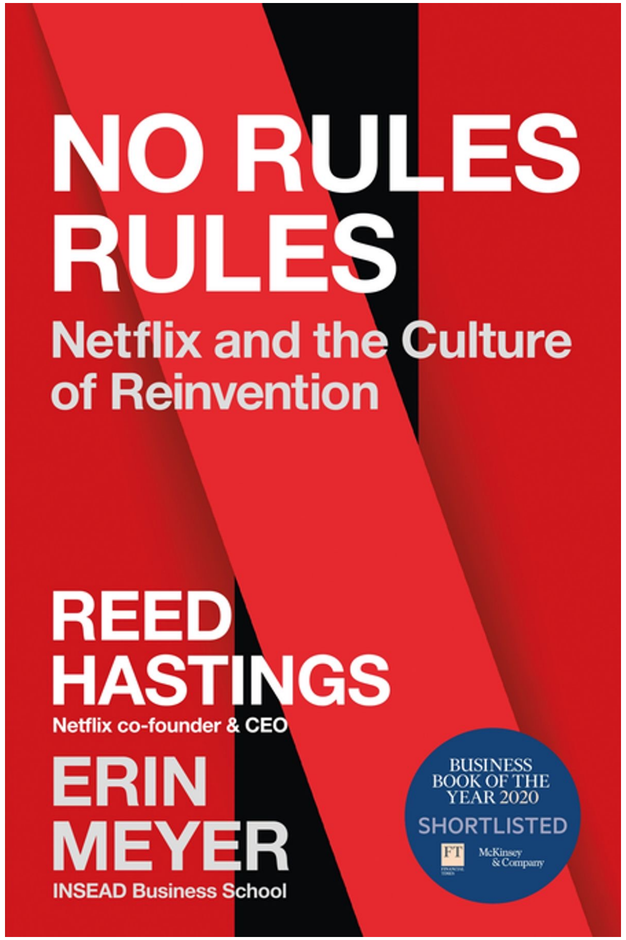
Innovation	Integrity
Netflix Inc	Blackrock Inc
Fossil Group Inc	Wynn Resorts Ltd
Nike Inc	Ambac Financial Group Inc
Lauder (Estee) Cos Inc	Big Lots Inc
Procter & Gamble Co	Intercontinental Exchange
Adobe Inc	Gap Inc
Salesforce.com Inc	Genworth Financial Inc
Acuity Brands Inc	U.S. Bancorp
Twitter Inc	News Corp
Facebook Inc	United States Steel Corp
Archer-Daniels-Midland Co	National Fuel Gas Co
Genuine Parts Co	Idexx Labs Inc
FleetCor Technologies Inc	Cooper Cos Inc (The)
Univision Communications	SBA Communications Corp
LKQ Corp	IDACORP Inc
Philip Morris International	ONEOK Inc
Cintas Corp	Ryder System Inc
Costco Wholesale Corp	CenterPoint Energy Inc
Emerson Electric Co	Williams Cos Inc
Texas Instruments Inc	Public Storage

Corporate Culture?

What is the purpose of having a corporate culture? What does it mean?

Definitions and prior literature inform nebulous nature

- A system of shared values and norms defining what is important, appropriate attitudes, and behaviors for organizational members (O'Reilly and Chatman, 1996)
- 'Path dependent and can be shaped by major corporate events (Weber et al., 1996; Guiso et al., 2015; Graham et al., 2018; Grennan, 2018)
- Important because employees will inevitably face choices that cannot be properly regulated ex-ante (O'Reilly, 1989; Kreps, 1990)
- Extant literature has limited large sample evidence, possibly due to nebulous nature creating measurement issues



Research Intent

What is the purpose of this article? What is a strong corporate culture?

Paper claims to address issues facing textual analysis

- Proposition of semi-supervised machine learning algorithm to measure corporate culture
- A methodological contribution to the accounting/finance literatures by introducing word embedding models to score corporate cultures
  - Assess management's alignment with corporate values, and ability to lead by example
  - Measure the true representation of corporate culture, applying less weighting to frequently occurring words
  - Explore implications of having a strong culture on business outcomes
- Innovate within the field of textual analysis through a better quantify semantics via vectorization, in addition to syntactic expressions
  - Previous methods have firm policy proxies explaining relationships with culture, or relying on surveys

Data & Methodology

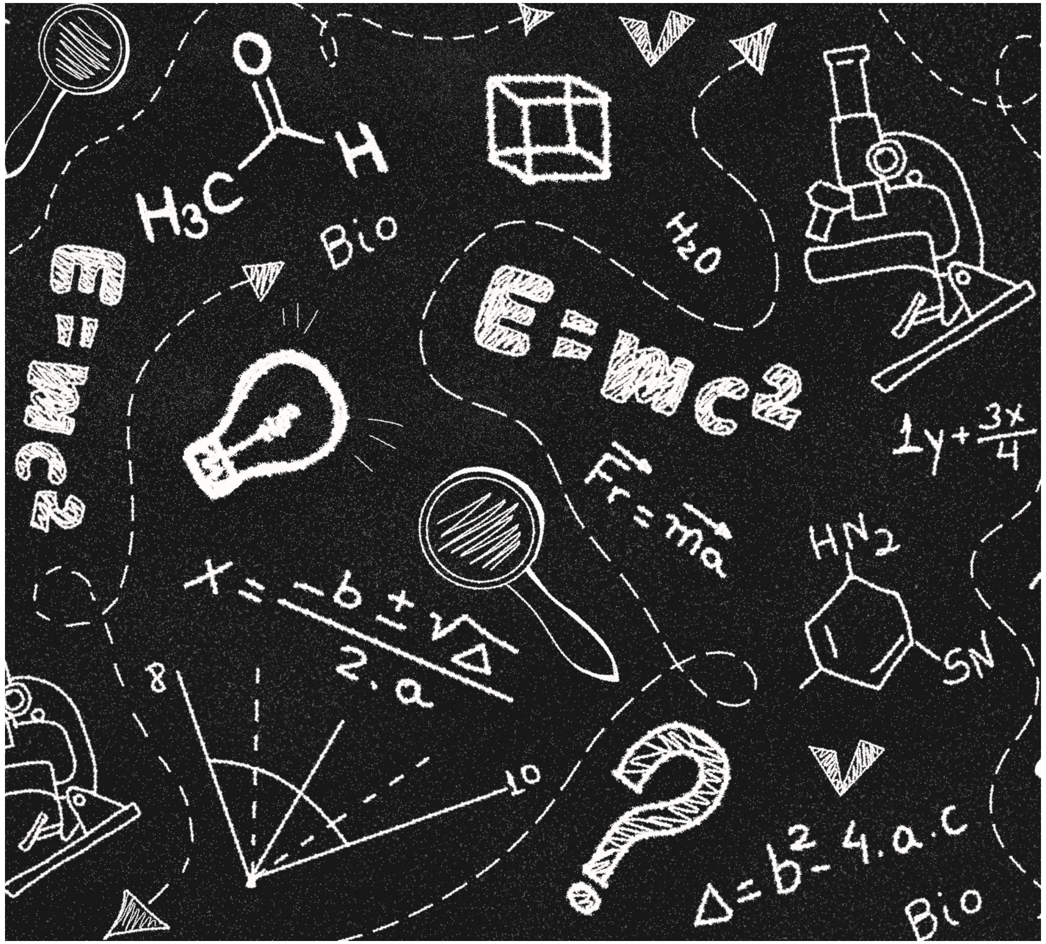
Connor McDowall

Overview of Methodology

This section provides a high-level overview of the methodology

The authors follow the subsequent implementation:

1) Data, Preprocessing and Parsing, and Learning Phrases
    1) Earnings call to score corporate culture
    2) Data, preprocessing and parsing, and learning phrases
2) Word Embedding, word2vec, and Model Training
3) Measuring corporate culture using word2vec
    1) Seed words
    2) Generating the culture dictionary
    3) Scoring corporate culture
4) Validating measures of corporate culture
    1) Validation Tests
    2) Corporate Culture and its markers
    3) Other ways of measuring corporate culture
    4) Addressing self-promotion in calls
    5) Words with multiple senses



Data, Processing and Parsing, and Learning Phrases

Why do organizations conduct earnings calls? What is their purpose?

Executives heavily influence culture

- The most influential factor in building a firm's current culture is the current CEO, consistent with results surveying top executives
- Prior studies have used CEO attributes and behaviours to proxy corporate culture
- Subsequently, earnings call transcript deemed a suitable external sources to measure corporate culture as prominently feature chief executives and other top executives
- Call emphasis business operations, and performance, without promoting or 'window dressing' corporate culture
- Q&A section most appropriate as less likely to be scripted/vetted by corporate lawyers and investor relations
- Methodology capable of learning copious amounts of culture value-related words/weighting scheme

Data, Processing and Parsing, and Learning Phrases

Why match the extracted metadata to the compustat database ?

Authors use a comprehensive dataset

- Transcripts extracted from Thompson Reuters' StreetEvents (SE) database for January 1st, 2001, to May 25th, 2018
- Each file contains the body of a call transcript and subsequent metadata; ticker symbol, company name, title of the event, and call date
- After matching with Compustat database:
  - 209,480 QA sections mapped to 64,511 firm-year observations
- Use the Stanford CoreNLP package to preprocess and parse text, segmenting documents into sentences and word, lemmatizing words into base forms, to extract general/corpus-specific phrases
  - Phrases (collocations) crucial for gathering information
  - Identify fixed, multi–word/compound expressions
  - Identify two/three-word phrases specific to corpus

	# firm-year obs.		# firm-year obs. removed	# transcripts	# transcripts removed
Match company names in call transcripts to GVKEY					
All conference call transcripts				391,091	
Earnings call transcripts				270,879	120,212
Transcripts matched with GVKEY	**66,371**			**221,209**	49,670
Including					
Perfect match with CRSP company name		21,627			
Perfect match with Compustat company name		7,355			
Perfect match with Compustat-CRSP merged		1,238			
Ticker matching if not subject to backfilling		559			
Manual matching if no perfect match		35,075			
Nonduplicated company name in brief files		517			
Transcripts without the QA section	65,247		1,124	214,295	6,914
Transcripts with fewer than 200 words in the QA section	**64,511**		736	**209,480**	4,815
Sample formation for Table 3					
After applying 3-year rolling average	84,144				
After imposing fiscal year $\leq$ 2018	76,232		7,912		
After matching with financial data	62,664		13,568		
Final sample	**62,664**				

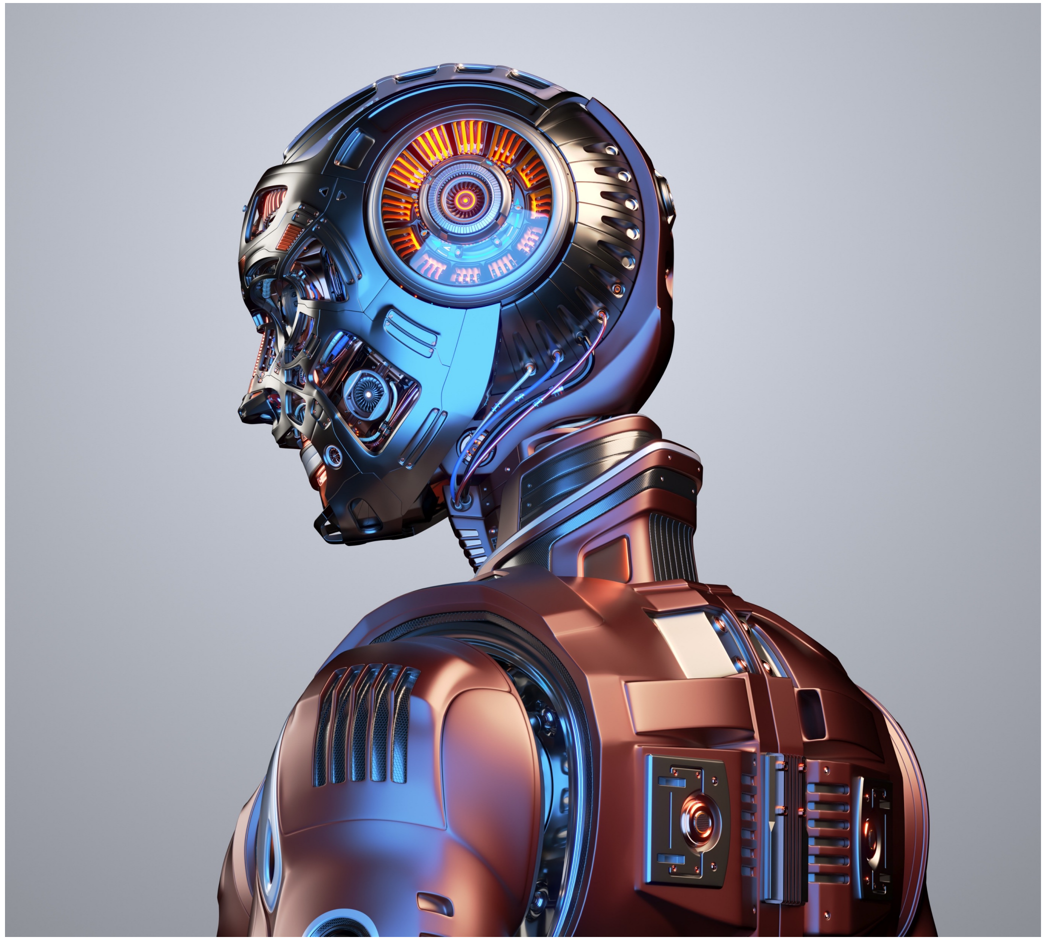
Why is it important to identify transcripts without Q&A sections?

Word Embedding, word2vec, and Model Training

What is machine learning?

Machine learning can be used for textual analysis

- Increasing reliance on automated textual analysis to extract information from corporate disclosures for research in accounting and finance
- A common method to measure sentiment is quantifying the reoccurrence of words with shared meaning, a laborious process from manual inspection and categorization of words
- Corporate culture; is often discussed in subtle, nuance fashions; can be an elusive, multi-dimensional construct; unrealistic to presume that experts could create and maintain dictionaries capable of adapting to the constant paradigm shifts in the business world
- A machine learning algorithm addresses the challenges



Introduction to Neural Networks

Neural networks inform Natural Language Processing Implementation

Introduction to Algorithms: Neural Networks

- A feedforward artificial neural network (ANN) is a series of layered perceptrons
- A linear threshold unit (LTU) feeds a weighted sum of input values into a step/activation function to determine the output. A perception is a single layer of interconnected LTUs
  - Activation functions: sigmoid, hypertangent, and linear
- Perceptions utilize a training algorithm to assess the strength of connections between perceptions
  - **Back propagation**, steepest descent, conjugate gradient, modified newton, and genetic algorithm etc.
- A perception makes predictions on an instance one at a time, re-enforcing the connection weights from incorrect LTU prediction to improve performance

LTU

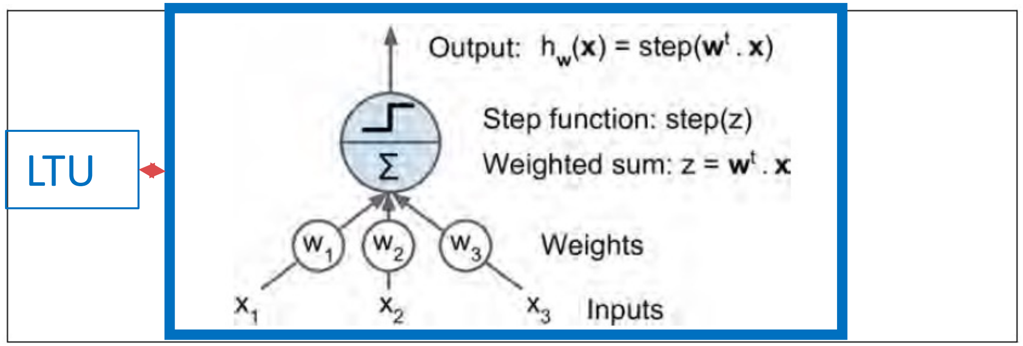


Figure 10-4. Linear threshold unit

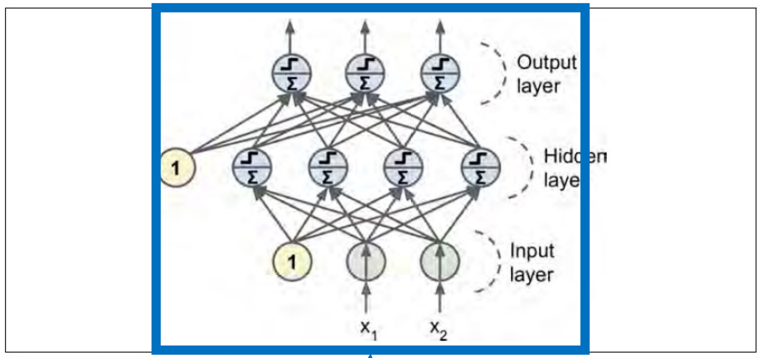


Figure 10-7. Multi-Layer Perceptron

Multi-layer Perception

Measuring Corporate Culture Using word2vec

What are some corporate values?

NLP innovate word embedding methods

- Word embedding represents semantics as a numerical vector, enabling vector arithmetic to determine relationships, assessing neighborhoods for similar meanings. This is complex given the number of combinations
- Natural language processing (NLP), word2vec, employs a 'neural network' to efficiently learn dense and low-dimensional vectors that can represent the meaning of words
- The five most-often mentioned values by the S&P 500 firms on their corporate websites, adding seed words in the transcripts, and unambiguously-related to the culture words
  1) Innovation (80%)
  2) Integrity (70%)
  3) Quality (60%)
  4) Respect (70%)
  5) Teamwork (50%)

Measuring Corporate Culture Using word2vec

How do you measure similarities between words?

Scoring corporate culture is feasible

- Scores the corporate culture by measuring the each of the five cultural values at the firm value.
- A weighted count, considering both term frequency, and inverse document frequency, accounts for both the importance of a word in a document and the significance of the word in the corpus.
- Provide summary statistics, measuring corporate culture using three year moving averages, with a final sample consisting of 7,501 firms and 62,664 firm-year observations
- Innovation and integrity are most and least frequently mentioned respectively

Measuring Corporate Culture Using word2vec

Do these associations surprise you?

A. Thirty most representative words for each cultural value in the culture dictionary

Innovation	Integrity	Quality	Respect	Teamwork
Creativity	Accountability	Dedicated	Talented	Collaborate
Innovative	Ethic	Quality	Talent	Cooperation
Innovate	Integrity	Dedication	Empower	Collaboration
Innovation	Responsibility	Customer_service	Team_member	Collaborative
Creative	Transparency	Customer	Employee	Cooperative
Excellence	Accountable	Dedicate	Team	Partnership
Passion	Governance	Service_level	Leadership	Cooperate
World-class	Ethical	Mission	Leadership_team	Collaboratively
Technology	Transparent	Service_delivery	Culture	Partner
Operational_excellence	Trust	Customer_satisfaction	Teammate	Co-operation

B. Thirty most frequently occurring words for each cultural value in the culture dictionary

Innovation			Integrity			Quality			Respect			Teamwork		
Word	%	*Cum.%*	Word	%	*Cum.%*	Word	%	*Cum.%*	Word	%	*Cum.%*	Word	%	*Cum.%*
Brand	4.24	4.24	Control	5.81	5.81	Customer	9.22	9.22	People	5.91	5.91	Partner	6.01	9.22
Technology	3.08	7.32	Management	4.93	10.74	Product	8.09	17.31	Team	5.10	11.00	Relationship	5.36	17.31
Focus	3.02	10.34	Careful	3.46	14.19	Client	5.99	23.30	Company	5.00	16.00	Discussion	5.22	23.30
Great	2.73	13.08	Honestly	2.71	16.90	Service	4.72	28.02	Hire	3.78	19.78	Together	4.61	28.02
Platform	2.53	15.61	Regulator	2.68	19.58	Build	4.09	32.11	Folk	3.61	23.39	Integrate	4.07	32.11
Ability	2.41	18.02	Honest	2.43	22.01	Deliver	3.42	35.54	Organization	3.39	26.78	Involve	3.77	35.54
Best	2.37	20.39	Safety	2.09	24.10	Network	3.30	38.84	Resource	3.11	29.89	Conversation	3.73	38.84
Design	2.19	22.58	Assure	2.01	26.11	Support	3.12	41.96	Employee	2.96	32.86	Integration	3.24	41.96
Create	2.18	24.76	Compliance	1.88	27.98	Quality	2.40	44.36	Management_team	1.91	34.77	Partnership	3.17	44.36
Solution	2.16	26.92	Trust	1.87	29.86	Sales_force	2.31	46.68	Train	1.88	36.65	Engage	2.65	46.68
Develop	2.12	29.04	Disciplined	1.82	31.68	Infrastructure	2.27	48.94	Training	1.81	38.46	Align	2.07	48.94
Success	2.00	31.04	Responsible	1.71	33.39	Supplier	2.21	51.16	Senior	1.80	40.26	Explore	1.79	51.16

Measuring Corporate Culture Using word2vec

What do you notice about the auto-correlations and correlation matrix?

B. Autocorrelations of corporate cultural values

Variable in year t	Obs.	Year $t-1$	Year $t-2$	Year $t-3$	Year $t-4$	Year $t-5$
Innovation	1,971	0.790	0.512	0.190	0.090	0.045
		[0.828]	[0.559]	[0.203]	[0.071]	[0.031]
		(0.151)	(0.301)	(0.441)	(0.475)	(0.500)
Integrity	1,971	0.695	0.361	−0.037	−0.085	−0.103
		[0.728]	[0.378]	[−0.071]	[−0.141]	[−0.160]
		(0.179)	(0.292)	(0.397)	(0.405)	(0.434)
Quality	1,971	0.738	0.417	0.052	−0.023	−0.051
		[0.776]	[0.442]	[0.029]	[−0.082]	[−0.116]

C. The correlation matrix

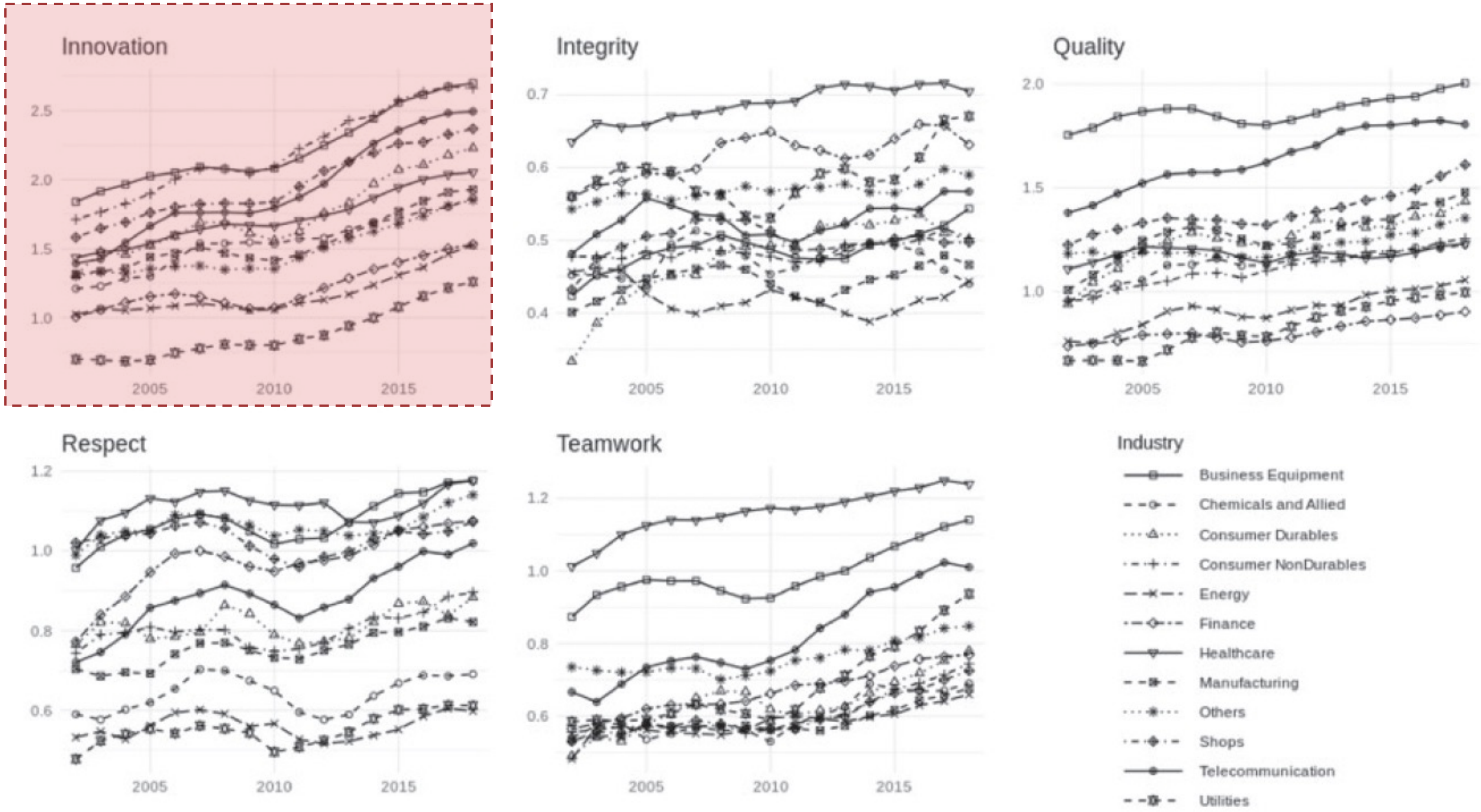
	Innovation	Integrity	Quality	Respect	Teamwork	Firm size	Leverage	ROA	Sales growth	Top-5 institutions
Innovation	1.000									
Integrity	0.109***	1.000								
Quality	0.490***	0.023***	1.000							
Respect	0.321***	0.269***	0.317***	1.000						
Teamwork	0.371***	0.276***	0.271***	0.258***	1.000					
Firm size	−0.186***	−0.010**	−0.261***	−0.255***	−0.309***	1.000				
Leverage	−0.282***	0.024	−0.276***	−0.170***	−0.199***	0.360***	1.000			
ROA	−0.105***	−0.130***	−0.069***	−0.093***	−0.352***	0.403***	−0.035***	1.000		
Sales growth	0.008*	−0.047***	0.017***	0.033***	−0.025***	0.057***	−0.076***	0.222***	1.000	
Top-5 institutions	0.059***	−0.096***	0.018***	0.033***	−0.081***	0.027***	−0.084***	0.145***	0.050***	1.000

Validation

Connor McDowall

Validating Measures of Corporate Control

What do you observe in the figure below?



Validating Measures of Corporate Control

What is an incremental R^2 measure?

Prior literature supports validation variables

- Innovation; ln(Patent), R&D Spending, Innovation Strength
  - Ln(Patent) is the natural logarithm of one plus the number of patents filed and eventually granted in one year
- Integrity (malfeasance in accounting and backdating executive options grants); Restatement, backdating
- Quality; product quality, product safety, top brand
- Respect; diversity, 'best employer'
- Teamwork; employment involvement, number of joint ventures/strategic alliances
- OLS (Ln(Patent), R&D Spending, diversity, number of JV/SA), Probit (Remainder)

A. Validating the cultural value of innovation

	ln(Patent) (1)	ln(Patent) (2)	ln(Patent) (3)
Innovation	0.183***	0.183***	0.098***
	(0.018)	(0.018)	(0.017)
Size	Yes	Yes	Yes
ROA	No	Yes	Yes
Ind FE/yr FE	No	No	Yes
Intercept	Yes	Yes	Yes
Obs.	25,298	25,298	25,298
R^2/pseudoR^2	.036	.036	.166
Incremental R^2	.0301	.0303	.0075
Incremental pseudoR^2			

Author's control for size, ROA, industry, and year effects

Validating Measures of Corporate Control

Can the authors justify the use of their measures?

Justifications loosely support measures

- Authors raise concerns regarding markers testing the corporate measures redundant from high correlations. They address these concerns through the following:
  - Corporate culture could be an aspiration yet to bear fruit in firm policy, performance, with firm culture
  - The markers are much narrower than what the value embodies
  - Data coverage and quality of corporate culture measures are far better than those for most markers
- Use other measures of corporate culture
  - Full transcripts/Glass Door/Topic Modelling
  - What are the issues with these methods?
- Investigate self promotion in calls using measures to detect positive/negative emotions, and word with multiple senses. High correlations in both investigations imply no significant role played

Corporate Finance Applications

Connor McDowall

Implications on Corporate Culture

Hypothesize the practical use of a strong corporate culture?

Authors hypothesize various business outcomes

- Surveys questioning North American CEOs and Chief Financial Officers (CFOs) provide a view corporate culture as one of the top-three factors affecting firm's value, while posturing cultural fit is integral to M&A success
- Authors attempt to empirically examine the implications of having a strong corporate culture on business. They explore:
    - Business Outcomes e.g., Tobin's Q
    - Performance in bad times
    - Mergers & Acquisitions
        - Fit and/or conflict
        - Acquisitiveness
        - Merger pairing
        - Post-merger acculturation

Implications on Corporate Culture

What is Tobin's Q? What does it measure?

	Tobin's q
	(8)
Strong culture$_{(t-1)}$	0.043***
	(0.009)
Firm-level controls	Yes
Ind FE/yr FE	Yes
Intercept	Yes
Obs.	48,750
R^2	.687
Strong culture$_{(t-3)}$	0.048***
	(0.009)
Firm-level controls	Yes
Ind FE/yr FE	Yes
Intercept	Yes
Obs.	36,954
R^2	.712
Strong culture$_{(t-5)}$	0.053***
	(0.010)
Firm-level controls	Yes
Ind FE/yr FE	Yes
Intercept	Yes
Obs.	27,302
R^2	.726

Do these positive correlations make sense?

	Abnormal return (1)	Abnormal return (2)
Strong culture	−0.012***	−0.004
	(0.003)	(0.004)
Strong culture × Financial crisis	0.028***	0.024***
	(0.005)	(0.005)
Strong culture × BP oil spill		
Firm-level controls	Yes	Yes
FF3 factor loadings	Yes	Yes
Yr FE	Yes	Yes
Firm FE	No	Yes
Intercept	Yes	Yes
Obs.	22,092	22,091
R^2	.018	.021

Is this feasible for financial companies?

Strong culture is an indicator variable that takes the value of one if the sum of a firm's five cultural values is in the top quartile across all Compustat firms in a year, and zero otherwise

Implications on Corporate Culture

What makes a successive acquisition?

Acquisitiveness, merger pairing, acculturation

- Authors form the following hypotheses"
  - Cultural fit: Differences in corporate cultures of firm-pairs are a key determinant of deal incidence
  - Acculturation: Predicts merging firms with different cultures will develop a jointly determined culture
  - Apply cultural similarity (cosine) and cultural difference measures (Euclidean distance) to explore hypotheses
- Explore a new dataset of 7,773 completed deals from Jan 1, 2003, to Dec 31, 2018
- Linear probability models (LPM) and Conditional logit models (Clogit) predict acquirers across three subsets - Compustat population, Industry/size matched, Industry/size/BM matched

A. Corporate cultural values and acquisitiveness

	Full sample
	LPM
Variable	(1)
Innovation	0.004**
	(0.002)
Integrity	−0.045***
	(0.005)
Quality	−0.008***
	(0.003)
Respect	0.015***
	(0.002)
Teamwork	−0.000
	(0.003)
Firm size	−0.002**
	(0.001)
Leverage	−0.028***
	(0.008)
ROA	0.137***
	(0.009)
Sales growth	0.054***
	(0.004)
Past return	0.023***
	(0.003)
Top-5 institutions	0.169***
	(0.011)
Ind FE/yr FE	Yes
Deal FE	No
Intercept	Yes
Obs.	53,545
R^2/pseudo R^2	.047

Assesses the probability of being an acquirer

What do the negative coefficients imply?

Implications on Corporate Culture

Is there sufficient evidence to support both hypotheses?

B. Cultural fit and merger pairing

Variable	Industry and size-matched Clogit (1)	Clogit (2)
Cultural similarity	4.305***	
	(0.902)	
Cultural distance		−0.496***
		(0.075)
Acquirer characteristics		
Firm size	2.634***	2.680***
	(0.210)	(0.210)
Leverage	−1.062***	−1.153***
	(0.342)	(0.350)
ROA	−0.077	−0.223
	(0.566)	(0.581)
Sales growth	0.355**	0.398**
	(0.169)	(0.168)
Past return	0.164	0.153
	(0.142)	(0.147)
Top-5 institutions	1.645***	1.665***
	(0.442)	(0.432)
Target characteristics		
Firm size	2.090***	2.064***
	(0.299)	(0.300)
Leverage	0.062	−0.113
	(0.307)	(0.307)
ROA	−0.585*	−0.605**
	(0.308)	(0.306)
Sales growth	0.321**	0.323**
	(0.141)	(0.141)
Past return	−0.053	−0.035
	(0.092)	(0.095)
Top-5 institutions	2.783***	2.818***
	(0.379)	(0.381)
Deal characteristics		
Same state	0.928***	0.925***
	(0.147)	(0.148)
HP similarity	26.551***	26.661***
	(2.058)	(2.035)
Deal FE	Yes	Yes
Obs.	5,682	5,682
Pseudo R^2	.295	.300

Cultural similarity examines the relation between cultural fit and acquirer-target firm pairing (binary;1,0), estimated from 594 completed deals.

Acculturation after deal completion, using OLS regressions, for one and three years after the deal, without engaging in another significant deal, using 492 and 335 completed deals, respectively. Target-specific values regressed on acquirer values in the year prior to deal announcement

C. Post-merger acculturation

	Innovation$_{t+1}$ (1)	Innovation$_{t+3}$ (2)	Integrity$_{t+1}$ (3)	Integrity$_{t+3}$ (4)	Quality$_{t+1}$ (5)	Quality$_{t+3}$ (6)	Respect$_{t+1}$ (7)	Respect$_{t+3}$ (8)	Teamwork$_{t+1}$ (9)	Teamwork$_{t+3}$ (10)
Acquirer innovation	0.854***	0.905***	0.030**	0.042**	0.026	0.059	0.049*	0.043	0.035*	0.072***
	(0.039)	(0.053)	(0.014)	(0.020)	(0.028)	(0.041)	(0.027)	(0.036)	(0.021)	(0.025)
Target-specific innovation	0.108***	0.108**	0.003	0.022	−0.010	−0.049	−0.022	−0.028	−0.002	−0.014
	(0.034)	(0.052)	(0.014)	(0.021)	(0.025)	(0.038)	(0.023)	(0.033)	(0.018)	(0.027)
Acquirer integrity	0.027	−0.050	0.552***	0.506***	−0.038	−0.077	0.073	0.026	0.043	−0.047
	(0.107)	(0.161)	(0.051)	(0.063)	(0.077)	(0.101)	(0.073)	(0.096)	(0.063)	(0.079)
Target-specific integrity	−0.038	−0.043	0.069*	0.112*	−0.002	0.040	0.074	0.065	0.045	0.067
	(0.086)	(0.132)	(0.041)	(0.058)	(0.070)	(0.100)	(0.061)	(0.101)	(0.048)	(0.068)
Acquirer quality	0.074	0.067	0.041*	0.044	0.841***	0.790***	0.077***	0.108**	0.073***	0.090**
	(0.048)	(0.083)	(0.022)	(0.033)	(0.032)	(0.048)	(0.029)	(0.053)	(0.026)	(0.037)
Target-specific quality	−0.001	0.064	−0.008	−0.003	0.099***	0.154***	−0.034	−0.034	0.015	0.027
	(0.035)	(0.052)	(0.016)	(0.023)	(0.028)	(0.041)	(0.026)	(0.037)	(0.020)	(0.030)
Acquirer respect	−0.104**	−0.196**	0.002	−0.001	0.035	0.014	0.766***	0.685***	−0.013	0.006
	(0.052)	(0.077)	(0.022)	(0.033)	(0.044)	(0.060)	(0.040)	(0.064)	(0.033)	(0.044)
Target-specific respect	0.085**	−0.012	−0.036*	−0.068**	0.024	0.044	0.094***	0.092**	0.025	−0.039
	(0.043)	(0.064)	(0.021)	(0.031)	(0.033)	(0.053)	(0.029)	(0.046)	(0.023)	(0.033)
Acquirer teamwork	0.064	0.079	−0.003	−0.038	−0.025	−0.036	−0.013	−0.071	0.684***	0.562***
	(0.076)	(0.105)	(0.031)	(0.042)	(0.044)	(0.067)	(0.051)	(0.082)	(0.042)	(0.049)
Target-specific teamwork	−0.071	−0.135*	0.029	−0.020	−0.001	−0.001	−0.016	−0.054	0.081***	0.200***
	(0.044)	(0.069)	(0.021)	(0.032)	(0.034)	(0.058)	(0.031)	(0.051)	(0.025)	(0.044)
Acquirer/target/deal controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ind FE/yr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	492	335	492	335	492	335	492	335	492	335
R^2	.806	.780	.538	.472	.807	.761	.746	.707	.717	.679

Concluding Comments

Connor McDowall

Conclusion

Can you measure corporate culture? Does it matter?

ML is useful for measuring corporate culture

- Introduce the word embedding model as a new approach to quantifying the meaning of expressions
- Propose a new semi-supervised machine learning approach for textual analysis to reap benefits from supervised and unsupervised
- Obtain scores for five corporate culture values: innovation, integrity, quality, respect, and teamwork
- Validate measures and attempt to correlate corporate culture to business outcomes, M&A Activity
- Machine learning holds promise for more applications in social science

Strengths & Weaknesses

What are additional strengths and weaknesses?

Paper has several strengths and weaknesses

- Strengths
  - Comprehensive dataset
  - Novel methodology to measures semantics within documentation
- Weaknesses
  - Validation tests not too thorough
  - Inconsistencies when applying corporate culture measures to business outcomes i.e., business performance
  - Industry/fixed effects explain changes in scores
  - Misalignment between autocorrelations and business performance

Literature Review & Future Research

There has been a limited number of articles implementing Natural Language Processing (NLP) algorithms to measure corporate culture

1. **Corporate Culture**
   O'Reilly, Chatman 1996; Graham et al, 2018
2. **Textual Analysis**
   Loughran, MacDonald 2016
3. **Collocations and Corporate Disclosures**
   Routledge, Sacchetto, Smith 2018
4. **Word Embedding Models**
   Harris, 1954
5. **Relationship between Culture and M&A**
   Graham et al., 2018
6. **Empirical Asset Pricing via Machine Learning**
   Shihao Gu, Bryan Kelly, Dacheng Xiu, 2020

Further research

1. Use methodology to measure sentiment in other areas e.g., Attitudes towards ESG
2. Improve evaluation of business performance, M&A activity
3. Make comparisons between organizational hierarchies to evaluate cultural adoption

Thank you