Predicting Stock Performance Using SEC Fillings and NLP

Joey(Junlin) Liu

Background

The **Efficient Market Hypothesis** claims that stock market movements are unpredictable, as all the known information about companies, interest rates, politics, economic trends, and more are already calculated into the price of the stock. Only **New Events** can effectively change the perceived value of a company and its stock.

Theoretical Form

$$P_t = E_t[M_{t+1}(P_{t+1} + D_{t+1})]$$

E_t = expected value given information at time t,
M_t = the stochastic_discount factor
D_t = the dividend the stock pays next period.

$$P_t = ME_t[P_{t+1}].$$

$$\log P_t = \log M + E_t[\log P_{t+1}]$$ Random walk with drift

New Events

SEC Fillings

https://www.sec.gov/edgar/searchedgar/companysearch.html

Past research has focused on using 10K and 10Q to predict stock price, and the results were impressive. This study will exclude above two kinds of fillings together with Form4 and rely on rest of the forms to make predictions.

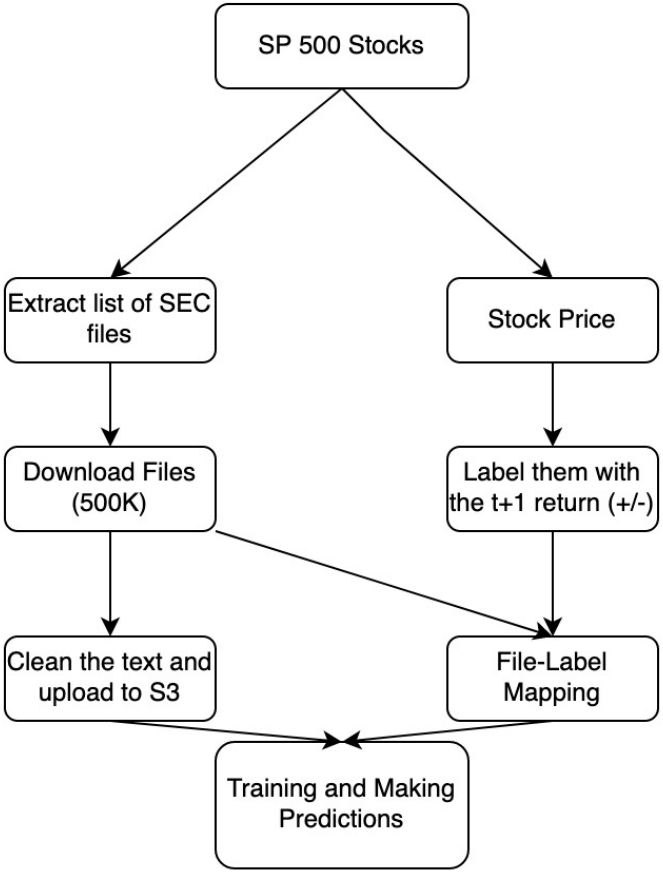
List of SEC forms:
https://www.sec.gov/forms

Form 4:
Not efficient in obtaining information
Browser Text

Table of Contents

- Data Preparation
- Modeling
  - Bag-Of-Words
  - TF-IDF
  - Financial Word Embeddings
  - RNN
  - RNN + Financial Word Embeddings
  - BERT
- Evaluation

Data Processing Flow



Data Preparation

- Raw Files (XML/HTML)

Samples:

https://www.sec.gov/Archives/edgar/data/100517/0001193125-15-006646.txt (html)

https://www.sec.gov/Archives/edgar/data/1090727/0001225208-15-001163.txt (xml)

https://www.sec.gov/Archives/edgar/data/1070750/0001086364-15-000082.txt (txt)

- ~500k files in total, each between 3KB – 5MB
- Reduce to less than 10% of original file

Data Preparation

- Raw Files (XML/HTML)
- Text with numbers and junk words

```
     UNITED STATESSECURITIES AND EXCHANGE COMMISSIONWashington, D.C. 20549SCHEDULE 13GUnder the Securities Exchange Act of
1934(Amendment No. 2)CHARTER COMMUNICATIONS(Name of Issuer)COMMON STOCK(Title of Class of Securities)16117M305(CUSIP
Number)December 31, 2014(Date of Event which Requires Filing of Statement)Check the appropriate box to designate the Rule pursuant
to whichthis Schedule is filed:[x]Rule 13d - 1(b)Rule 13d - 1(c)Rule 13d - 1(d)           1Name of Reporting PersonT. ROWE PRICE
ASSOCIATES, INC.52-05569482Check the Appropriate Box if a Member of a GroupNOT APPLICABLE3SEC Use
Only_____4Citizenship or Place of OrganizationMARYLANDNumber of Shares Beneficially Owned by Each
Reporting Person With5Sole Voting Power*289,6406Shared Voting Power*-0-7Sole Dispositive Power*874,0408Shared Dispositive
Power-0-9Aggregate Amount Beneficially Owned by Each Reporting Person874,04010Check Box if the Aggregate Amount in Row (9)
Excludes Certain SharesNOT APPLICABLE11Percent of Class Represented by Amount in Row 90.7%12Type of Reporting PersonIA*Any shares
reported in Items 5 and 6 are also reported in Item7.           Item 1(a)Name of Issuer:Reference is made to page 1 of this
Schedule13GItem 1(b)Address of Issuer\'s Principal Executive Offices:12405 POWERSCOURT DRIVE, ST. LOUIS, MO 63131Item 2(a)Name of
Person(s) Filing:(1)T. Rowe Price Associates, Inc. (PriceAssociates)(2)Attached as Exhibit A is acopy of an agreement between the
Persons Filing (as specified hereinabove) that this Schedule 13G is being filed on behalf of eachof them.Item 2(b)Address of
Principal Business Office:100 E. Pratt Street, Baltimore, Maryland21202Item 2(c)Citizenship or Place of Organization:
(1)Maryland(2)Item 2(d)Title of Class of Securities:Reference is made to page 1 of this Schedule13GItem 2(e)CUSIP Number:
16117M305Item 3The person filing this Schedule 13G is an:XInvestment Adviser registered under Section 203 of the Investment
Advisers Act of 1940Investment Company registered under Section 8 of the Investment Company Act of 1940Item 4Reference is madeto
Items 5-11 on the preceding pages of this Schedule 13G.           Item 5Ownership of Five Percent or Less of a Class.Not
Applicable.XThis statement is being filedto report the fact that, as of the date of this report, the reporting person(s) has
(have) ceased to be the beneficial owner ofmore than five percent of the class of securities.Item 6Ownership of More than Five
Percent on Behalf of AnotherPerson(1)Price Associates does not serve as custodian of the assets of any of its clients;
accordingly, in each instance only the clientor the client\'s custodian or trustee bank has the right to receive dividends paid
```

Data Preparation

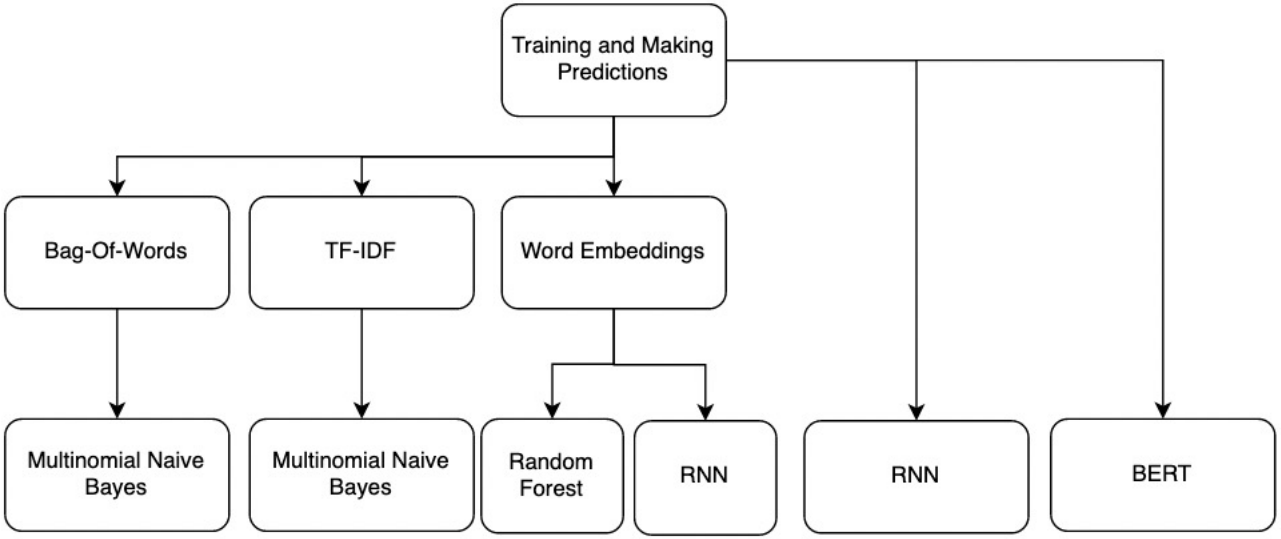
- Raw Files (XML/HTML)
- Text with numbers and junk words
- Further cleaning with stopwords and pretrained embeddings

```
earliest event registrant specified principal executive telephone including area code changed appropriate box filing intended
simultaneously satisfy filing obligation registrant communications pursuant material pursuant communications pursuant communications
pursuant holding company primary subsidiary provide investor update related preliminary financial operational results fourth quarter year
investor update incorporated information including furnished shall filed purposes subject liabilities shall deemed incorporated reference
document filed pursuant shall expressly set forth specific reference issued press release reporting operational press release attached
incorporated information including furnished shall deemed filed purposes subject liabilities shall deemed incorporated reference
registration statement document filed pursuant tothe shall expressly set forth specific reference dated issued dated herewith requirements
registrant duly caused report signed behalf bythe undersigned hereunto duly dated issued dated herewith investor update provides guidance
certain forward looking statements information investor update contains preliminary financial operational results fourth quarter forward
looking statements consolidated system available seat miles increased estimated compared period prior consolidated domestic decreased
approximately international increased estimated versus fourth quarter estimates consolidated system increased passenger revenue available
seat mile increased versus fourth quarter guidance negatively impacted percentage points related certain interline recorded fourth quarter
expects cargo revenue million million expects revenue million year consolidated increased expects cargo revenue million revenue billion
fourth quarter expects consolidated cost excluding profit business special increase expects consolidated excluding profit business
expenses special increase year expects record approximately million business expense fourth quarter approximately million business revenue
associated business activities recorded estimates consolidated fuel price fourth quarter including hedge closed percentage points hedge
adding approximately gallon hedge hedged fuel early settlement hedge expects approximately million cash settled hedge losses fourth
```

Data Preparation

- Raw Files (XML/HTML)
- Text with numbers and junk words
- Further cleaning with stopwords and pretrained embeddings
- Split training (2.5GB) and test data(0.5GB)

Modeling FLow



Modeling Baseline

- A naïve discriminative approach based on the type of forms

Form Type	-1	1	Number of Files	Pred
424B2	49%	51%	27557	1
8-K	50%	50%	27477	1
SC 13G/A	46%	54%	21521	1
FWP	54%	46%	6333	-1
3	51%	49%	4938	-1
SC 13G	46%	54%	4450	1
425	53%	47%	2949	-1
DEFA14A	49%	51%	2839	1
4/A	49%	51%	2129	1
DEF 14A	48%	52%	2030	1

	precision	recall	f1-score	support
-1.0	0.46	0.14	0.21	15206
1.0	0.52	0.86	0.65	16759
accuracy			0.51	31965
macro avg	0.49	0.50	0.43	31965
weighted avg	0.49	0.51	0.44	31965

Bag-Of-Words + Multinomial Naïve Bayes

abandoned	abandonment	abbreviated	abbreviations	abetted	abide	abilities	ability
0	0	0	0	0	0	0	20
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

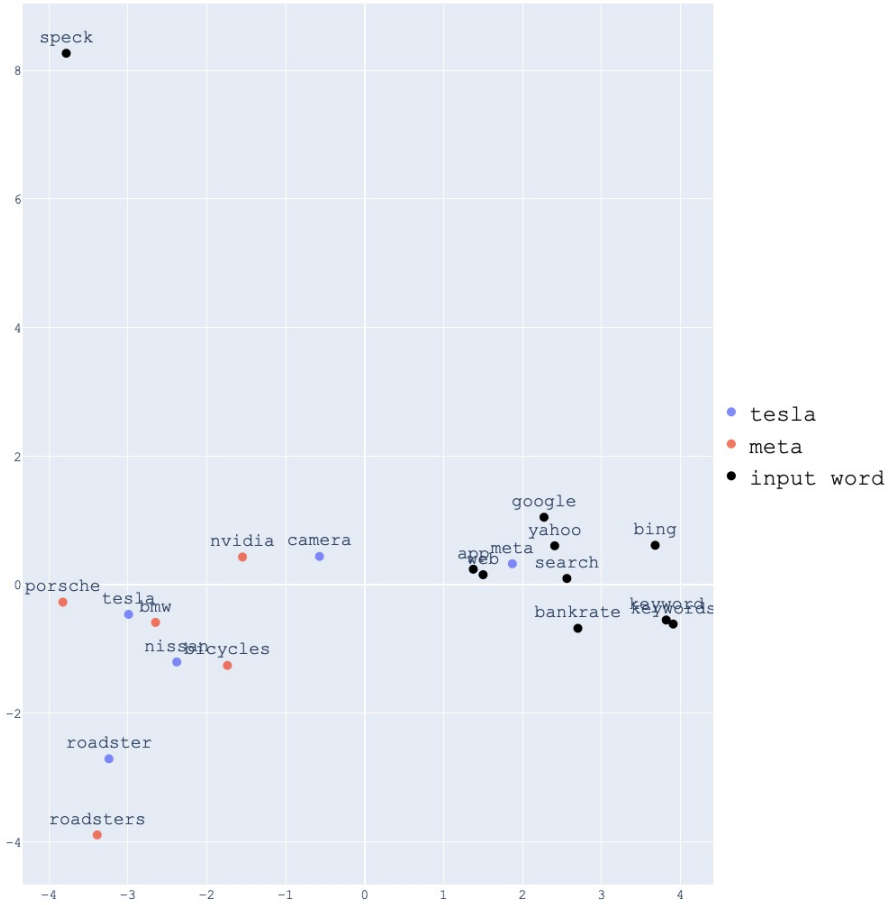
	precision	recall	f1-score	support
0.0	0.50	0.72	0.59	16086
1.0	0.53	0.30	0.39	16483
accuracy			0.51	32569
macro avg	0.52	0.51	0.49	32569
weighted avg	0.52	0.51	0.49	32569

TF-IDF + Multinomial Naïve Bayes

abandoned	abandonment	abbreviated	abide	ability	able	abroad
0.0	0.0	0.0	0.0	0.177265	0.0	0.0
0.0	0.0	0.0	0.0	0.000000	0.0	0.0
0.0	0.0	0.0	0.0	0.000000	0.0	0.0

	precision	recall	f1-score	support
0.0	0.47	0.52	0.50	16086
1.0	0.48	0.44	0.46	16483
accuracy			0.48	32569
macro avg	0.48	0.48	0.48	32569
weighted avg	0.48	0.48	0.48	32569

Pretrained Financial Embedding



Learning Word Embeddings from 10-K Filings for Financial NLP Tasks

Saurabh Sehrawat
e-mail: saurabhsehrawat@gmail.com

Abstract

In this paper, we generate word embeddings learned from corpus of 10-K filings by corporates in U.S. to S.E.C from 1993 to 2018 using word2vec model implemented in PyTorch. Word Embeddings learned from a general corpus of articles from Google News, Wikipedia etc. are readily available online for researchers to use in their models but embeddings learned from 10-K filings are not publicly available. We publish the word embeddings learned from 10-K filings on GitHub for other researchers to use in their NLP tasks such as document classification, document similarity, sentiment analysis, readability index etc. on 10-K filings or other financial documents. We show that using these learned word embeddings we can differentiate between different types of sentiment words in the widely used Loughran-McDonald word lists and generate average similarity scores between them. We also present an application of word embeddings where we can quantitatively track changes in 10-K documents using the learned embeddings.

Pretrained Financial Embedding

- Average embedding vectors of each document
- Embedding dimension of 300
    - Results a 1 * 300 vector for each document
      = 300 features
- Random Forest

	precision	recall	f1-score	support
0.0	0.50	0.60	0.55	29695
1.0	0.50	0.40	0.44	29459
accuracy			0.50	59154
macro avg	0.50	0.50	0.49	59154
weighted avg	0.50	0.50	0.49	59154

RNN

	precision	recall	f1-score	support
0.0	0.50	0.61	0.55	16086
1.0	0.52	0.41	0.46	16483
accuracy			0.51	32569
macro avg	0.51	0.51	0.51	32569
weighted avg	0.51	0.51	0.51	32569

LSTM Structure

```
embedding_size = 32
model.add(Embedding(len(tokenizer.index_word)+1, embedding_size))
model.add(LSTM(32, dropout=0, recurrent_dropout=0))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

```
model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 32)          1172064

 lstm (LSTM)                 (None, 32)                8320

 dense (Dense)               (None, 32)                1056

 dense_1 (Dense)             (None, 1)                 33

=================================================================
Total params: 1,181,473
Trainable params: 1,181,473
Non-trainable params: 0
```

RNN + Pretrained Financial Embedding

LSTM with pretrained embedding layer

	precision	recall	f1-score	support
0.0	0.51	0.58	0.54	16086
1.0	0.53	0.46	0.49	16483
accuracy			0.52	32569
macro avg	0.52	0.52	0.52	32569
weighted avg	0.52	0.52	0.52	32569

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 500, 300)	10988100
lstm (LSTM)	(None, 32)	42624
dense (Dense)	(None, 32)	1056
dense_1 (Dense)	(None, 1)	33

Total params: 11,031,813
Trainable params: 43,713
Non-trainable params: 10,988,100

BERT

Because of the limitation of 512 subword tokens, if a document is too long, it only keeps the beginning 1500 characters and the last 1500 characters

	precision	recall	f1-score	support
0	0.50	0.97	0.66	29695
1	0.47	0.03	0.06	29459
accuracy			0.50	59154
macro avg	0.49	0.50	0.36	59154
weighted avg	0.49	0.50	0.36	59154

Evaluation

Feature	Model	Precision	Recall	Accuracy
Baseline	-	0.49	0.50	0.51
Bag-Of-Words	Multinomial Naïve Bayes	0.52	0.51	0.51
TF-IDF	Multinomial Naïve Bayes	0.48	0.48	0.48
Pre-trained Embedding	Random Forest	0.50	0.50	0.50
RNN	-	0.51	0.51	0.51
RNN + Pre-trained Embedding	-	0.52	0.52	0.52
BERT	-	0.49	0.50	0.50