

[Home](#) (<http://blog.nycdatascience.com/>) [Featured](#) (<http://blog.nycdatascience.com/category/featured/>)  
[Faculty](#) (<http://blog.nycdatascience.com/category/faculty/>)  
[Meetup](#) (<http://blog.nycdatascience.com/category/meetup/>)  
[Students' Work](#) (<http://blog.nycdatascience.com/category/student-works/>)  
[Data Science News and Sharing](#) (<http://blog.nycdatascience.com/category/data-science-news-and-sharing/>)

[Signup](#)[/student-signup/](#) [Login](#) (/wp-login.php)  
**SUBSCRIBE TO OUR NEWS**Sign up to our newsletter for up  
exclusive promotions. First Name Last Name Email[Subscribe](#)

## Browse

- [Home](#) (<http://blog.nycdatascience.com/>)
- [R Visualization](#) (176 posts) (<http://blog.nycdatascience.com/categories/r-visualization/>)
- [R Shiny](#) (161 posts) (<http://blog.nycdatascience.com/categories/r-shiny/>)
- [Web Scraping](#) (135 posts) (<http://blog.nycdatascience.com/categories/web-scraping/>)
- [Machine Learning](#) (87 posts) (<http://blog.nycdatascience.com/categories/machine-learning/>)
- [Capstone](#) (54 posts) (<http://blog.nycdatascience.com/categories/capstone/>)
- [Faculty](#) (43 posts) (<http://blog.nycdatascience.com/categories/faculty/>)
- [Meetup](#) (126 posts) (<http://blog.nycdatascience.com/categories/meetup/>)
  
- [r-bloggers.com](#) (<http://www.r-bloggers.com/>)
- [Blog Post Guide](#) (<http://blog.nycdatascience.com/work/blog-post-guide/>)
- [How to Add a Blog Post With Shortcode](#) (<http://blog.nycdatascience.com/works/how-to-add-blog-with-shortcode/>)

# Predicting Profit Warnings: NLP applied to Conference Call Transcripts Analysis

Published by [Carlos Salas Najera](#) ([http://blog.nycdatascience.com/author/carrossn/](http://blog.nycdatascience.com/author/carlossn/)) at March 31, 2017

The objective of the following research is to introduce the reader into the reader into NLP (Natural Language Processing) and how it can be used to predict profit warning risk. "Profit warning" is defined as an event in which a company advises its earnings will not meet analyst expectations. The profit warning event triggers a rather acute stock correction that can be exploited by investors able to identify industries and companies about to trim their quarterly/yearly earnings guidance.

NLP literature has grown notably over the last couple of years covering mainly sentiment analysis of SEC reports and conference calls transcripts. However, there is a significant scarcity of papers based on NLP and corporate events like Profit Warnings. Particularly, this research seeks to yield some answers to the some of the next questions:

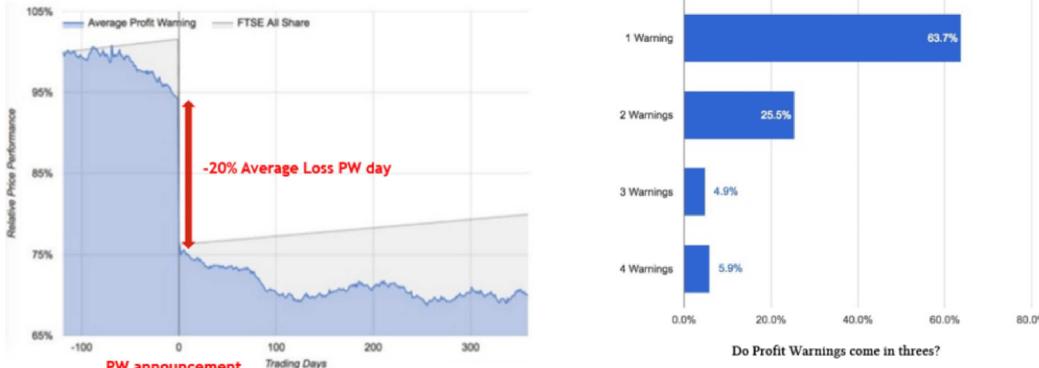
- *Are NLP tools useful to help equity research analysts?*
- *Do conference call transcripts contain additional information?*
- *Can we build tools that help flag profit warning risk?*
- *Can we predict Profit Warnings using NLP tools?*

Successful adoption of NLP tools could boost the productivity of the average equity research analyst both in the sell-side and buy-side. Currently a significant amount of resources are wasted in the investment community as the average analyst is ordered to read, digest and analyze manually more than 40 company's conference call transcripts per quarter in order to extract useful investing insights from them. The fall of the sell-side equity research business, as well as the underperformance of pure fundamental investment strategies in the buy-side, is enough evidence pointing towards the urgent need by active investment managers of embracing more innovative ways of processing data in order to cope with the pace imposed by quantitative strategies and passive "smart-beta" instruments.

## Dissecting Profit Warnings

Profit warning is defined as an event in which a company advises its earnings will not meet analyst expectations. A profit warning is usually announced two or more weeks before an earnings announcement. The reader can acquire a quick understanding of the consequences of a profit warning looking at bullet points and the couple charts below. The charts shows the average return from an analysis conducted in 2016 using 245 profit warnings between January 2013 and early August 2016 in the UK stock market (source: "The Profit Warning Survival Guide"):

- *Pre-event weak price momentum:* On average, prices begin falling by 6% in the 6 months before the warning, underperforming the FTSE-All-Index by 7.6%.
- *Sudden event price adjustment:* The average price decline on the day of the profit warning was -19.2%.
- *Sticky negative sentiment:* A noticeable further decline followed for two-to-three months after the warning, possibly coinciding with further earnings news.
- *Persistent flattish performance:* Over 12 months later there was, on average, no significant reversal in the price decline. 44.5% of stocks lost an additional 10% or more from the day after the profit warning until 1 year later.
- *Standalone or Multiple Profit Warnings?* More than 1 in 3 profit warnings is likely to be followed by another, which may explain some of the ensuing post-announcement underperformance.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/PW.png>)

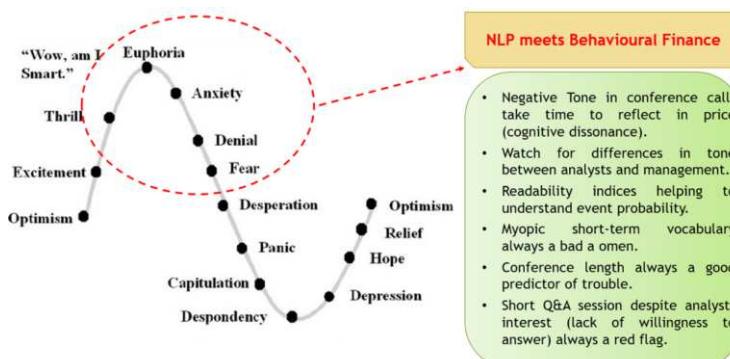
The table below shows different practitioner and scholar papers with regards profit warnings. Different authors have found over the last two decades to find patterns around profit warning announcements that disproved perfect capital market theories. For instance, Kearns and Whitley (2005) find that 75% of the profit warning stocks experience further margin deterioration during the year of the announcement and are more prone to increase financial leverage, trim CapEx (Capital Expenditure) and reduce their dividend payout. Chang and Watson (2007) point out that small caps underperformance in profit warning events is even more acute and that insiders (management, largest investors) sell before the event. Aubert and Louhichi (2009) find that profit warning stocks experience abnormally high volatility and trading volumes around the event date.

Author	Year	Title
George Bulkley, Renata Herreras	2002	Stock Returns Following Profit Warnings
Helbok, Walker	2003	On the Willingness of UK Companies to Issue Profit Warnings
Kearns, Whitley	2005	The balance-sheet information content of UK company profit warnings
Clare	2005	The information in UK company profit warnings
Chang, Watson	2007	What are insider trades around profit warnings really telling us?
Francoeur, Labelle, Martinez	2008	Governance and the decision of issuing a Profit Warning
Aubert, Louhichi	2009	Profit warnings and abnormal stock returns: the French case
Tummurthu, Wang, Lions	2010	The relationship between the profit warning and stock returns: Empirical evidence in EU markets
Naliaka	2011	Effect of Profit Warnings on Share Price of Companies listed in the Nairobi Stock Exchange
Heesters	2011	Stock Returns Following Profit Warnings: Evidence from the Dutch Stock Market
Sphor	2015	Can Analysts Forecast Profit Warnings?
Page, Hobson, Naamani	2016	The Profit Warning Survival Guide

(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/PWRes-1.png>)

## NLP: Loving the Alien

NLP (Natural Language Processing) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages and, in particular, concerned with programming computers to fruitfully process large natural language corpora. NLP research is particularly scarce, especially in Finance, with the bulk of literature having been released over the last five years.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/BFNL.P.png>)

## Tags

Big Data

(<http://blog.nycdatascience.com/big-data/>)

D3.js (<http://blog.nycdatascience.com/d3.js/>)

data science

(<http://blog.nycdatascience.com/data-science/>)

data visualization

(<http://blog.nycdatascience.com/data-visualization-2/>)

Demo Day

(<http://blog.nycdatascience.com/demo-day/>)

gbm

(<http://blog.nycdatascience.com/gbm/>)

ggplot2

(<http://blog.nycdatascience.com/ggplot2/>)

googleVis

(<http://blog.nycdatascience.com/googleVis/>)

Hadoop

(<http://blog.nycdatascience.com/hadoop/>)

Kaggle

(<http://blog.nycdatascience.com/kaggle/>)

Logistic Regression

(<http://blog.nycdatascience.com/logistic-regression/>)

machine learning

(<http://blog.nycdatascience.com/machine-learning/>)

Maps

(<http://blog.nycdatascience.com/maps/>)

neural network

(<http://blog.nycdatascience.com/neural-network/>)

Neural networks

(<http://blog.nycdatascience.com/neural-networks/>)

nlp (<http://blog.nycdatascience.com/nlp/>)

The table below shows selected papers that help to prepare our Profit Warning NLP model. Among the most important findings, Chen, De, Hu and Hwang (2014) demonstrate that the frequency of negative Words compiled by Loughran and McDonald (2011) is a good predictor of strong or weak earnings. Matsumoto, Pronk and Roelofsen (2006) highlight that the length of the conference call transcript, especially its Q&A part, also plays an important role setting the earnings tone. McKay Price, Doran, Peterson, Bliss (2011) confirm that Q&A Management textual tone is important, especially for non-dividend paying companies. Zhou (2014) points out that underperforming stocks' CEOs are more prone to blame external factors during their conference call speech. Bushee, Gow, Taylor (2016) underscore that stock prices reflect good conference call tones sooner than negative tones, which is in line with behavioural finance reported biases such as cognitive dissonance. Last but not least, Borochina, Ciconi, Delisle and McKay Price (2017) find that management-Analysts tone differences sparked uncertainty, which is important to explain stock underperformance. The table below contains these and other NLP-related research papers used in the development of the Profit Warning NLP model:

Author	Year	Title	Topic
Bowen, Davis, Matsumoto	2000	Do Conference Calls Affect Analysts' Forecasts?	Conference Calls
Bushee, Matsumoto, Miller	2001	Open versus Closed Conference Calls	Conference Calls
Kohlbeck, Magill	2002	The Impact of Concurrent Conference Calls on the Information Content of Earnings Announcements	Conference Calls
Brown, Hillegeist, Lo	2003	Conference Calls and Information Asymmetry	Conference Calls
Matsumoto, Pronk, Roelofsen	2006	Length, Discourse, and Tone	NLP / Annual Reports
Li	2006	Annual report readability, current earnings, and earnings persistence	NLP / SVM / Annual Reports
Holander, Pronk, Roelofsen	2008	Does Silence Speak? An Empirical Analysis of Disclosure Choices during Conference Calls	Conference Calls
Kogan, Rosedgård, Lewis, Sagi, Smith	2009	Predicting Risk from Financial Reports with Regressions	NLP / Conference Calls
Doran, Peterson, McKay Price	2010	Earnings Conference Call Content and Stock Price: The Case of REITs	NLP / Conference Calls
Larcker, Zakołyska	2011	Detecting Deceptive Discussions in Conference Calls	NLP / Conference Calls
Mayew, Sharp, Venkatachalam	2011	Using Earnings Conference Calls to Identify Analysts with Superior Private Information	NLP / Conference Calls
Bassamir, Farkas, Pachta	2011	The Effect of Conference Calls on Analysts' Forecasts - German Evidence	NLP / Conference Calls
Mayew, Venkatachalam	2011	The Power of Voice: Managerial Affective States and Future Firm Performance	NLP / Conference Calls
McKay Price, Doran, Peterson, Bliss	2011	Earnings conference calls and stock returns: The incremental informativeness of textual tone	NLP / Conference Calls
Loughran, McDonald	2011	Measuring Readability in Financial Disclosures	NLP / Annual Reports
Blau, Delisle, McKay Price	2012	do Investors Interpret Earnings Conference Call Tone Differently than Investors at Large? Evidence from The Effect of Manager-specific Optimism on the Tone of Earnings Conference Calls	NLP / Conference Calls/Shorts
Davis, Ge , Matsumoto, Zhang	2012	Oral financial reporting: a rhetorical analysis of earnings calls	NLP / Conference Calls
Crawford Camiciottoli	2012	irs Detect Managers' Lack of Spontaneity? Adherence to Pre-determined Scripts during Earnings Conference Calls	NLP / Conference Calls
Lee, Zhou	2014	The Blame Game	NLP / Conference Calls
Brockman, Li, McKay Price	2014	Differences in Conference Call Tones: Managers versus Analysts	NLP / Conference Calls & Social Media
Chen, De, Hu, Hwang	2014	Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media	NLP / Conference Calls
Frankel, Jennings, Lee	2015	Using Unstructured and Quantitative Methods to Explain Accruals	NLP / Conference Calls
Jha, Bhagat, Montague	2015	Finding Value in Earnings Transcript Data with AlphaSense	NLP / Conference Calls
Brochet, Lourenç, Serfesheim	2015	Speaking of the short-term: disclosure horizon and managerial myopia	NLP / Conference Calls
Leu, Hwang, Chen	2015	Founders vs Professional CEOs	NLP / Conference Calls & Social Media
Call, Sharp, Shohfi	2016	Buy-Side Analysts' Participation on Public Earnings Conference Calls	NLP / Conference Calls
Jung, Wong, Zhang	2016	Buy-Side Analysts and Earnings Conference Calls	NLP / Conference Calls
Druz, Wagner, Zeckhauser	2016	Reading Managerial Tone: How Analysts and the Market Respond to Conference Calls	NLP / Conference Calls
Bushee, Gow, Taylor	2016	Linguistic Complexity in Firm Disclosures: Obsfuscation or Information?	NLP / Conference Calls
Brochet, Narajan, Yu	2016	Capital market consequences of language barriers in the conference calls of non-U.S. firms	NLP / Conference Calls
Brockman, Ciconi, Li, S. McKay Price	2017	Words versus Deeds: Evidence from Post-Call Manager Trades	NLP / Conference Calls
Borochina, Ciconi, Delisle, McKay Price	2017	The Effects of Conference Call Tone on Market Perceptions of Value Uncertainty	NLP / Conference Calls

(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/NLPres.png>)

## Web scraping: The Quest for Data

The NLP Profit Warning model data was obtained using web scrapping techniques in Python. Python is the favorite choice for web scrapping for three main reasons i) it allows to manipulate strings very easily, ii) it posses many scrapping libraries that makes web harvesting much easier and intuitive, and iii) Python is a compiled language thus turning it into a much more powerful tool than interpreted languages like R. Two main Python libraries were used to scrape the data: Scrapy and Selenium.

The data required to conduct this analysis came mainly from two sources:

- RTT News ([www.rttnews.com](http://www.rttnews.com)): RTT News is a content provider, delivering comprehensive and timely information on a wide variety of topics. The company provides open data about US profit warnings released over the last 9 months.
- Seeking Alpha ([seekingalpha.com](http://seekingalpha.com)): Seeking Alpha is a well-known s platform for investment research, with broad coverage of stocks, asset classes, ETFs and investment strategy. This website contains publicly available conference call transcripts for US stocks and ADRs (American Depository Receipt).



+200 stocks information for 1Q17 and 2016

([https://blog.nycdatascience.com/wp-content/uploads/2017/03/PW\\_Scrape.png](https://blog.nycdatascience.com/wp-content/uploads/2017/03/PW_Scrape.png))

## EDA: Data description and Feature Engineering

NYC

(<http://blog.nycdatascience.com/>)

NYC Open Data

(<http://blog.nycdatascience.com/> open-data/)

Open Data

(<http://blog.nycdatascience.com/> data/)

pandas

(<http://blog.nycdatascience.com/>)

python

(<http://blog.nycdatascience.com/>)

2/)

python machine learning

(<http://blog.nycdatascience.com/>)

machine-learning/)

python web scraping

(<http://blog.nycdatascience.com/>)

web-scraping/)

python webscraping

(<http://blog.nycdatascience.com/>)

Python Workshop

(<http://blog.nycdatascience.com/>)

workshop/)

R (<http://blog.nycdatascience.com/>)

random forest

(<http://blog.nycdatascience.com/>)

R Programming

(<http://blog.nycdatascience.com/>)

R Shiny

(<http://blog.nycdatascience.com/>)

r studio

(<http://blog.nycdatascience.com/>)

R Visualization

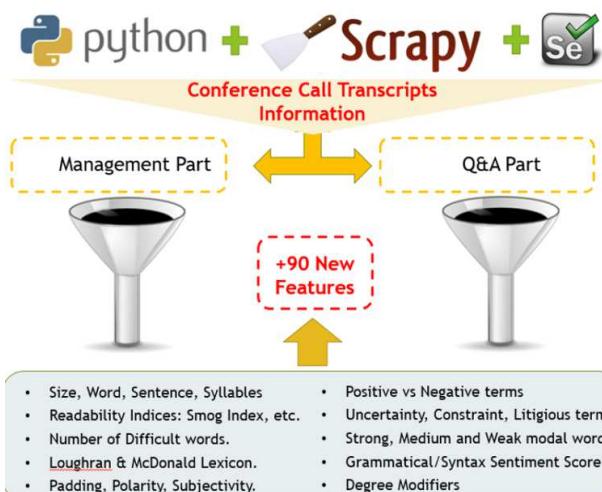
(<http://blog.nycdatascience.com/>)

Transcripts and fundamental data for more than 200 stocks was obtained, yet a narrower sample was used in the NLP analysis in order to control for specific factors (event time, sector, industry, etc):

- Only covers US companies are considered in the analysis thus ADRs are excluded.
- The training set only contains stocks within the Industrial sector, mainly capital goods companies.
- Event Analysis (profit warning occurrence) for dates between 1Q17 and 4Q16.
- Conference call transcripts analyzed are those 3 quarters before a profit warning event triggers: 1Q16, 2Q16 and 3Q16 calendar periods.
- 93 conference call transcripts analysed containing 42 transcripts from future profit warning stocks and 51 transcripts from healthy stocks.

Because of the fact that language overtones and lexicon may vary from one sector to another, it was necessary to choose a particular industry to build the foundation of the NLP model. In addition, time differences in topic importance, and lack of available profit warning data for periods beyond nine months, were a hindrance that limited the NLP analysis to a one-year time span. Lastly, NLP tools and libraries are so far more developed to appraise English-written texts, for which reason ADRs were excluded from the training set as previous research showed that NLP precision and accuracy vary differently depending on the mother tongue of the management team speaker.

Each conference call was split in two parts: "Management Discussion" (MD) and "Q&A". As mentioned in the research discussion of this post, several scholar and practitioner authors have found both parts can deliver significant NLP input. On the one hand, "Management Discussion" is focused only on the message transmitted by the management team - with no analysts interference - thus allowing to obtain a pure measure of the language complexity, semantics, lexicon and overall sentiment conveyed by the company's CEO and his team. On the other hand, "Q&A" is very helpful when seeking to obtain the overall sentiment of the conference call or to gather speech tone differences between management and analysts.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Process-1.png>)

More than 90 features were created out from the whole conference call transcripts and its MD and Q&A parts. Python NLP-related libraries used in this analysis were textstat, NLTK, VADER, pySentiment, spaCy and Gensim. Several NLP dimensions were measured in order to generate reliable and significant predictor categories related to text physical properties (size, number of words, number of syllables etc), text complexity (readability indices like Smog Index, padding), lexicon complexity (number of difficult words, Brown Dictionary), semantic and syntactic sentiment indices. The two main challenges when creating new NLP features are shown right below:

1. *Beyond semantics meaning:* classifying sentences or paragraphs using only word sentiment labeling is dangerous. For instance, both sentences "I am very happy" and "I am not very happy" will be classified as "positive" due to the presence of the word "happy". Degree modifiers like "not very" and also POS (Part-of-Speech) tagging that incorporates an adequate syntax meaning are among the key items to take care of when running NLP analysis. The use of Python libraries such as NLTK or VADER are helpful to solve degree modifying and POS issues, yet still there are parts of the text that become more challenging as it is always more challenging to treat short texts, i.e. flash news or social media brief posts, than long documents or transcripts.
2. *Sentiment dictionary choice matters:* using a standard English dictionary to classify words as negative or positive is not recommended when the text to be analyzed is finance-related. Loughran and McDonald (2011) provide a clear demonstration that applying a general sentiment word list to accounting and finance topics can lead to a high rate of misclassification: 75% of the negative words in the Harvard IV TagNeg dictionary of negative words are typically not negative in a financial context. For example, words like "mine", "cancer", "tire" or "capital" are often used to refer to a specific industry segment. These words are not predictive of the tone of documents or of financial news and simply add noise to the measurement of sentiment and attenuate its predictive value.

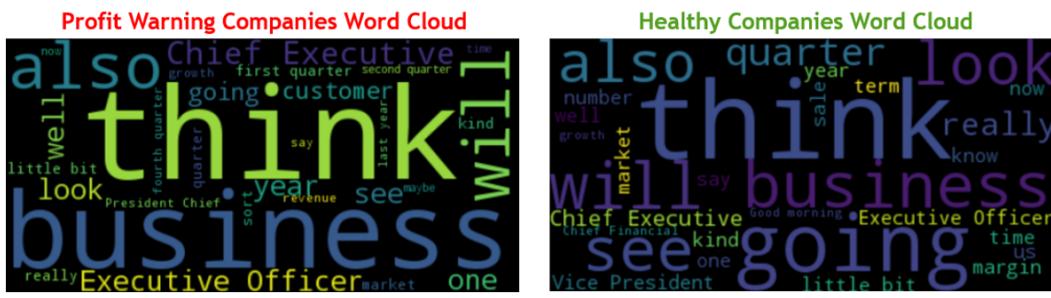
<b>R Workshop</b>	( <a href="http://blog.nycdatascience.com/r-workshop/">http://blog.nycdatascience.com/r-workshop/</a> )
<b>Scrapy</b>	( <a href="http://blog.nycdatascience.com/scrapy-visualization/">http://blog.nycdatascience.com/scrapy-visualization/</a> )
<b>Selenium</b>	( <a href="http://blog.nycdatascience.com/sentiment-analysis/">http://blog.nycdatascience.com/sentiment-analysis/</a> )
<b>Shiny</b>	( <a href="http://blog.nycdatascience.com/shiny/">http://blog.nycdatascience.com/shiny/</a> )
<b>Sports</b>	( <a href="http://blog.nycdatascience.com/sports-statistics/">http://blog.nycdatascience.com/sports-statistics/</a> )
<b>SVM</b>	( <a href="http://blog.nycdatascience.com/svm/">http://blog.nycdatascience.com/svm/</a> )
<b>Tableau</b>	( <a href="http://blog.nycdatascience.com/tableau-visualization/">http://blog.nycdatascience.com/tableau-visualization/</a> )
<b>web scraping</b>	( <a href="http://blog.nycdatascience.com/web-scraping/">http://blog.nycdatascience.com/web-scraping/</a> )
<b>webscraping</b>	( <a href="http://blog.nycdatascience.com/webscraping/">http://blog.nycdatascience.com/webscraping/</a> )
<b>XGBoost</b>	( <a href="http://blog.nycdatascience.com/xgboost/">http://blog.nycdatascience.com/xgboost/</a> )

Fortunately, Loughran and McDonald (LM) created a financial dictionary with customized lists of negative and positive words specific to the accounting and financial domain. LM dictionary has the additional benefit of showing additional dimensions of interest beyond the traditional dichotomy positive/negative. Among others, two noteworthy additions are the "Uncertainty" word list that attempts to measure the general notion of imprecision (without an explicit reference to risks), and the "Litigiousness" word list that may be used to identify potential legal problem situations. LM dictionary is used in the Profit Warning NLP model in order to adapt NLP sentiment classification methods to a corporate and finance-related world.

## EDA: Unsupervised Machine Learning

When conducting EDA (Exploratory Data analysis) as previous state to a more in-depth NLP analysis, it's always good to conduct some preliminary descriptive analysis to get familiar with the data and extract clues that could be useful later when developing Machine Learning models. Word Cloud was one of the first methods implemented, although without too much success. The two charts below show Word Cloud analysis for both profit warning company transcripts (left) and healthy companies (right) with no significant differences shown in their top-30 word lexicon.

The only relevant insight extracted from Word Cloud analysis came from comparing "uncertainty" lexicon for both classes (using LM financial dictionary tagging system): profit warning companies use more balanced vocabulary when talking about "uncertainty" than healthy companies. These findings are in line with Lee (2014) who highlights lack of spontaneity - managers following a carefully prepared script with more complex vocabulary and lexicon diversity - as one of the main factors in underperforming stocks.

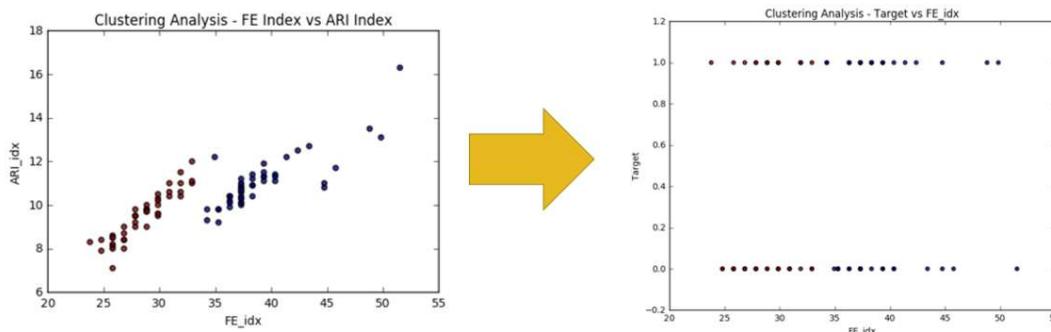


(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/WC.png>)

The results from clustering analysis were also misleading sometimes as the couple of plots below highlight. The first plot compares two well-known readability indices in linguistics:

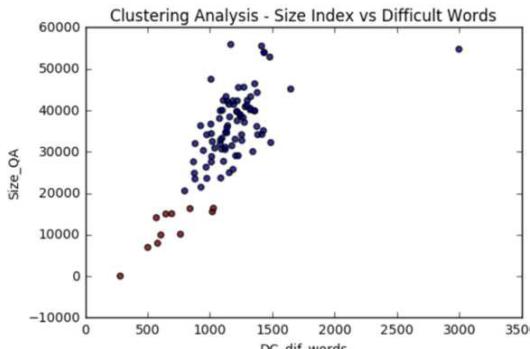
- *FE\_idx* (*x-axis*): Flesch Reading Ease index (1948) was created mainly to analyze DoD (Department of Defense) materials as well as Life Insurance documents. Originally the index ranges from 0-100 (very easy to very confusing) but it has been reversed to homogenize other readability indicators.
- *ARI\_idx* (*y-axis*): Automatic Readability Index (1967) was developed for general text readability purposes. ARI formula outputs a number which approximates the grade level needed to comprehend the text.

First impression is misleading as when using 2 clusters it seems to be two differentiated groups, particularly when using *FE\_idx*. Nevertheless, the second chart shows a lot of overlapping between *FE\_idx* (*y-axis*) and the response variable (*y-axis*, "1" is profit warning and "0" healthy company). Hence, whatever subgroups the clustering method is suggesting, they have nothing to do with the probability of being a profit warning company.

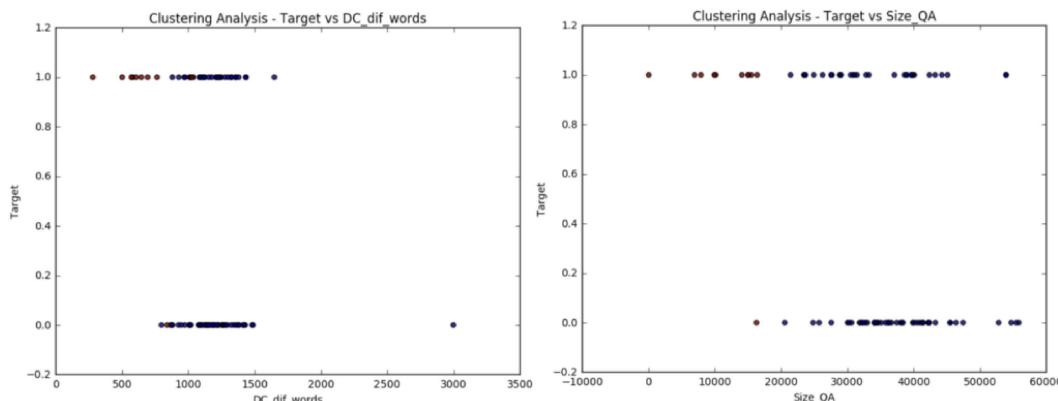


(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Clustering.png>)

More fruitful insights were extracted from clustering two other features: Size Q&A part and Difficult words. Once again the first chart below shows in red and blue how there seems to be 2 different subgroups. The next step is again to confirm that these two clusters are related to the probability of profit warning (Target):



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Clus2.png>)

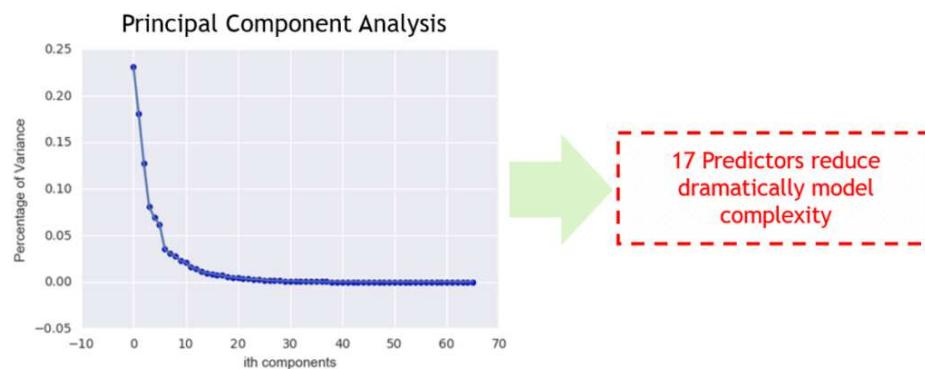


(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Clus3.png>)

Comparing both Size\_QA and DC\_dif\_words against our Target variable provides some comfort as it seems both variables are partially explaining the difference between profit warning (1) and non-profit warning(0) companies with two main takeaways:

- First takeaway is that companies about to issue a profit warning use a more simple "plain" English language than healthy ones. Bushee, Gow, Taylor (2016) demonstrate how management teams can use complex information to convey information (positive) or obfuscation (negative).
- Second takeaway is in line with Hollander, Pronk and Roelofsen (2008) evidence upon negative forward returns and managers's silence so our findings above about Size\_QA confirm that shorter QA sessions are more likely to happen in profit warning candidates.

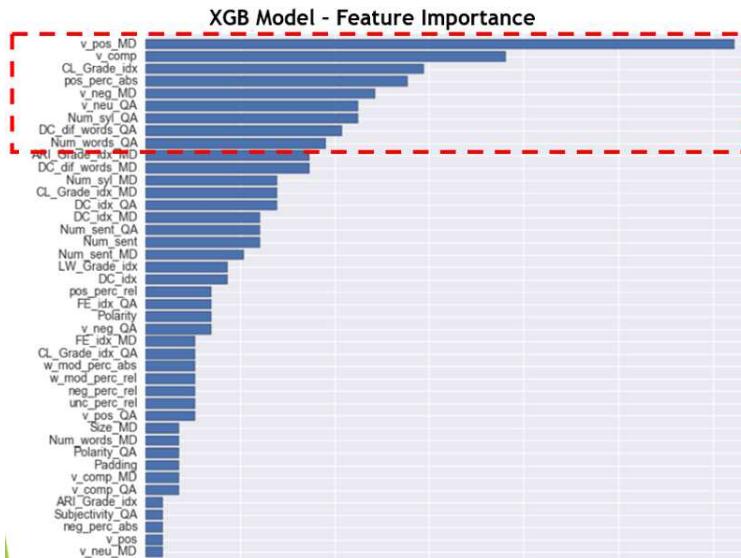
With more than 90 features created from the training data set, PCA (Principal Component Analysis) played a pivotal role to summarize feature information and ensure the elimination of any residual multicollinearity effect. The chart below shows how only 17 PCs (Principal Components) out of 93 predictors can explain more than 95% of the total variance. These PCs were later used in a logit model that delivered one of the best classification accuracy rates while ensuring a reasonable complexity in order to minimize overfitting risk.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/PCA.png>)

PCA variable transformation may be difficult to interpret as new non-correlated eigenvectors (PCs) are created from the original features. Hence, another alternative to analyze variable importance while keeping our original predictors intact are Ridge Regression, Lasso Regression or Gradient Boosting variable importance analysis. The bar chart below shows feature importance from a tree-based XGB model (Extreme Gradient Boosting) - one of the best Machine Learning methods used in the NLP Profit Warning model - with the next features standing out significantly:

- The percentage of positive sentiment in the management initial speech (`v_pos_MD`) is a game-changer feature along with the overall transcript sentiment score (`v_comp`). These two features were generated using NLTK and VADER libraries, which incorporate empirically derived measures for the impact of sentiment at sentence-level text. VADER incorporates word-order sensitive relationships between terms such as degree modifiers aka intensifiers, booster words, or degree adverbs.
- Coleman Liau Index (`CL_Grade_idx`) is the most decisive of all the readability indices. This index was created in 1975 with the purpose of determine text books readability. `CL_Grade_idx` outcome is tantamount to the grade of education for readers required to understand properly a particular text.
- The percentage of positive terms (`pos_perc_abs`) using the Loughran-McDonald financial dictionary played an important role to classify a company as a profit warning candidate. Profit warning management teams start lowering expectations very subtly via their vocabulary two or three quarters before the official announcement is released. Different tones of positiveness are very difficult to detect for human readers, for which reason NLP tools are extremely helpful in this area to wipe out behavioral biases such as conservatism bias (slowness incorporating new information), cognitive dissonance (reluctance to accept negative parts of the transcript) or confirmation bias (paying attention only to good news on the transcript)
- Particular Q&A session indicators are important: the number of difficult words, syllables and words along with Q&A neutral tone scoring are among other explanatory related to both text complexity and sentiment that add remarkable value when classifying companies using a XGB model.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/VI.png>)

## Machine Learning – Testing Models

Two main measures were used to decide what models were more effective to classify the training set observations as "profit warning" or "healthy":

- Misclassification error rate: The proportion of misclassified observations. This error rate is calculated using the next formula:

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i).$$

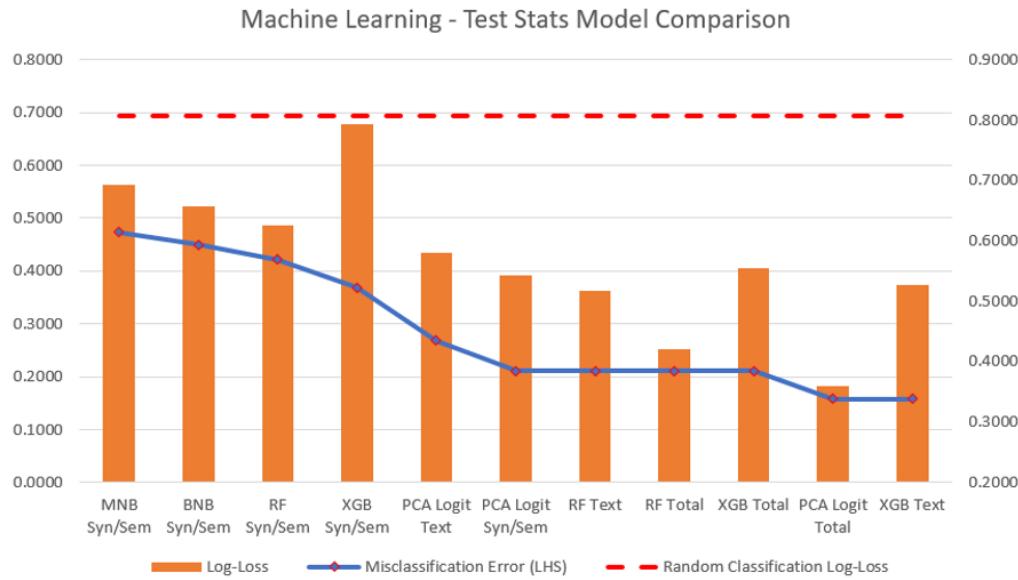
(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/ME.png>)

- Log-Loss error rate: it measures the difference between the distribution of actual labels and the classifier probabilities. A best case classifier with 100% accuracy will have a 0 log-loss, while a classifier that assigns randomly each observation to a k = 2 labels (profit warning or healthy) will have a log-loss of  $-\log(1/2) = 0.69315$ . The formula to calculate log-loss is shown below:

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}),$$

(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/LL.png>)

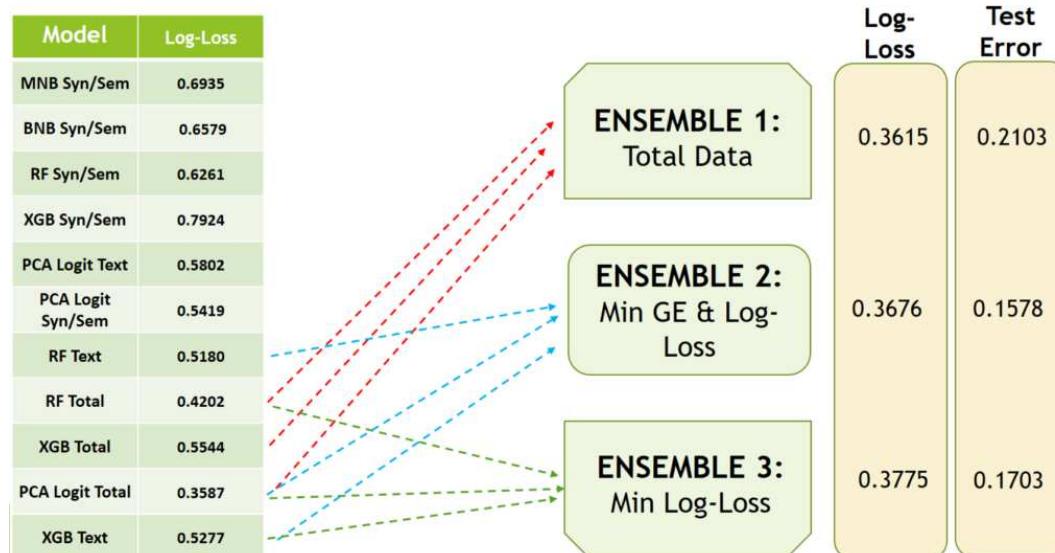
More than 30 different models were tested using Bernoulli Bayes, Multinomial Bayes, Logit Regression, Tree-based (Bagging, Random Forest, Boosting, Extreme Gradient Boosting), SVM (Support Vector Machine) with the 11 models displayed underneath passing the first preliminary acid test: outperform a random classification method tantamount to yield a log-loss error rate below 0.69315 (dashed red line). Models named as "Total" use the whole set of features or PCs (Principal Components) calculated from the total number of predictors, whereas models named "Text" have only text complexity predictors as input (readability indices, number of difficult words, etc) and those finished as "Syn/Sem" used only explanatory variables that convey syntax, semantic and sentiment. "PCA Logit Total" and "XGB Text" are the two best models with accuracy rates above 80%.



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Models-4.png>)

During the second stage the more powerful models from each type were combined using Python's Brew library, a comprehensive tool to ensemble and stack predicting models in order to enhance their stand-alone predicting ability. A majority voting rule criterion is used thus class assignment hinges on the prediction for the observation as predicted by the majority of models. To attain a much better predicting performance, model ensemble is conducted using three different criteria:

1. Ensemble 1: The top three models using "Total" predictors as defined two paragraphs above are used. This ensemble criterion intends to utilize as much data as possible in the classification process and uses RF (Random Forest) Total, XGB (Extreme Gradient Boosting) Total and PCA Logit (Principal Components) Total.
2. Ensemble 2: The top three models minimizing both Generalization Error and Log-Loss. Generalization Error aka "leakage" is the difference between the misclassification rate obtained from the test data set and the training data set. Models selected are RF Text, PCA Logit Total and XGB Text.
3. Ensemble 3: The top three models with lowest Log-Loss error rate. Models selected are RF Total (0.4202), PCA Logit Total (0.3587) and XGB Text (0.5277).



(<https://blog.nycdatascience.com/wp-content/uploads/2017/03/Ensem-1.png>)

"Ensemble 2: Min GE & Log-Loss" is the most balanced model with the second best Log-Loss error rate (0.3676) and lowest misclassification rate (0.1578 or 84.22% accuracy). "Ensemble 1: Total Data" yields the lowest Log-Loss but backfires in terms of test error (0.2103 misclassification rate). Although "Ensemble 2" results are slightly worse than the individual model "PCA Logit Total", ensemble models are preferred as they provide a much better diversified method of prediction than relying on a single approach.

An interesting and practical observation is that despite having a misclassification rate of approximately 15%, "Ensemble 2" portfolio management effectiveness as tool to shortlist short candidates is still high. The solution to this conundrum is as it follows: a stock classified as "profit warning" might eventually not issue a profit warning announcement, yet results from authors like Jha, Blaine, Montague (2015) provide evidence that stocks mislabelled as "profit warning" are to be at least companies with weak earnings momentum prospects that are more likely to underperform in the short term.

## NLP & Profit Warnings – Answers and Future Projects

At the beginning of this post several questions were set out with regards NLP utility for measuring and predicting Profit Warning risk. The lesson learnt is that NLP tools are helpful to boost productivity, save precious time and minimize investing behavioral biases as they allow analysts and Portfolio Managers to unveil additional information embedded into earnings conference call transcripts.

There are several expansions to be implemented in future analysis such as the inclusion of other sectors and industries, extension of the time span to at least five years, tagging and standalone analysis of management team different members (CEO, CFO, COO, etc) and analysts as well as the creation of more complex modeling methods such as neural networks.

To sum up, NLP tools either on a standalone basis, or complementing quantitative model based on numerical features (profit margin percentile level, sales growth, accrual ratios, etc), may enhance significantly a stock screening process aimed at identifying short candidates due to be suffering from profit warning risk or at least earnings weakness in the foreseeable future.

[Click here to check code in Github \(\[https://github.com/Carlossn/Python/tree/master/NLP\\\_Prof\\\_Warning\\\_Prediction\]\(https://github.com/Carlossn/Python/tree/master/NLP\_Prof\_Warning\_Prediction\)\)](#)

[Click here to visit my Blog \(<http://www.planetary-ascension.com/blog>\)](#)



**Carlos Salas Najera (<http://blog.nycdatascience.com/author/carlossn/>)**

Carlos is a passionate individual for investments and technology with long/short equity portfolio management experience that has worked over the last 10 years in combining his fundamental, quantitative and data science skills in order to deliver superior returns. His core strength lies on his understanding that investments is a continuous learning process that requires a regular disciplined effort and relentless mindset. Throughout the last five years, Carlos has been running investment mandates summing up more than 1 billion dollars, training junior analysts, enhancing his technological skills and delivering double-digit returns for his investors. He earned his MBA at AFI Business School, and degree in Business & Economics at the University of Alicante. In addition, he is also a CFA and graduated in Game Theory and Competitive Strategy at the London School of Economics. Currently, Carlos is augmenting his skills at the NYC Data Science Academy to increase the amount of non-structured data to be used in his investment models and develop machine-learning tools that tackle issues related to behavioral finance errors.

## Related posts



July 25, 2017



July 25, 2017

**Day Performance Similarities of Multiple Listed Companies - A Case Study**  
[\(http://blog.nycdatascience.com/student-works/day-performance-similarities-multiple-listed-companies-case-study/\)](http://blog.nycdatascience.com/student-works/day-performance-similarities-multiple-listed-companies-case-study/)

Read more

[\(http://blog.nycdatascience.com/student-works/day-performance-similarities-multiple-listed-companies-case-study/\)](http://blog.nycdatascience.com/student-works/day-performance-similarities-multiple-listed-companies-case-study/)

**How safe is New York public school drinking water?**  
[\(http://blog.nycdatascience.com/student-works/ny-public-school-water/\)](http://blog.nycdatascience.com/student-works/ny-public-school-water/)

Read more

[\(http://blog.nycdatascience.com/student-works/ny-public-school-water/\)](http://blog.nycdatascience.com/student-works/ny-public-school-water/)

**What school is the best "bang for your buck"? An analysis of the College Scorecard.**  
[\(http://blog.nycdatascience.com/student-works/college-scorecard/\)](http://blog.nycdatascience.com/student-works/college-scorecard/)

Read more

[\(http://blog.nycdatascience.com/student-works/college-scorecard/\)](http://blog.nycdatascience.com/student-works/college-scorecard/)

## 2 Comments



**Derek says**  
 Reply to [Carlos Salas Najera's comment](#) on July 18, 2017 at 1:30pm (<http://blog.nycdatascience.com/student-works/predicting-profit-warnings-nlp-applied-conference-call-transcripts-analysis/#comment-4314>)

Great post Carlos!



**Carlos Salas Najera says**  
 Reply to [Derek's comment](#) on July 18, 2017 at 3:20pm (<http://blog.nycdatascience.com/student-works/predicting-profit-warnings-nlp-applied-conference-call-transcripts-analysis/#comment-4315>)

Thanks Derek... you gave me a good edge teaching me R 😊

## Leave a Reply

Your email address will not be published. Required fields are marked \*

### Comment

Name \*

Email \*

Website

© 2017 NYC Data Science Academy Blog. All Rights Reserved.