ESTIMATING AND PREDICTING A COMPANY'S SALES GROWTH: NLP + ML ON FINANCIAL STATEMENTS

Sibo Ding
Apr 24, 2023

TABLE OF CONTENTS

MOTIVATION

Sales: One key financial indicator for companies and investors

Financial statements: The most authoritative and influential documents for companies

Natural language processing (NLP) + Machine learning (ML): Two powerful tools to analyze textual data

Thus, it is valuable to estimate and predict the sales growth using NLP and ML to analyze financial statements

COMPANY SELECTION: COCA-COLA

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

2 reasons:

A world-wide food company listed on Dow Jones Industrial Average

Stock price fluctuates → Easier to estimate the impact of financial statements on different growing patterns

DATA COLLECTION

Time span: 1994 – 2002 (36 quarters)

Quarterly reports (10-Q & 10-K) (“.txt” file)

Sources: SEC EDGAR & Company’s website

https://www.sec.gov/edgar/browse/?CIK=21344&owner=exclude
https://investors.coca-colacompany.com/filings-reports/annual-filings-10-k?page=4

Quarterly sales data on Bloomberg

No sales for 2001Q4 → Substitute it with the average of 2001Q3 and 2002Q1

NLP PREPROCESSING

E.g., “Accounting and 180 _Rd”

01

Lower case

“accounting and 180 _rd”

02

Tokenize

“accounting” “and” “180” “_rd”

03

Retain words starting with a letter using regex

“accounting” “and”

04

Remove stop words

“accounting”

05

Stem and lemmatize

“account”

NLP PREPROCESSING: OUTPUT

Bag-of-Words (BOW): Number of occurrences of each word

	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

MACHINE LEARNING INTUITION

Regression: Use factors to explain a number. After knowing these factors, use them to predict that number.

REGRESSION ON BOW

Linear regression, Random Forest regressor

28 quarters (80%) training + 8 quarters test (same in classification)

Compute mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

	Linear regression	**RF regressor**
MSE	0.005587	0.001619
Square root of MSE	0.0747 (7%)	0.0402 (4%)

Square root to have the same comparable unit
as the outcome variable (sales growth)

MACHINE LEARNING INTUITION

Regression: Use factors to explain a number. After knowing these factors, use them to predict that number.

Classification: Use factors to explain a few categories. After knowing these factors, use them to predict those categories.

CLASSIFICATION ON BOW

Classify sales growth (outcome variable) into 2 / 3 / 4 classes

Naïve Bayes, Logistic regression, Random Forest classifier

Compute predictive accuracy: accurate predictions out of 8 (test size)

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

Softmax function for multi-class logistic regression

MACHINE LEARNING INTUITION

Regression: Use factors to explain a number. After knowing these factors, use them to predict that number.

Classification: Use factors to explain a few categories. After knowing these factors, use them to predict those categories.

Clustering: Group similar data together.

CLUSTERING ON BOW

K-means & Hierarchical clustering

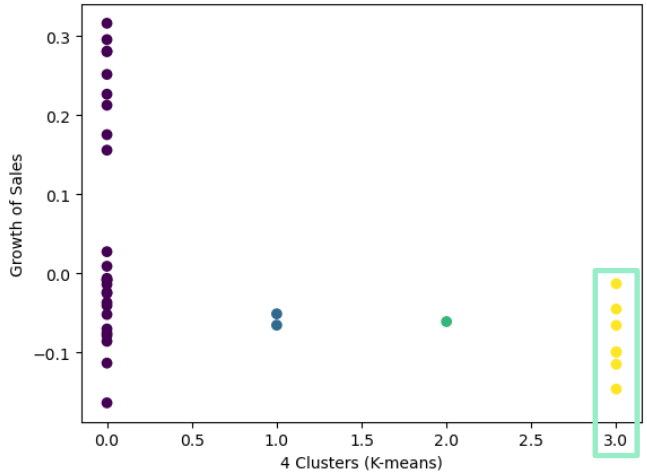
36 quarters of financial statements

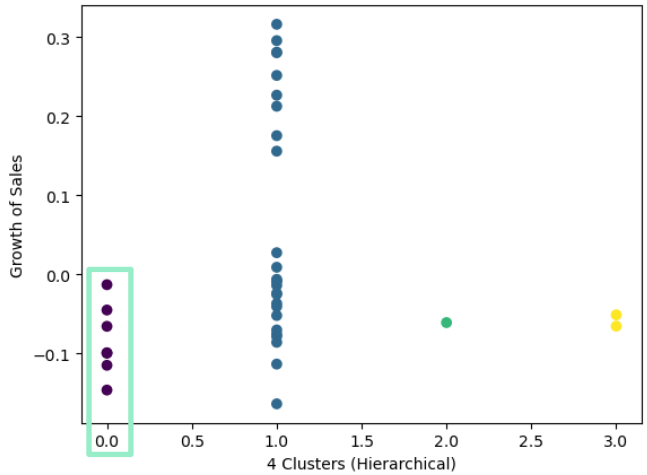
Try from 2 to 8 clusters

Evaluate the similarity of sales growth in each cluster

→ No desirable outcomes

negative





SENTIMENT ANALYSIS

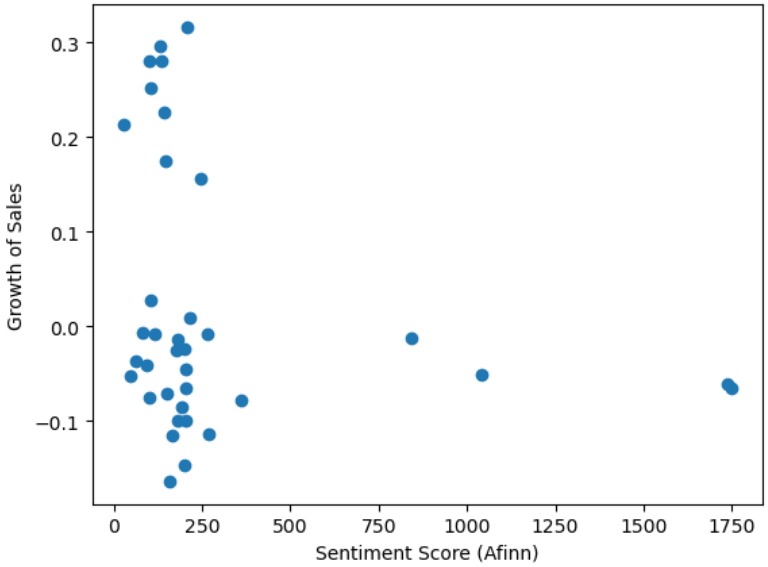
Afinn lexicon

Positive words have higher sentiment score
e.g., amazing: 4

Multiply the sentiment score of each word by its count (BOW)

Sum up and get 1 sentiment score for each quarter

→ No significant relationship



TF-IDF

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

Higher if a word appears more in a document

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

Higher if a word appears in fewer documents

$$tf - idf = tf \times idf$$

Manually calculate it using Pandas on BOW

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

https://baike.baidu.com/item/tf-idf/8816134

TF-IDF: OUTPUT

	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

Input it into all models

→ No increase in performance

FURTHER RESEARCH

- Other NLP methods to convert text to number
- Other NLP packages
- Other ML models
- Data for more quarters
- Collect other types of data → Build composite models
- Other financial indicators as the outcome variable
- Other companies

THANKS!