Topics in Data Science: Applied Machine Learning for Financial Modeling

Group 4

Chenpeng Guan - cg3165 Rongyu Li - rl3098

Kassiani Papasotiriou - kp2535 Jiaxiang Wang - jw3864 Jinzhu Yang - jy3024

Text Mining on Earnings Call Transcripts

Spring 2020

Business Value-Add

An earnings call is a teleconference, or webcast, in which a public company discusses the financial results of a reporting period. It is of utmost importance to investors to assess the management's presentation and get a sense of the company's future performance. Therefore, creating a machine learning system that can quickly analyse and evaluate the sentiment of a EC transcript can be profitable and useful to parties of interest. In this project we extracted both textual and numerical information from JP Morgan's earnings call transcripts and created a scalable method to predict future movement of stock prices.

Short Model Description

Our goal was to predict the effect that the EC had on JP Morgan's stock price. In order to capture that effect we modeled our response feature as the change in slope of the stock price before and after the call. To predict this, we used two predicting features which we extracted from the transcript: text score and EPS spread. The former metric captured the performance of the company on a specific quarter based on textual analysis and the latter measured the gap between the consensus expected EPS and realized EPS reported on the call.

Data Preprocessing

In this section we will discuss the preprocessing we applied to the three variables of interest as mentioned above.

Price Trend

The response feature, the change of the price slope before and after the call, was created from the historical stock price using a multi-step approach. We first identified the date of the earning call and extracted out the 5-day price changes before and after the date. Using these 5-day prices we then performed a time-series linear regression to find the linear coefficient and used it as the price trend. After performing the regression steps for both before and after 5-day prices, we calculated the trend spread by simply subtracting the after trend with before trend.

Transcript Cleaning

We made the assumption in our model that not all speakers' words are of equal importance. More specifically, we assumed that in order to assess the company's performance it suffices to focus on the words of the CEO and CFO. We therefore filtered out all the other speakers present in the call (such as the call operator, analysts, etc.)

Earnings Per Share Spread

The other predicting feature, EPS spread, was created by subtracting the realized EPS announced during the earnings call by the corresponding consensus expected EPS.

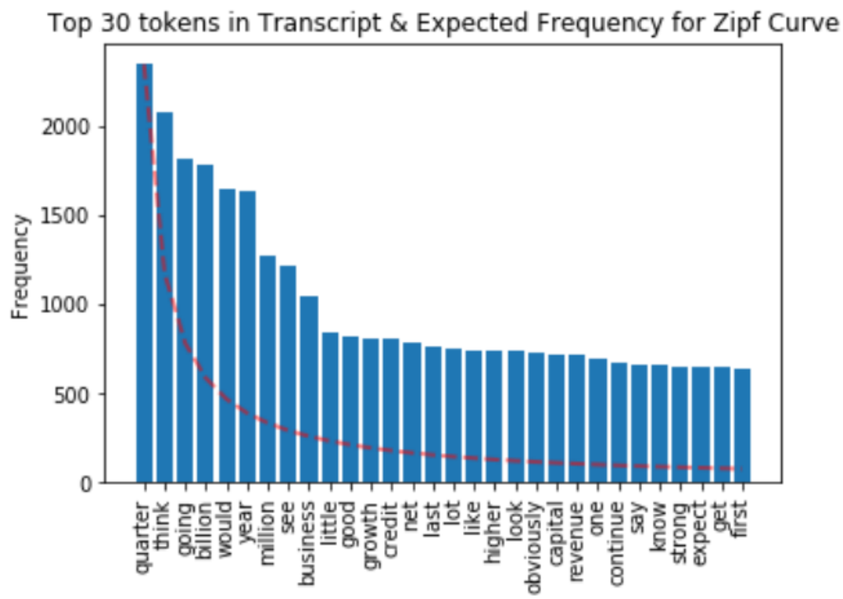
Methodology

Text Statistical Analysis

In order to get a better sense of the contents of EC transcripts we performed a textual statistical analysis. More specifically, we examined the frequency of words that appear in the JP Morgan transcripts and used word cloud and Zipf's curve as visualization.

As the first step, we did a preliminary cleaning to the text in order to remove high frequency words with no particular meaning such as stopwords and the names of the speakers. We then displayed the text frequency using the Zipf's curve.

As Zipf's Law stated, the frequency of any word is inversely proportional to its rank in the frequency table. In other words, the occurrence of the most frequent word should be approximately twice as often as that of the second most frequent word. In the graph presented below, the red line in the plot shows the expected word frequency and the blue bar shows the actual word frequency, which indicates that several actual occurrences of words did not strictly follow Zipf's distribution in our case.



In order to focus more on “sentiment significant'' words, we repeated the frequency analysis by narrowing the words scope to the positive and negative classes in the Loughran and McDonald Financial Dictionary (LMFD).[1] By comparing the frequency of the positive and negative words (as defined in LMFD) appearing throughout the transcripts, we concluded that the CEOs and CFOs of JPMorgan tend to use more positive language in their presentations and tend to avoid words with negative tone. Bellow is the resulting word cloud:

---

[1] The Loughran and McDonald lexicon is a collection of words, along with their sentiments, tuned for applications in finance. It consists of seven categories: positive, negative, strong modal, weak modal, uncertainty, litigious, constraining. For more reference, visit https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists

Sentiment Analysis

In this section we outline the natural language processing techniques we applied on the transcripts to compute features for the predictive task. We will describe the two main sentiment analysis techniques we attempted, explain their advantages and weaknesses, and motivate our final choice between them.

Before we delve into these methods, let us define the concept of sentiment polarity that we use for scoring a transcript. Given a word, we defined its sentiment polarity as a continuous value in [-1 ,1], with -1 being absolutely negative, 1 being absolutely positive and 0 being neutral. On a sentence level we averaged the sentiment of its words, and on a transcript level we averaged the sentiment of its sentences. We explored two general methodologies to extract the sentiment of the transcripts: (i) rule-based sentiment analysis using WordNet's lexicon, and (ii) rule-based sentiment analysis of our own design, using numeric-based parsing.

Method 1: Sentiment Analysis based on WordNet's Lexicon

WordNet's subjectivity lexicon for English adjectives is a lexicon that contains over 2900 English adjectives along with their sentiment polarity and subjectivity score. We used the Python package TextBlob as a wrapper around this lexicon, with the added functionality of transforming the adjectives to adverbs (e.g. "terrible" becomes "terribly") and negation (e.g. "not bad" has a positive sentiment). We tried to improve the sentiment scoring in the following two ways:

Using the Loughran and McDonald lexicon

As mentioned in the Text Statistical Analysis section, the Loughran and McDonald lexicon is a collection of words, along with their sentiments, tuned for applications in finance. We combined this lexicon with the scoring of WordNet to focus on scoring only "sentiment important" sentences.

Sentence-level filtering

The transcripts of the executive officers, while informative for the most part, contained sentences such as "Good morning, everyone." and "Now please turn to Page 3". We filtered these sentences out using a rule-based approach and then scored the resulting transcripts with the aforementioned method.

Shortcomings of Method 1

Thorough analysis of the transcripts with the above methods revealed two shortcomings. First, these lexicons operate best on short text passages. WordNet-based sentiment analysis has proven to be very effective on tweets, while the Loughran and McDonald lexicon is very effective on short headlines about stock exchange. However, there is a large domain gap between these

domains and the transcripts at hand. The transcripts are spoken words as opposed to written, which are often quite lengthy and with broad topics that often include multiple subjects.

The second shortcoming lies in the fact that there is a lack of sentiment in the transcripts themselves since they are meant to be delivered in a strictly professional setting. The transcripts mostly outline factoids and numeric figures, and do not interpret these facts and figures with emotionally loaded words. In this context, the “positive” and “negative” words are often very subtle (e.g. “up” and “down”) and are not included in the above lexicons.

Method 2: Sentiment Analysis with Numeric Parsing (**SANP**)

The second method we attempted was based on parsing the numbers that were present in the EC transcripts and the sentiment of the words that surround them. The key assumption for this method is that a number surrounded by positive words (such as “up”, “grew”, “increased”, etc.) is considered a good sign for the company’s performance in that specific quarter, and equivalently, a number surrounded by negative words implies a lower performance. We therefore scored a transcript based on the number of positive and negative words that surrounded all the numbers that were present. Motivated by the shortcomings of method 1, we constructed a lexicon that is fine-tuned for these transcripts as our base case scoring scenario and as a second step, and enhanced it with the Loughran and McDonald lexicon. The results of this scoring method were used for the final model prediction.

Correlation Analysis of Variables

In this section we performed a statistical analysis on our predicting features, with a purpose of assessing the significance of EPS spread and results from the sentiment analysis by exploring the relationship between each of them and the trend spread.

Sentiment Score vs Trend Spread

Using simple linear regression, we ran a statistical analysis between each of the sentiment scores and the trend spread.

The results of this analysis are shown in the table below.

Method	R2	p-value
WordNet only	0.058537535	0.09756435
WordNet + Loughran McDonald	0.004392581	0.65446625
SANP only	0.08265421	0.047548642
SANP + Loughran McDonald	0.16119258	0.004679046

From the result, we observed that our custom sentiment analysis method produces a higher correlation, therefore we choose that for feature generation.

EPS vs Price Trend

Below this the result of the SLR between EPS spread and trend spread.

Method	R2	p-value
EPS	0.1909	0.001903

Prediction

Model Selection

We decided to pursue prediction using both regression and classification methods, with the response feature being trend spread for regression and 0-1 conversion of price spread (negative spread as 0 and positive spread as 1) for classification. Given the simplicity of our assumed model, we concluded that using multiple linear regression (**MLR**) and decision tree regressor (**DTR**) as the regression models and logistic regression (**LogR**) and decision tree classifier (**DTC**) as the classification models would be sufficient for our prediction purpose.

The performance of a model can only be conclusive when it is generalized over many runs. Therefore, we layered in an additional cross validation process in the step of calculating scores and accuracies for each model. This process splitted the original data as train and test data

iteratively for 20 times, and each split was then fed into the corresponding models for training and testing. For each model, we then took the average of the 20 score results and used it as the generalized performance measure. By this approach we were able to draw a more confident conclusion on the performance of each model.

Performance

There are two scenarios of data that we used for predicting, one where the sentiment score feature was calculated using the restricted lexicon SANP method and the other one using SANP and Loughran McDonald lexicon method.

The metric we used to assess the predicting power of the regression models is R^2 and the metric we used to evaluate our classification model is accuracy score. Listed below is the summary of scoring results for each model.

Base Case Model:

Predicting feature: EPS and SANP

Response feature: Price spread

Performance

Classification	Accuracy
DTC	0.565
LogR	0.585

Regression	R2
DTR	-2.229
MLR	-0.971

Improved Model:

Predicting feature: EPS and SANP with lexicon

Response feature: Price spread

Performance

Classification	Accuracy

DTC	0.49
LogR	0.54

Regression	R2
DTR	-2.135
MLR	-0.788

Note that a model that constantly predicts the expected value of the response feature will yield a R^2 of 0. In this case we can observe that the regression models performed poorly under both scenarios. In a two-feature-classifiers scenario like this, a model that randomly guesses the outcome would yield an accuracy of 0.5 on average, therefore we can conclude that none of the classification models have a strong predicting power, with the base case scenario being a little better than the improved one.

Improvements

As we can see from the Prediction Section above, there is plenty of room for improvements on both the data preprocessing side and the model selection side. A more complicated model with more independent variables related to stock price would probably be a better predictor of the price trend.

Future improvements could also involve a sentiment analysis scorer trained on financial texts. The TextBlob model that was used initially to score the transcripts is trained in tweets and reviews which by default have a very different language than an EC transcript. Therefore using such a model in our case didn't yield the most insightful results. Moreover, we believe that building a sentiment lexicon tailored to a specific company could improve the sentiment scores.

Therefore, the above two improvements could result in a more sophisticated and accurate model for sentiment analysis.

References

Historical stock prices:

Yahoo Finance - https://finance.yahoo.com/quote/JPM/history/

Consensus Expected EPS: https://www.estimize.com/jpm/fq2-2020?metric_name=eps&chart=historical

Earning Call Transcripts: Factiva