**SEEDLINK LEADERSHIP RESEARCH PROJECT**

**A computational stylometry approach to assessing leadership traits in human languages**

**Abstract:**

Variations in natural language correlate significantly with aspects of personality (Fast and Funder, 2008; Gill et al., 2009; Yarkoni, 2010; and Pennebaker and King, 1999). Computational Stylometry is a technique in Natural Language Processing that allows the study of communication styles of individuals based on the linguistic analysis of the text. The communication styles help us study additional characteristics, such as the personality traits of individuals. These characteristics are important for assessing leadership traits because CEOs personality and communication style correlate with company’s performance (Choudhury, Prithwiraj and Wang, Dan and Carlson, Natalie and Khanna, Tarun, 2019). Therefore, the goal of this research is to understand what language features characterize successful leaders.

**Why Computational Stylometry?**

The approach of Computational Stylometry in our research is inspired by the challenge of author attribution. In Author attribution, techniques in computational stylometry helps to study certain traits of the author, such as personality and writing style to identify author’s identity (Daelemans). Similarly, in this research, we aim to study the subconscious markers in the natural language of CEOs that can help explain the variability in their leadership effectiveness

Transcripts of earnings conference calls enable researchers to study language of CEOs which corresponds to their traits and behavior, e.g. optimism (Davis et al., 2015), personality (Gow et al., 2016), deception (Burgoon et al., 2016), and managerial style (Davis et al., 2012). Our Approach to Computational Stylometry involves rigorous text analysis of the language of S&P 500 companies CEOs during quarterly earnings conference calls. The Text analysis includes Parts of Speech (P.O.S) tags nGrams, Function Stop words nGrams, language complexity measures, vocabulary richness, and lexical & stylometric features. We combine our text analysis of CEOs language with Machine Learning Algorithms to learn about the correlations between linguistic features and company’s performance

**Data Source:**

1. **Earning Conference Calls**

The earnings conference calls transcripts were obtained using web scrapping techniques in python. A total of 22,508 conference calls transcripts of S&P 500 companies were collected containing the language of acting CEO, CFO, Employee, and Analysts. These conference calls transcripts were generated from the following two sources:

1) Seeking Alpha (<https://seekingalpha.com/>)

2) Common Crawl (<http://commoncrawl.org/>)

1. **Companies Stock Price Data**

The stock data for these S&P 500 companies was collected to cover the company's performance during the CEOs tenure. The companies showing abnormal variations in the stock prices were eliminated from the dataset. This Financial Data was obtained through the following source:

1) Intrinio (<https://intrinio.com/>)

**METHODOLOGY:**

We approach our research problem in a supervised machine learning manner. In supervised machine learning, the machine learns from the input data using the ground truth values. In our research, we use the CEOs ranking (defined below) as these ground truth values and CEOs language during conference calls as the input data. We then compare the linguistic features of CEOs language during conference calls against the CEOs ranking, and therefore try to study what linguistic features correlate significantly with CEO ranking and therefore effective leadership traits.

**CEO Ranking and its computation**

The CEOs Ranking is a measure of the performance of the CEO’s firm over his/her complete tenure. CEOs performing consistently well have a greater CEO ranking compared to CEOs performing well over a brief period.

All the S&P 500 companies were categorized into their industrial sectors to analyze the industry growth rate per year of each industry. Afterwards, each company stock price was corrected for its respective industry. This was done to analyze company outperformance or underperformance by industry standards. The CEOs were ranked using the yearly growth rate of industry corrected stock prices during their tenure. The computation for this growth rate is described below:

The expected adjusted stock price of a company *pt* is frequently modeled to grow exponentially over time:



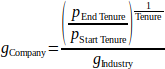
Where *pt=0* is the adjusted stock price in an arbitrary year, *t* is the number of years since *t*=0 and *g* is the annual growth rate, for example *g*=1.05 indicates that the adjusted stock price is growing by five percent year over year.

We model the growth rate of the adjusted stock price as the product of industry and company effects, i.e.:

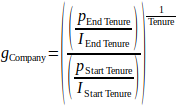


The industry-adjusted growth rate during the CEO tenure is then estimated as:









**Data Pre-Processing:**

The language of the CEOs required pre-processing because it was obtained as the transcripts, but not as written responses. Therefore, the use of commas, semicolons and other alpha numeric character was not an accurate representation of the CEOs language structure and therefore they were removed.

**Feature Extraction from CEOs Language**

The goal of the research project is to extract language features of Firm Leaders (CEOs) that characterize successful leaders. Our approach is to extract transferable features from earning conference calls which can be applied across the language of the CEO, such as an interview. Here are some of the features we extracted:

***Stylometric Features***

Stylometric features allow us to study the structure of the language. An analysis of the Stylometric features is inspired by author attribution- identifying the author based on text. These features are transferable across the language in different context and allow us to identify key structures that separate the speaking/writing style of an individual from the other. We aim to solve a classification problem in this research and try to find language patterns that are common within successful and unsuccessful people respectively. Some of the Stylometric Features we extracted were:

* *Adjective Rate:* The number of adjectives per sentence.
* *Adverb Rate:* The number of adverbs per sentence.
* *Syllable Rate*: The average number of syllables per sentence.
* *Sentence Length:* The average length of each sentence.
* *Function Word Rate:* The average number of function words per sentence.

***Lexical Features***

Lexical Features allow us to analyze the common patterns in the language. We extracted the following lexical features:

* *Unigrams:* DistinctParts of Speech (POS) and Function Stop words that occur across CEO language.
* *Bigrams:* A group of two Parts of Speech (POS) and Function Stop words that occur sequentially across the CEO language corpus.
* *Trigrams:* A group of three Parts of Speech (POS) and Function Stop words that occur sequentially across the CEO language corpus.
* Pronoun Count: A term frequency of Singular and Plural pronouns across CEOs language.

***Readability Scores***

The readability measure is the ease with which the communicator conveys his/her ideas. Research has shown that the communication style of the CEO has a significant impact on the firm performance. (Insert Reference here). Therefore, Calculating the Readability scores were crucial to understanding the communication style of the CEOs. The Readability scores measure from the text are as follows:

* Flesch Reading Ease
* Flesch-Kincaid Grade Level
* Gunning Fog Index
* Dale Chall Readability Formula
* Shannon Entropy
* Simpson's Index
* Smog Index
* Coleman Liau Index

***Vocabulary Richness***

* Hapax Legomenon
* Hapax DisLegemena
* Honores R Measure
* Sichel’s Measure
* Brunets Measure W
* Yules Characteristic K
* Shannon Entropy
* Simpson’s Index

***Term frequency Inverse Document Frequency (Tdidf) word ngram Vectorizer:***

We analyze the Parts of Speech (POS) Tags and Function Stop words ngrams using the Tfidf Vectorizer from sklearn library. The Tfidf Vectorizer converts the word representations into number representation based on the term frequency of the Parts of Speech (POS) ngrams and function stopwords ngrams in our corpus text.

**Machine Learning Models**

The research problem was solved as a classification problem. We categorized CEOs outperforming and underperforming over their tenure into two separate classes and analyzed their language during conference calls against the class category. A total of over 100,000 features were extracted and used to train the classifier. A balanced dataset with the language of 350 CEOs with outperformance and 350 CEOs with underperformance were included in the dataset.

The classification models implemented were the following:

1) Support Vector Classifier (Sklearn Library)

2) XGBoost Classifier (Sklearn Library)

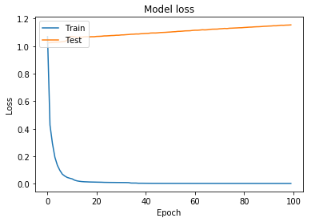
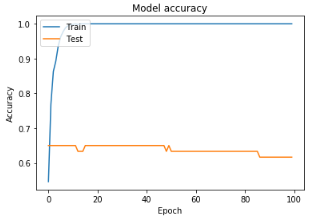
3) Multinomial Naïve Bayes (Sklearn Library)

4) Dense Neural Network (Keras Library)

**RESULTS:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Cross Validation Accuracy** | **Test Accuracy** |
| **SVC** | **63.5** | **67.5** |
| **XGBoost** | 60.5 | 63.5 |
| **Naive Bayes** | 58.3 | 59.2 |
| **Neural Network** | 60.5 | 62.5 |

**Validation and Training Loss with Dense Neural Network:**



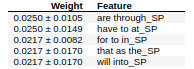
The Model Loss and Accuracy graphs explain that the Neural network largely overfit the training data and therefore did not perform well on the test data. Given the nature of our problem, we learned that a dense neural network would almost always overfit the training data unless the number of features were significantly reduced.

**Hyper-parameter tuning on tfidf word ngram and Machine Learning Models (Improving Classifier’s Accuracy):**

We created pipelines containing the tfidf word ngram model and the Machine Learning Classifier to apply hyper-parameter tuning on the entire framework. We used GridSearchCV to perform fine tuning of our parameters, such as maximum number of features, C-penalty in SVC, and n-gram range. The GridSearchCV tries all different combinations of the parameters provided and selects the parameters which provide the best accuracy of the model.

**Permutation Importance to explain Feature Importance**

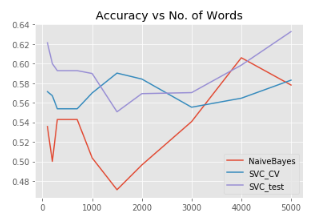
After we have computed the features and obtained an accuracy measure from the machine learning model, we need to measure the feature importance. We measure the feature importance by a technique called Permutation Importance ([Link Here](https://www.kaggle.com/dansbecker/permutation-importance)) . Provided several features computed from the language of the CEO, permutation importance randomly shuffles the values of a single feature across the CEOs and make prediction using the remaining features. We then use these predictions and the true target values to calculate how much loss the function suffered from shuffling. This performance deterioration measures the importance of the feature we shuffled. We then restore the order of the shuffled feature and perform permutation importance over the next feature. Here are the results we obtained using permutation importance. The Weight corresponding to each feature represents the percentage in the total accuracy that feature accounts for. Here’s an example:



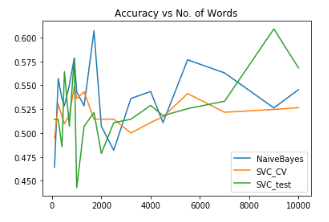
**Number of Words Vs Accuracy Visualization**

The Classifier was trained separately on Function Stop words nGrams, Parts of Speech (POS) nGrams, and on the Lexical features for the word range 100-5000. Here are the results:

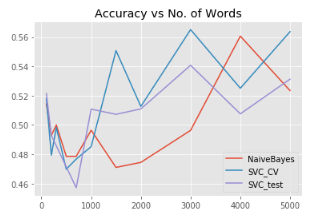
**Function Stop words nGrams accuracy**



**Parts of Speech (POS) nGrams Accuracy**



**Lexical Features Accuracy**



**LIMITATIONS AND CHALLENGES**

The main challenge of our research is that we study language features which are transferable across the CEOs language regardless of the context. Therefore, analyzing the content and meaning of CEOs language during conference calls does not help our case because we are more interested in the structure of language. This is the main challenge is to study the language features to identify correlations between CEOs language features and company’s performance. Secondly, it was also challenging to provide a strong explanation for certain nGrams of POS tags and Function Stop words to have strong predictive power of leaders' success.

**FUTURE WORK**

In future, we would like to study the overlaps of predictive features between successful and unsuccessful CEOs. This will help us develop an explanation for the predictive power of these features. Moreover, more techniques in author attributions that provide direct motivation to our research problem could be tested to further improve the accuracy of the classifier.

**References**

Daelemans, Walter. "Explanation in computational stylometry." *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Berlin, Heidelberg, 2013

Choudhury, Prithwiraj, et al. "Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles." *Available at SSRN 3392448* (2019)

Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., ... & Spitzley, L. (2016). Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. *Journal of Language and Social Psychology*, *35*(2), 123-157.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2012). The effect of managerial “style” on the tone of earnings conference calls. In *CAAA Annual Conference. Retrieved from* [*http://www*](http://www/)*. usc. edu/schools/business/FBE/seminars/papers/ARF\_9-21-12\_GE. pdf*.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, *20*(2), 639-673.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, *94*(2), 334.

Gill, A. J., Nowson, S., & Oberlander, J. (2009, March). What are they blogging about? Personality, topic and motivation in blogs. In *Third International AAAI Conference on Weblogs and Social Media*.

Gow, I. D., Kaplan, S. N., Larcker, D. F., & Zakolyukina, A. A. (2016). *CEO personality and firm policies* (No. w22435). National Bureau of Economic Research.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, *44*(3), 363-373.