Predicting Stock Price Changes from Earnings Calls and What (not) to say (when)

Vikram Rao Sudarshan
Cornell University
vikram@cs.cornell.edu

December 8, 2014

Abstract

We[1] analyze corporate earnings calls of companies to observe patterns in speech content. In particular, usage of legal and management jargon, and positive and negative words yields useful insights on how companies communicate their performance. In addition, we use these features to predict the direction of stock change of the company, as well as stock volatility.

1 Introduction

Companies are mandated to regularly communicate their performance to their shareholders. These quarterly Earnings Calls (ECs) are usually held as conference phone calls in which company executives report the performance of the company in a prepared statement and then answer particular, detailed questions asked by analysts from investment banks. These earnings calls are later analyzed by the analysts who use the call and prior knowledge to predict the company's performance in the next quarter. Since such analysis is expensive and time-consuming, automatic prediction of stock price changes from the text of earnings calls would be very useful. It is argued qualitatively in Rehm (2013) that the Q&A is essential to a productive earnings call.

The availability of data and the usefulness of earnings calls has vastly increased following the Sarbanes-Oxley Act (2002) which mandates quarterly and annual communication from companies. Automated analysis of financial documents has thus been done in a number of settings. Kogan et al. (2009) showed that using unigram and bigram features improved accuracies in predicting stock volatility over a baseline of historical volatility. Loughran and McDonald (2011) manually constructed a financial sentiment lexicon more representative of financial words than the widely used Harvard General Inquirer lexicon. Tsai and Wang (2013) considered the task of ranking companies by volatilities. Tsai and Wang (2014) showed that expanding the financial sentiment lexicon by using continuous word vector representations improved performance on volatility ranking and regression tasks. However, they all use Form-10Ks which are performance reports but do not have a live question and answer component that could contain information absent in a prepared document. Leidner and Schneider (2010) used earnings calls to construct a graph of risk patterns to be used for risk alerting.

In another line of work, Lavrenko et al. (2000) identified news stories associated with trends in financial time series. Devitt and Ahmad (2007) performed document level sentiment analysis on financial news.

Differently from prior work, we focus on the task of predicting *stock change direction* using *earnings calls* rather than *volatilities* using *Form-10Ks*, and show that treating prepared statements and Q&A separarely improves performance and gives better insights. To our best knowledge, we are the first work to treat speeches and Q&A separately and analyze the importance of differences between them. In addition, we predict over a *60 day window* whereas earlier work predict volatility over a long *12 month window*. The binary classification task is of predicting whether the stock change corresponding to an earnings call is in the top/bottom quarter of all stock changes by using only textual features. We show that using TF-IDF unigram features alone gives a baseline predictive accuracy of over 55%, which is significant given the challenging task of predicting stock changes. Running a topic model (LDA) on the earnings calls and using topic proportions alone as features

[1] Joint work with Jon Kleinberg, Lillian Lee and Sendhil Mullainathan

gave better performance than using unigram features. The topic modeling approach has the additional advantage of being highly interpretable and amenable to longitudinal studies on the performance of specific subsectors of industry corresponding to specific topics.

To facilitate comparison with prior work, we also report mean squared errors (MSE) on volatility prediction. The qualitative results on the distinction between the prepared speech and Q&A holds in this case also.

2 Dataset

Earnings Calls are transcribed and made available online by many websites. We downloaded 45k transcripts from a popular site and stock prices from Yahoo! Finance. After discarding calls with unavailable stock prices, we were left with 36k transcripts involving 4233 firms.

For the prediction task between the top and bottom 25% of the data, our final dataset involves 18004 documents (2667 test from 2013, 15337 train from earlier). For the volatility prediction task, we used the full dataset involving 36k documents.

3 Prediction Tasks

We are given a date-annotated document set $\mathcal{D}$ of tuples (d_i, f_i, x_i) where d_i is the date of the earnings call, f_i is the firm and x_i are the text features.

3.1 Stock-change direction

Given a pair of day offsets (b, e), we form the labelled dataset $\mathcal{L}^{(b,e)}$ by annotating $\mathcal{D}$ with stock changes

$$s_i^{(b,e)} = \frac{stock(f_i, d_i + e) - stock(f_i, d_i + b)}{stock(f_i, d_i + b)}$$

The classification task $\mathcal{C}_p^{(b,e)}$ involves predicting between the top/bottom p^{th} percentiles of the stock changes, training on past instances while testing on current instances.

We focus on the $\mathcal{C}_{25}^{(-10,50)}$ prediction task. While the task can be performed using any percentiles, we stick to 25 in order to have separable classes and be able to interpret the model well.

As a baseline, we use unigram text features using a vocabulary of 5947 words. We filtered out esoteric (sparse, industry-sepecific) words by removing words that occur in earnings calls of at most 500 companies and ignoring named entities annotated through Stanford NER. Speeches and the QA section were treated as separate documents and TFIDF vectors were computed as follows:

$$\begin{aligned} \text{TF}(w,d) &= \frac{n_{w,d}}{\max_w(n_{w,d})} \\ \text{IDF}(w,d) &= \log\left(\frac{N}{\sum_{d'} \mathbb{I}(n_{w,d'} > 0)}\right) \\ \text{TFIDF}(w,d) &= \text{TF}(w,d).\text{IDF}(w,d) \end{aligned}$$

Also, we ran LDA with 50 topics treating speeches and QA rounds as separate documents to get topics and represent each document by its topic probability vector. No preprocessing was done in this case. The number of topics was selected to balance redundancy among topics and the specificity of the topics.

We also trained two different 50 topic LDA models - one for all speeches and one for all Q&A rounds (LDAS) instead of having the same model generating both speeches and Q&A rounds. While thas the same number of features as above and accommodates the fact the speeches and Q&A sections can be talking about different topics, the two sets of 50 features used in the classification are semantically unrelated. We show later that doing this loses information valuable to the classification task.

3.2 Volatility

Define return $r_t(f) = \frac{stock(f,t)}{stock(f,t-1)} - 1$. The volatility of f between times t_1 and t_2 is the empirical standard deviation of the set of numbers

$$R_{t_1}^{t_2}(f) = \{r_t(f) | t_1 \leq t \leq t_2\}$$

The regression task involves predicting the log-volatility of the stock for the 12 months following the earnings call $v^{(12)}$.

As a baseline, we used one feature - the volatility of the stock for the 12 months preceding the earnings call $v^{(-12)}$. We compare this with the LDA features described above.

4 Experiments

4.1 Stock-change direction

Using TFIDF features (BoW) and topic proportions for the 50 topics (LDA and LDAS) for the speech and QA section, we used an SVM to predict between the top and bottom 25% of the stock changes (the data points corresponding to the middle 50% were discarded). The results are shown in the **S+QA** column. BoW thus had 11894 features while LDA had 100 features.

To highlight the differences and importance of the speech and QA section, three other prediction tasks were performed - using only speech features (**S**), only QA features (**QA**) and using both features but swapping the features for the speech and QA section in the test data with probability 0.5 ((**S+QA**)$_P$). The classification accuracies are reported in terms of AUC (Area Under the Curve) and summarized in Table 1. ROC curves follow in Figure 1.

	S+QA	**(S+QA)**$_P$	**S**	**QA**
BoW	0.587	0.566	0.556	**0.593**
LDA	**0.607**	0.579	0.598	**0.607**
LDAS	0.548	0.548	0.534	**0.553**

Table 1: Performance on classification (test 2013)

When using LDA features, performance was similar when using either the speech or the QA section alone, or both. However, the permutation test gave lower performance, indicating that there are significant differences in the weights of the topics for the speech and the QA section. However, when using the baseline BoW features, the QA section alone is most predictive of stock change. The speech has least predictive power. Also, permuting the speech and QA section reduced performance significantly.

The message here is twofold. First, that the QA section in earnings calls holds significant signal for predicting stock changes, which we are to the best of our knowledge the first work to leverage. Second, the good and bad features (words and topic proportions) differ significantly between the speech and QA section as seen by the reduced performance on permuting the corresponding features 50% of the time.

In addition, having separate topic models for the two sections worsens performance, thus underscoring the fact that *differences* in the same topic between the speech and Q&A round are important rather than the topic proportions themselves.

4.2 Volatility

Similar experiments were performed for the regression task on 12-month volatility using Support Vector Regression similar to Kogan et al. (2009). Results for the mean squared error in log volatility are shown in Table 2. The same qualitative results on the importance of treating the speech and Q&A separately and using the the same topic model for both of them hold here.

	S+QA	**(S+QA)**$_P$	**S**	**QA**
$v^{(-12)}$: 0.267	-	-	-	-
LDA	**0.347**	0.565	0.367	0.364
$v^{(-12)}$+LDA	**0.207**	0.282	0.210	**0.207**
LDAS	**0.340**	0.481	0.369	0.366
$v^{(-12)}$+LDAS	0.211	0.239	**0.210**	0.212

Table 2: Performance on Volatility (test 2013)

5 Discussion

LDA topics with the highest and lowest weights (in both the speech and QA sections) are shown in Table 3, and the top words for some of the topics are shown in Table 4.

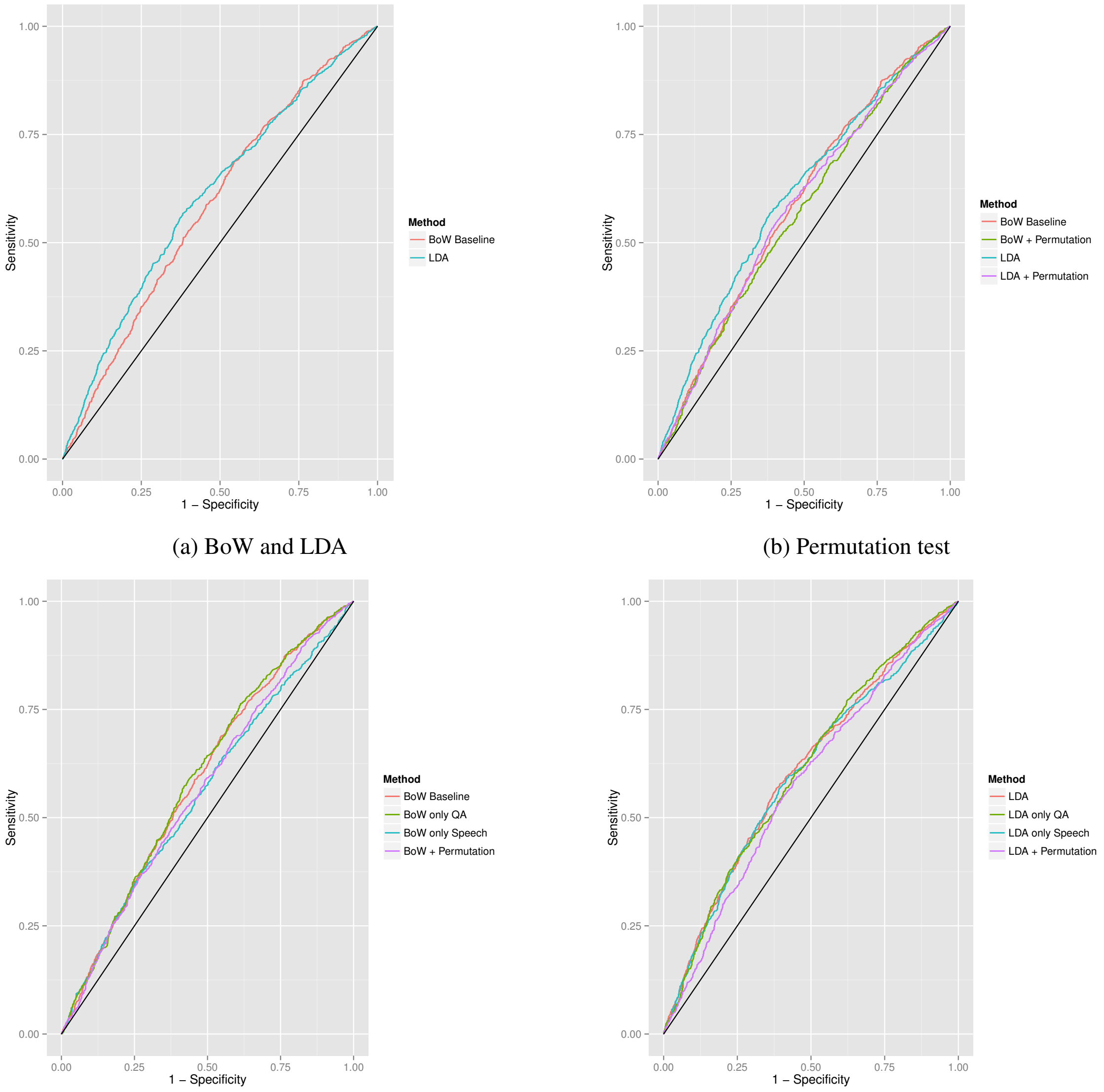
+ Speech	*Negative*, Positive-1, Financial terms, *International*
+ QA	Product jargon, Management+, Oil rigs, Tech/Cloud
- Speech	Positive-2, Advertising, Product terms, Transportation
- QA	*Negative*, Legal terms, Losses/Insurance, Education

Table 3: Top Features

The top predictive features agree with our intuition. Saying negative things in the speech is good (accepting the truth) whereas the same is bad in the QA section, where you are presumably being forced to admit negative performance. In addition, Financial terms and positive management jargon is good, as is being international. Legal terms and talk of losses and insurance is bad in the QA section.

6 Conclusion and Future Work

We proposed the task of predicting stock change direction from the text of earnings calls and showed that the QA section is interesting and informative. A Bag of Words approach yields over 55% AUC and using topic models yields better performance while enhancing interpretability of the model. We also predicted long-term stock



(a) BoW and LDA

(b) Permutation test

(c) BoW - S, QA, S+QA

(d) LDA - S, QA, S+QA

Figure 1: ROC Curves

volatility and show the same qualitative results to hold even in this setting.

Earnings Calls hold much promise for predicting stock changes and are also a valuable data source for studying hedging, lying and other phenomena. We hope that this work leads to other interesting work in this domain.

Acknowledgments

We'd like to thank Jon Kleinberg, Lillian Lee and Sendhil Mullainathan with whom we started exploring this work, and David Mimno for motivating a fresh perspective on the topic (no pun intended) and for the interesting discussions.

References

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi and Noah A. Smith 2009. *Predicting Risk from Financial Reports with Regression* NAACL.

Ming-Feng Tsai and Chuan-Ju Wang 2014. *Financial Keyword Expansion via Continuous Word Vector Representations* EMNLP.

Jochen L. Leidner and Frank Schilder 2010. *Hunting for the Black Swan: Risk Mining from Text* ACL.

Ming-Feng Tsai and Chuan-Ju Wang 2013. *Risk ranking from financial reports* Advances in Information Retrieval.

Tim Loughran and Bill McDonald 2011. *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks* The Journal of Finance.

Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen and James Allan 2000. *Language Models for Financial News Recommendation* CIKM.

Ann Devitt and Khurshid Ahmad 2007. *Sentiment Polarity Identification in Financial News: A Cohesion-based Approach* ACL.

Werner Rehm 2013. *Three Steps to a More Productive Earnings Call* McKnisey Report.

Negative	Management+	Financial terms	Losses/Insurance	Legal terms	Positive-1	Positive-2
cost	performance	million	business	company	quarter	billion
lower	company	cash	quarter	process	year	business
quarter	focus	debt	loss	case	growth	growth
costs	management	flow	capital	companies	operating	year
environment	key	capital	rate	future	strong	cost
decline	years	ebitda	insurance	current	share	eur
economic	opportunities	balance	million	additional	earnings	net
reduction	team	interest	ratio	stock	increased	half
levels	position	credit	portfolio	important	basis	end
economy	progress	facility	investment	agreement	margin	increase
impact	investments	sheet	year	potential	fourth	group
conditions	improve	free	losses	legal	impact	trading
reduced	initiatives	adjusted	growth	situation	due	revenues
reduce	support	sale	rates	plan	rate	good
negative	strategy	liquidity	book	related	billion	results
business	core	assets	premium	provide	cash	costs
savings	level	loss	higher	specific	points	positive
improvement	investment	notes	risk	timing	full	volume
actions	long-term	paid	underwriting	shareholders	guidance	performance
challenging	strategic	accounting	premiums	tax	continued	reduction
impacted	focused	transaction	lines	state	segment	margin
difficult	forward	financing	life	litigation	income	financial
reductions	important	agreement	reinsurance	final	adjusted	strong
quarters	profitability	expense	significant	information	higher	ratio
pricing	benefits	remaining	products	difficult	prior	profit

Table 4: Top Words