HKU
BUSINESS
SCHOOL
港大經管學院

Estimating and Predicting a Company’s Sales Growth: Using Natural Language Processing and Machine Learning on Financial Statements

Sibo Ding

April 2023

Motivation

Sales is one key financial indicator for companies and investors. Financial statements are the most authoritative and influential documents for companies. Natural language processing (NLP) and machine learning are two powerful tools to analyze textual data. Therefore, it is valuable to estimate and predict sales growth using NLP and machine learning to analyze financial statements.

Company Selection: Coca-Cola

I select Coca-Cola (KO) mainly considering two reasons. First, it is a worldwide food company listed on the Dow Jones Industrial Average. Second, its stock price fluctuates. Therefore, it is easier to estimate the impact of financial statements on different growing patterns.

Data Collection

The time span is from 1994 to 2002, including 36 quarters. Quarterly reports (10-Q and 10-K) are downloaded from SEC EDGAR and the company's website. Quarterly sales data are downloaded from Bloomberg. There is no sales for 2001Q4, so it is substituted with the average of 2001Q3 and 2002Q1.

NLP Preprocessing

As original financial statements are not appropriate for analysis directly, some preprocessing needs to be done. An example "Accounting and 180 _Revenue" demonstrates the following steps.

First, as most small letters have the same meaning as capital letters, all letters are lowercased, so the example is converted to "accounting and 180 _revenue".

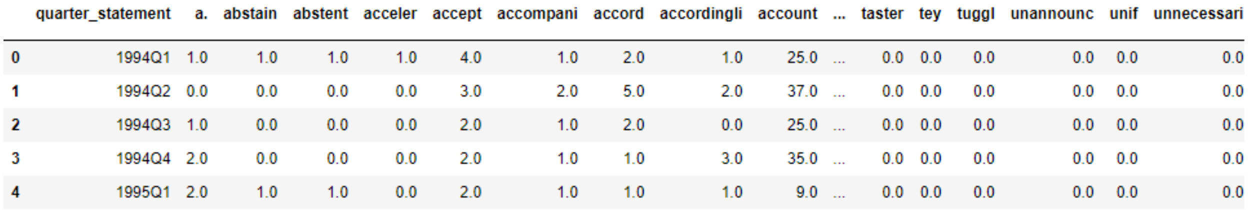
Second, whole paragraphs are split (tokenized) into words, so the example becomes four words: "accounting" "and" "180" "_revenue".

Third, as words starting with a number or a special character often do not have useful information in these financial statements, they are removed so "accounting" "and" are retained.

Fourth, as stop words (such as "the", "and", "in") do not carry useful information, they are removed so "accounting" remains.

Fifth, many words in English are derived from a root word (e.g., "normality" is derived from "norm"). For easier analysis, all derived words are converted (stemmed and lemmatized) into their root forms, so "accounting" is converted to "account".

After all these steps, I count and output the number of occurrences of each word, which is also called Bag-of-Words (BOW) in Figure 1.



	quarter_statement	a.	abstain	abstent	acceler	accept	accompani	accord	accordingli	account	...	taster	tey	tuggl	unannounc	unif	unnecessari
0	1994Q1	1.0	1.0	1.0	1.0	4.0	1.0	2.0	1.0	25.0	...	0.0	0.0	0.0	0.0	0.0	0.0
1	1994Q2	0.0	0.0	0.0	0.0	3.0	2.0	5.0	2.0	37.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2	1994Q3	1.0	0.0	0.0	0.0	2.0	1.0	2.0	0.0	25.0	...	0.0	0.0	0.0	0.0	0.0	0.0
3	1994Q4	2.0	0.0	0.0	0.0	2.0	1.0	1.0	3.0	35.0	...	0.0	0.0	0.0	0.0	0.0	0.0
4	1995Q1	2.0	1.0	1.0	0.0	2.0	1.0	1.0	1.0	9.0	...	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1. Bag-of-Words

Regression on BOW

Intuitively, regression is using factors to explain a number. After knowing these factors, we use them to predict that number. Here, I use BOW to predict sales growth. I set 28 quarters (80%) as training data and 8 quarters as test data, and this division stays the same in classification. I use three models: linear regression, random forest regressor, and XGBoost regressor. Then I compute the square root of mean squared error (RMSE) for each model. It measures the inaccuracy of predictions compared to actual values, and it has the same comparable unit as the outcome variable (sales growth). The results are in Table 1.

	Linear regression	**RF regressor**	**XGB regressor**
MSE	0.005587	0.001619	0.004380
RMSE	0.0747 (7%)	0.0402 (4%)	0.0662 (6%)

Table 1. Regression Results

7% RMSE when using linear regression to predict sales growth is a bit large. 4% RMSE using RF regressor is marginally acceptable.

Classification on BOW

Intuitively, classification is using factors to explain a few categories (not a number this time). After knowing these factors, we use them to predict those categories. Here, since the outcome variable (sales growth) is continuous, I artificially classify it into 2, 3, or 4 classes, and use BOW to predict them. I use four models: Naïve Bayes, logistic regression, random forest classifier, and XGBoost classifier. Then I compute the predictive accuracy for each model, which is the number of accurate predictions out of test size 8. The results are in Table 2.

	Naïve Bayes	**Logistic regression**	**RF classifier**	**XGB classifier**
2 classes by 0	0.5	0.875	0.875	0.875
2 classes by median	0.75	0.75	0.75	0.625
3 even classes	0.5	0.75	0.625	0.75
4 even classes	0.375	0.625	0.5	0.5

Table 2. Classification Results

Logistic regression predicts most accurately, followed by RF classifier, XGB classifier and Naïve Bayes. Using logistic regression, 87.5% (7/8) for predicting 2 classes is quite accurate, however, each class has a wider range. 62.5% (5/8) for predicting 4 classes is marginally acceptable because each class has a narrower range.

Clustering on BOW

Intuitively, clustering is grouping similar data together. I use two models: k-means clustering and hierarchical clustering. I cluster financial statements of all 36 quarters into 2, 3, …, 8 clusters, and evaluate the similarity of sales growth within each cluster. Figure 2 shows the results of sales growth in 4 k-means clusters.

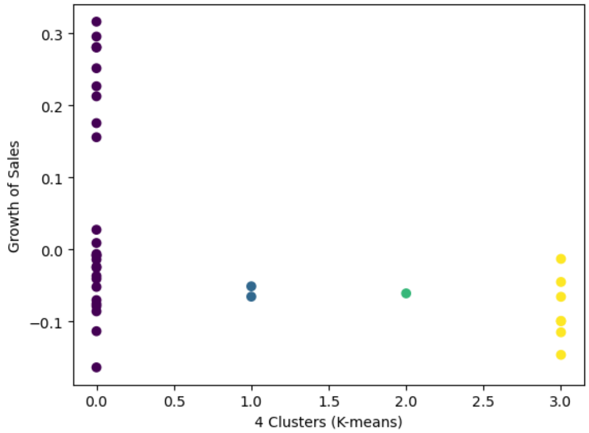


Figure 2. Sales Growth in 4 K-means Clusters

Sales growths within each cluster are not significantly similar (clustered). Cluster 3 only contains negative sales growths, but this is not robust as Cluster 0 also contains many negative sales growths. Other than 4 clusters, different numbers of clusters also do not show desirable outcomes.

Sentiment Analysis

The rationale of sentiment analysis is simple. In Afinn lexicon, positive words have higher sentiment scores, for example "amazing" is 4. I multiply the sentiment score of each word by its count (which is BOW), then sum up all multiplications in one document to get one sentiment score for each quarter. Figure 3 shows the relationship between sales growth and sentiment score.

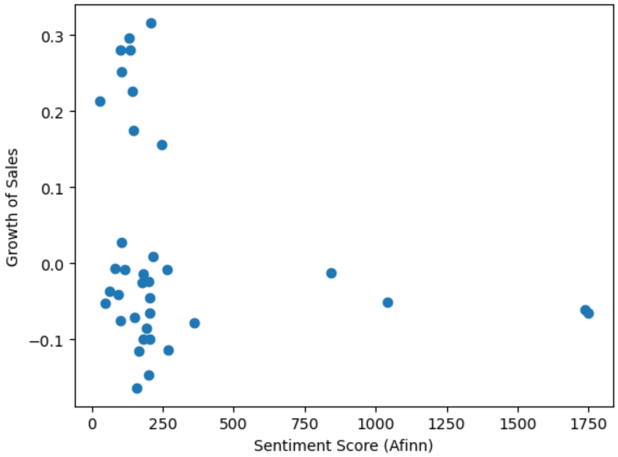


Figure 3. Relationship between Sales Growth and Sentiment Score

There is no significant relationship between sales growth and the sentiment score of financial statements.

tf-idf

tf-idf (term frequency–inverse document frequency), similar to BOW, measures how important a word is in a collection of documents. The formulas and rationales are as follows,

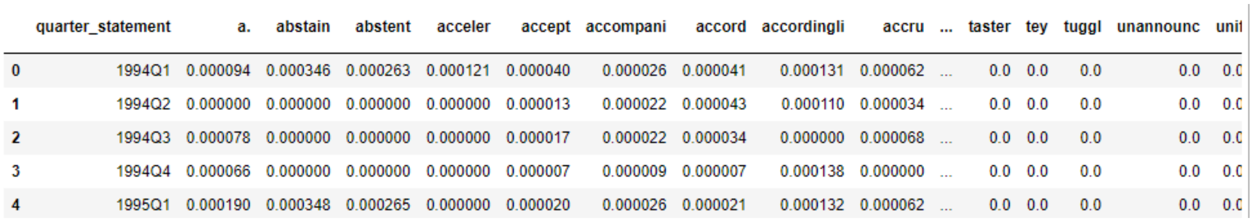
$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} = \frac{Count\ of\ a\ word\ in\ a\ document}{Total\ word\ count\ in\ this\ document}$$

Because the denominator is fixed, tf will be higher if a word appears more in a document.

$$idf(t, D) = \log\frac{N}{|\{d \in D: t \in d\}|} = \log_{10}\frac{Total\ number\ of\ documents}{Number\ of\ documents\ where\ a\ word\ appears}$$

Because the numerator is fixed, idf will be higher if a word appears in fewer documents.

tf-idf is the multiplication of tf and idf, shown in Figure 4.



	quarter_statement	a.	abstain	abstent	acceler	accept	accompani	accord	accordingli	accru	...	taster	tey	tuggl	unannounc	unit
0	1994Q1	0.000094	0.000346	0.000263	0.000121	0.000040	0.000026	0.000041	0.000131	0.000062	...	0.0	0.0	0.0	0.0	0.0
1	1994Q2	0.000000	0.000000	0.000000	0.000000	0.000013	0.000022	0.000043	0.000110	0.000034	...	0.0	0.0	0.0	0.0	0.0
2	1994Q3	0.000078	0.000000	0.000000	0.000000	0.000017	0.000022	0.000034	0.000000	0.000068	...	0.0	0.0	0.0	0.0	0.0
3	1994Q4	0.000066	0.000000	0.000000	0.000000	0.000007	0.000009	0.000007	0.000138	0.000000	...	0.0	0.0	0.0	0.0	0.0
4	1995Q1	0.000190	0.000348	0.000265	0.000000	0.000020	0.000026	0.000021	0.000132	0.000062	...	0.0	0.0	0.0	0.0	0.0

Figure 4. tf-idf

I input tf-idf into all previous models, but there is no increase in performance.

Case Study: Highest Sales Growth

To understand some characteristics of an individual quarterly statement, I investigate the quarter with the highest sales growth. The highest sales growth is 31.6% in 2002Q2. Since random forest regressor is the most accurate model previously, I use it to predict this sales growth, and the result is 27.8%.

The 10 most frequent words in this statement are “compani", “million”, “oper”, “incom", “net”, “month”, “share”, “account”, “six”, “consolid”. Several points are noticeable here. First, words are being stemmed and lemmatized, so they are not in normal formats. Second, financial statements have certain templates, so frequent words in this statement may also appear frequently in other statements. Third, frequent words are closely related to business operations, so they may not carry useful information in specific situations.

To better visualize the relative frequency of words, I create a word cloud, shown in Figure 5.

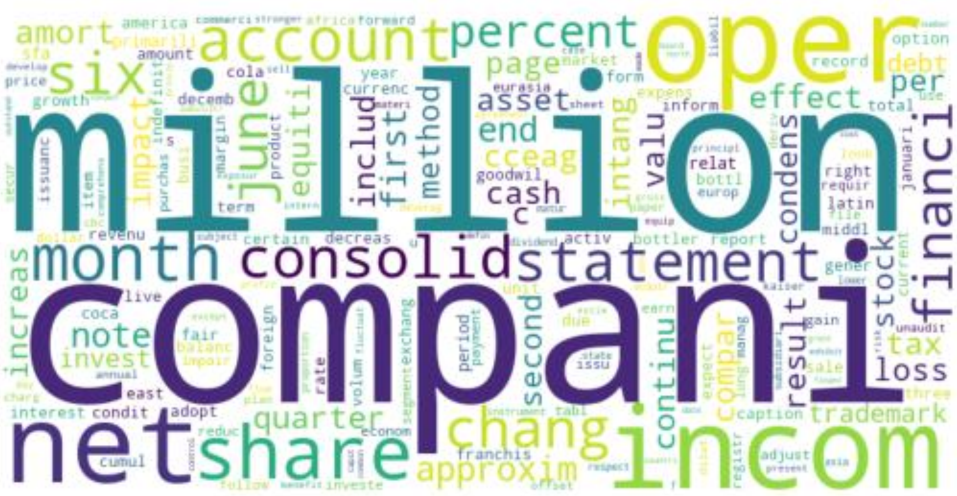


Figure 5. Word Cloud of the Statement with the Highest Sales Growth

Further Research

Beyond what this research has covered, further research can use other NLP methods to convert text to number. It can also use other NLP packages and other machine learning models. It can collect data for more quarters or collect other types of data to build composite models. Beyond the sales growth of Coca-Cola, it can also focus on other financial indicators or other companies.

References

Coca-Cola Stock Price

https://www.google.com/finance/quote/KO:NYSE?window=MAX

SEC EDGAR

https://www.sec.gov/edgar/browse/?CIK=21344&owner=exclude

Coca-Cola Website

https://investors.coca-colacompany.com/filings-reports/annual-filings-10-k?page=4

tf-idf

https://en.wikipedia.org/wiki/Tf%E2%80%93idf
https://baike.baidu.com/item/tf-idf/8816134