**CEO’s Twitter behaviour and its impact**

* *Introduction to our study*
* *CEOs tweets Data Extraction from Twitter*
* *Data Processing*
* *Feature Engineering*
* *Statistical Analysis of variation of company performance with our extracted features*
* *Conclusion and further developments*

*#Introduction portion not finished. Need your help !*

Social capital of a CEO and his natural behaviour

Characteristics of CEO that impact our target variable under study, for e.g Ethics impact Financial statements/restatements as fraudulent ..

And how does this behaviour impacts multiple company variables.

How ethical the CEO is as a person? Etc..

Used ExecComp and CompuStat data and extracted tweets of 3860 CEOs of 2155 companies during time period of 2007-2018, using Python tweet extraction repository by **Jefferson Henrique**, though this kind of extraction is completely blocked now. Twitter API can be used instead.

Tweets extracted using CEO name and 5-yr time period(2014-2018).

too sparse which may bias the estimation

of time trend effect.

Variables extracted (for each tweet)–->

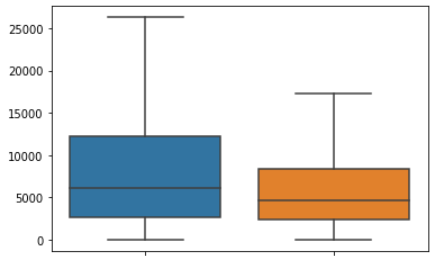
* Tweet text
* Retweets on that tweet
* Time and date
* Mentions and hashtags

Now, this data enabled to generate many basic insights along our study

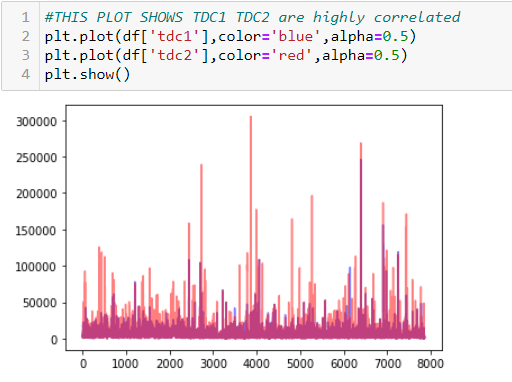
* When time line reduced down to 5 yrs

|  |  |
| --- | --- |
| Total number of CEOs | 2421 |
| On twitter? Yes (10 %) | 240 |
| no | 2181 |

* Mean total compensation of CEOs on twitter > Mean total compensation of CEOs not on twitter
  + Suggests that better CEOs believe in social media presence

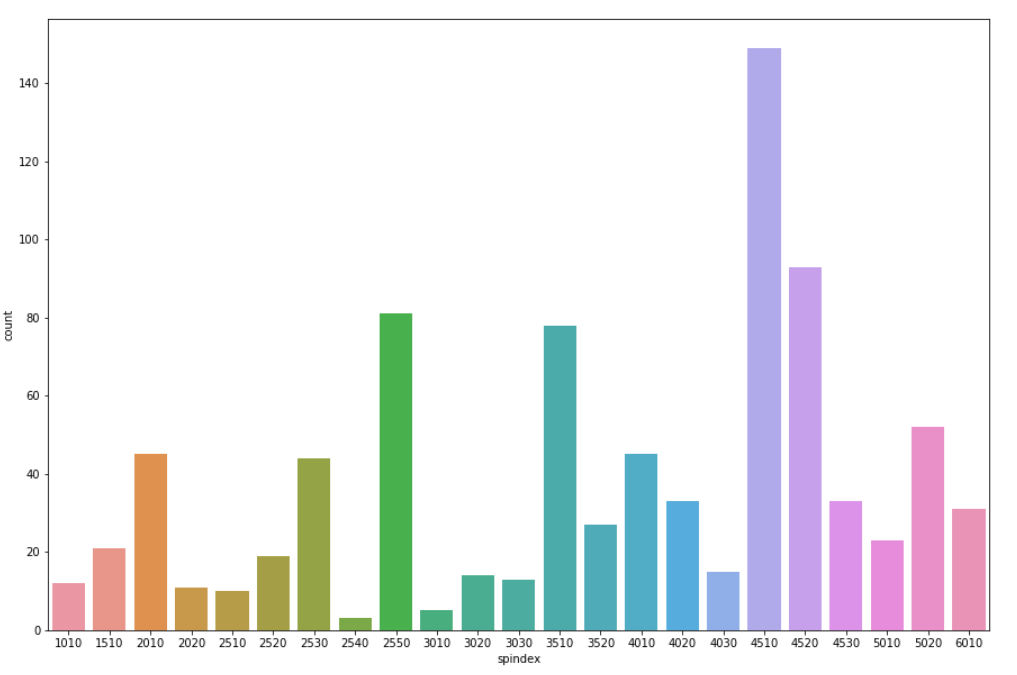


* High correlation of TDC1 and TDC2 – corr = 0.7688567372912828 as expected



*Purple colour shows overlap*

* Social media presence of CEOs for different industries classified on SPindex [firm-year observations]
  + ***Software and Services*** Sector has the maximum share of CEOs on twitter



**Sentiment Analysis part --** [**Colab notebook**](https://colab.research.google.com/drive/1JqNW-mHzEHhoGn3vozyTusdwTF0wFMqe)

done basic cleaning of our twitter corpus (1 tweet as a complete document or can say data point) and deployed the Transformers library which has the pre-trained state-of-the-art NLP models available for different use cases, our use case being sentiment analysis.

[Transformer Library Reference](https://huggingface.co/transformers/)

The sentiment scores obtained from this transformers library are taken to be pretty accurate cause those model are trained over humongous data and have proven results.

We expect this sentiment score to be key variable impacting our target variable during our model development. Also the extracted context can be somewhat justified using this sentiment score evaluated. Say positive tones like – motivation or happy must have +ve sentiment score

# ****Tones & Contexts score part --**** [Colab notebook](https://colab.research.google.com/drive/1j3JUqbDGsh2T7hOLQV2aiV2QUcvOesUM)

\*one issue as discussed before was the selection of words for representing different contexts and tones. For once I have taken from Merriam Webster thesaurus considering key tone and contextual categories based on personal understanding and somewhat web re-searching.

Now for justification of this selection, two things can be done –

* Evaluation and further development in categories selection by any crowd sourcing website ([Amazon mturk](https://www.mturk.com/))
* P-value analysis and Correlation study of our results and selected categories, which is kind of a verification task.

For this part of our study, we do the following tasks over our corpus🡪

1. Preprocessing of our twitter corpus
2. tokenizing
3. incorporating word embeddings (the contextual ones) using 'allennlp' module

one vector is generated for every word, which will be further aggregated via PCA and respresented by the principal component, for further calculation. This principal component aggregation done both for data points as well as contexts and tone set of words.

And we calculate cosine similarity and euclidean distance of these vectors with embedding vectors of different tones and contexts (obtained in same way as for tweets) in a 1024 dimensional space of ELMo vectors.

The sentiment scores raised to power three to make the distribution sparse.

And cosine similarity scores reduced to square root, to reduce centricity of scores.

\*After all these data engineering and feature extraction operations, some data consistency checks (basic SQL operations) were done with respect to dates and companies etc, because multiple independent datasets were used.

After all this functions done and final dataset compiled, multi linear regression was performed, relating our input variables and ROA as target.

By now, our data points reduced to 477 data points only, over which we need to perform our analysis.

We have our one data point as –

Scores of a particular CEO over a single financial year defined by that company.

Now I was planning to approach the problem in this manner

* Taking all data points and analysing impact of our scores on target
* With some initial values, taken from above analysis, forecast the target variable for a company over specified time period and CEO social media variables.

BUT, due to lack of data, and data having no structure at all starting from the beginning, it is impossible to develop a reasonable predictive model for our required target variables.

Although there is still lot of scope for improvement in our language model and feature generation.

* Could have used LIWC and other packed repos for context use cases
* Use of ***Sentiwordnet*** for sentiment analysis etc.

Also, a customized BERT model can be trained for our tasks, if we scale more data.

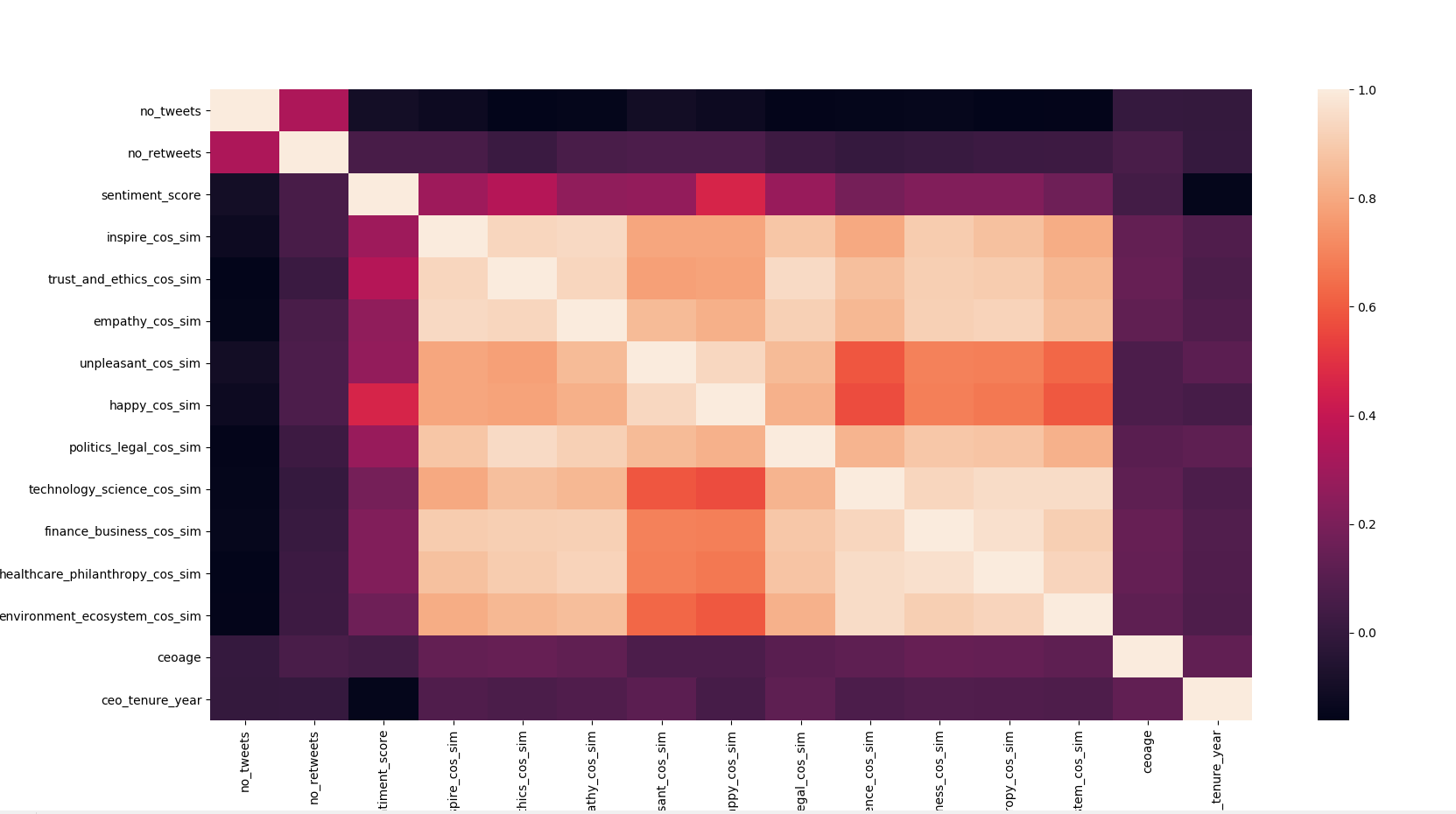
And given we acquire a good amount of structured data, the above mentioned approach will work just fine for any target variable we choose.

Also there could be more input features we can think of, but are some that I used,

* just to calculate P-values of different features using Linear regression
* and feature importance using XGboost model

This is a normal probability plot of residuals of predictions from our model. Apart from a very few points, the residuals seem to follow normal distribution. The reason behind this is that our data is quite unstructured.

One of the problems come from this high-correlation in our similarity features generated.



Anyways, the P-value analysis gives some helpful insights for observing how different features impact our model for different target variables.

Here is a glimpse of P-value analysis (95% significance level)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.no | Feature\target variable | lead\_tob\_q | lead\_roa\_ebit | lead\_roa | emp\_turnover |
| 1 | no\_tweets |  |  |  |  |
| 2 | no\_retweets |
| 3 | sentiment\_score |
| 4 | inspire\_cos\_sim |
| 5 | trust\_and\_ethics\_cos\_sim |
| 6 | empathy\_cos\_sim |
| 7 | unpleasant\_cos\_sim |
| 8 | happy\_cos\_sim |
| 9 | politics\_legal\_cos\_sim |
| 10 | technology\_science\_cos\_sim |
| 11 | finance\_business\_cos\_sim |
| 12 | healthcare\_philanthropy\_cos\_sim |
| 13 | environment\_ecosystem\_cos\_sim |
| 14 | ceoage |
| 15 | ceo\_tenure\_year |

For different target variables (picked from data shared by Sir) –

* Our model gives importance to different features as target variables change, therefore signifying the importance of extracted features, these features seem to perform well

*This is progress by far, need to do further developments – after scope discussion with Sir.*