EN.601.666 Information Retrieval and Web Agents

Project Writeup

CLASSIFICATION OF 1-DAY RETURN BASED ON EARNING CALL

May 17, 2022

URL: https://github.com/g-1f/irwa_project
Yifei Gu
ygu52@jh.edu

1 Overview

This project aims use earnings calls for stocks to classify 1 day return after the earning call. The assumption is that the earning call for each individual stock contributes to the volatility and return of the stock within a short time period. In this project, some features are extracted from the prepared remarks part of earning call transcript as predictors for classifying 1 day return.

2 Instructions

To run the code:

- `python scrape.py`
  The script scrapes earning call presentation transcript for the desired listing of stocks, and generate a meta-data data frame that include `stocks`, `earnings_url`, `date`, `path` etc...

- `python meta_data.py`
  The script generate additional meta data include `trailing eps`, `trailing pe`, `forward pe`, `market cap` and `target`.

Feature extraction includes:

- `python sentiment_analysis.py` generates sentiment features.
- `python readability_score.py` generates readability features.
- `python svd.py` generates svd matrix.

For building data set and model:

- `python preprocessing.py` pre-process the above features for model.
- `model.ipynb` jupyter-notebook for building classifier.

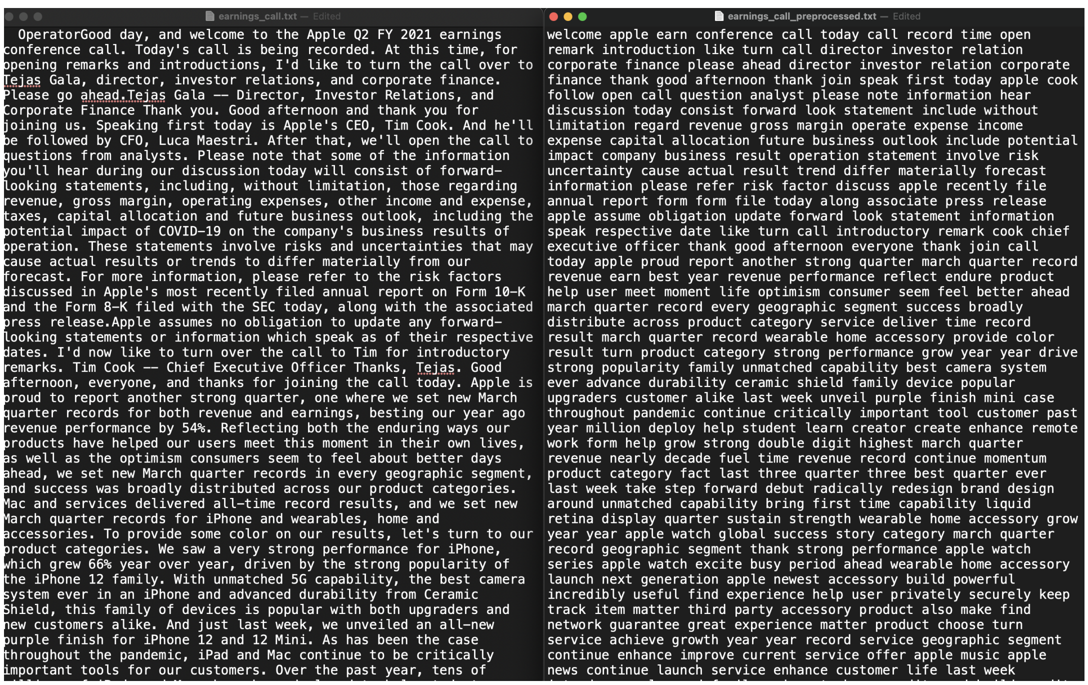
3 About this project

The scraping script scrapes `prepared remarks` for each earning call from The Motey Fool. Additionally, it scrapes `tickers` and their `sectors` from **Nasdaq-100** and **SP-500** Urls. Namely:

- It provides an accessible and easy way to scarpe earning calls. Note that `https://seekingalpha.com/` was previously the go-to option before banning web scrappers.
- It can scrape earning calls based on desired input tickers.

The features extracted consists of follows: sentiment analysis, SVD for document-term matrix, and readability, specifically:

- pre-process earning call transcript with bag-of-words model and filter irrelevant words based on frequency, then find sentiment score:
  - use `VADER` and `TextBlob` to find polarity scores and sentiment score for pre-processed transcript.
  - Loughran McDonald-SA-2020 sentiment word list specialize word sentiment in financial reports. Sentiments score includes `Litigious, Positive, Negative, Constraining, Uncertainty, StrongModal, WeakModal` are calculated with bag-of-words model.



earnings_call.txt — Edited

OperatorGood day, and welcome to the Apple Q2 FY 2021 earnings conference call. Today's call is being recorded. At this time, for opening remarks and introductions, I'd like to turn the call over to Tejas Gala, director, investor relations, and corporate finance. Please go ahead.Tejas Gala -- Director, Investor Relations, and Corporate Finance Thank you. Good afternoon and thank you for joining us. Speaking first today is Apple's CEO, Tim Cook. And he'll be followed by CFO, Luca Maestri. After that, we'll open the call to questions from analysts. Please note that some of the information you'll hear during our discussion today will consist of forward-looking statements, including, without limitation, those regarding revenue, gross margin, operating expenses, other income and expense, taxes, capital allocation and future business outlook, including the potential impact of COVID-19 on the company's business results of operation. These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast. For more information, please refer to the risk factors discussed in Apple's most recently filed annual report on Form 10-K and the Form 8-K filed with the SEC today, along with the associated press release.Apple assumes no obligation to update any forward-looking statements or information which speak as of their respective dates. I'd now like to turn over the call to Tim for introductory remarks. Tim Cook -- Chief Executive Officer Thanks, Tejas. Good afternoon, everyone, and thanks for joining the call today. Apple is proud to report another strong quarter, one where we set new March quarter records for both revenue and earnings, besting our year ago revenue performance by 54%. Reflecting both the enduring ways our products have helped our users meet this moment in their own lives, as well as the optimism consumers seem to feel about better days ahead, we set new March quarter records in every geographic segment, and success was broadly distributed across our product categories. Mac and services delivered all-time record results, and we set new March quarter records for iPhone and wearables, home and accessories. To provide some color on our results, let's turn to our product categories. We saw a very strong performance for iPhone, which grew 66% year over year, driven by the strong popularity of the iPhone 12 family. With unmatched 5G capability, the best camera system ever in an iPhone and advanced durability from Ceramic Shield, this family of devices is popular with both upgraders and new customers alike. And just last week, we unveiled an all-new purple finish for iPhone 12 and 12 Mini. As has been the case throughout the pandemic, iPad and Mac continue to be critically important tools for our customers. Over the past year, tens of

earnings_call_preprocessed.txt — Edited

welcome apple earn conference call today call record time open remark introduction like turn call director investor relation corporate finance please ahead director investor relation corporate finance thank good afternoon thank join speak first today apple cook follow open call question analyst please note information hear discussion today consist forward look statement include without limitation regard revenue gross margin operate expense income expense capital allocation future business outlook include potential impact company business result operation statement involve risk uncertainty cause actual result trend differ materially forecast information please refer risk factor discuss apple recently file annual report form form file today along associate press release apple assume obligation update forward look statement information speak respective date like turn call introductory remark cook chief executive officer thank good afternoon everyone thank join call today apple proud report another strong quarter march quarter record revenue earn best year revenue performance reflect endure product help user meet moment life optimism consumer seem feel better ahead march quarter record every geographic segment success broadly distribute across product category service deliver time record result march quarter record wearable home accessory provide color result turn product category strong performance grow year year drive strong popularity family unmatched capability best camera system ever advance durability ceramic shield family device popular upgraders customer alike last week unveil purple finish mini case throughout pandemic continue critically important tool customer past year million deploy help student learn creator create enhance remote work form help grow strong double digit highest march quarter revenue nearly decade fuel time revenue record continue momentum product category fact last three quarter three best quarter ever last week take step forward debut radically redesign brand design around unmatched capability bring first time capability liquid retina display quarter sustain strength wearable home accessory grow year year apple watch global success story category march quarter record geographic segment thank strong performance apple watch series apple watch excite busy period ahead wearable home accessory launch next generation apple newest accessory build powerful incredibly useful find experience help user privately securely keep track item matter third party accessory product also make find network guarantee great experience matter product choose turn service achieve growth year year record service geographic segment continue enhance improve current service offer apple music apple news continue launch service enhance customer life last week

Figure 1: original transcript vs pre-processed

- Dimension reduction with SVD:
  - calculate Tf-idf document term matrix
  - get first 4 dimensions of term representation for all earning call documents.

Path	svd_1	svd_2	svd_3	svd_4
/Users/gyf/Desktop/earning call analysis/CTAS/2022/Q1/	0.02294622101477207	-0.006085854424916311	-0.01023835806864517	0.002366072187086217
/Users/gyf/Desktop/earning call analysis/CTAS/2022/Q2/	0.02639244296472701	-0.007086554000573436	-0.0201888972141872	0.00041388349949480634
/Users/gyf/Desktop/earning call analysis/CTAS/2021/Q4/	0.02226903627485426	-0.013167334330072942	-0.013014858963548712	-0.008472851225689857
/Users/gyf/Desktop/earning call analysis/WELL/2022/Q1/	0.024724563381259505	0.0036650555395668825	0.008087858157894154	-0.017491698898400816
/Users/gyf/Desktop/earning call analysis/WELL/2021/Q1/	0.02252081753848908	0.020919541472658072	0.010943298884167653	-0.01148572156035765
/Users/gyf/Desktop/earning call analysis/WELL/2021/Q2/	0.022914525232205173	0.023081347778520698	0.004621070597237323	-0.013607580161141981
/Users/gyf/Desktop/earning call analysis/WELL/2021/Q4/	0.025712780297647297	0.010495126761064085	-0.005772120660678381	-0.011291433904716508
/Users/gyf/Desktop/earning call analysis/WELL/2021/Q3/	0.023984268955591248	0.011132899240735898	0.005480514097485518	-0.014943784121480006
/Users/gyf/Desktop/earning call analysis/VZ/2021/Q1/	0.028470793637884923	-9.400704056429323e-05	-0.007253442604846614	-0.01330362633826202
/Users/gyf/Desktop/earning call analysis/VZ/2021/Q2/	0.027109857451502135	0.00020172239448817806	-0.012030946938085733	-0.015054920561068727
/Users/gyf/Desktop/earning call analysis/VZ/2021/Q3/	0.025629365201837343	-0.011872816197960082	-0.01288586738144202	-0.014124457073074805
/Users/gyf/Desktop/earning call analysis/AMZN/2022/Q1/	0.024693543577156276	-0.010465286384873302	0.008942525247202572	-0.012104213374343649
/Users/gyf/Desktop/earning call analysis/AMZN/2021/Q1/	0.02185810716281909	-0.01999815973434758	0.005315836958206393	-0.02166827490877166
/Users/gyf/Desktop/earning call analysis/AMZN/2021/Q2/	0.019047389012951287	-0.03282012022386174	-0.0006276875297257448	-0.017787465119081976
/Users/gyf/Desktop/earning call analysis/AMZN/2021/Q4/	0.02149037932892842	-0.01822544664246636	0.0035222676854189225	-0.022652988504201582
/Users/gyf/Desktop/earning call analysis/AMZN/2021/Q3/	0.018560174522906994	-0.037425243933493044	0.017806703954004983	-0.02574700129428252
/Users/gyf/Desktop/earning call analysis/DXC/2022/Q1/	0.029924860868662837	-0.008624185175184087	-0.004984072763775243	-0.02241074796859939
/Users/gyf/Desktop/earning call analysis/DXC/2022/Q2/	0.02889501894828952	-0.011847035900338306	0.0037711696624677105	-0.022609371587081616
/Users/gyf/Desktop/earning call analysis/DXC/2022/Q3/	0.028451076657210302	-0.014629849111143232	-0.004265290098509432	-0.02357316544304251
/Users/gyf/Desktop/earning call analysis/DXC/2021/Q4/	0.028877965907042915	-0.018635602077321272	-0.0031963671951644955	-0.025190926577998863
/Users/gyf/Desktop/earning call analysis/CNP/2021/Q1/	0.02019853291408187	0.019456537905516828	0.05464099074617022	-0.004680968953919976
/Users/gyf/Desktop/earning call analysis/CNP/2021/Q2/	0.021631261706554585	0.010720581155711874	0.04917769150886834	-0.010424344891378097
/Users/gyf/Desktop/earning call analysis/CNP/2021/Q3/	0.021706163870583866	0.020489717469437153	0.04595380844674584	-0.007841807880884129

Figure 2: SVD demo

- Readability score:
  - Various readability metrics were used for feature generation, includes:
    - Flesch-Kincaid Reading Ease
    - Dale Chall Readability
    - Gunning Fog
    - SPACHE

In short, the project:

- **calculate sentiment** of earning calls with bag-of-words models
- **reduce dimension** of earning calls with tf-idf weighting.
- **calculate readability** of earning calls.
- **predict** whether to **hold, buy, sell** a stock given earning call. (more on evaluation part)

4 Limitations

There are multiple limitations of this project, namely:

- It only uses earning call features as predictors, in the real world, earning call & earnings is only one component of that moves the market. There are thousands of other things that could move the market for example: for the exact same earning call under different interest rates, the return could be significantly different. Thus, a lot of other features could potentially be added.

- The target is the categorized 1 day return. The stock market is highly volatile under different market conditions and thus irrational. The 1 day return suffers from noise significantly, some better metrics like 3-day, 1-week returns could be used as well. (I tried 5 days and the performance was not good).

- I only used bad-of-words model for sentiment analysis. A more comprehensive, n-gram model that tokenize financial term well should be used to improve the accuracy of sentiment scores. This is the same case for SVD.

- During this project, I noticed difficulty of getting some financial data, for example, it is very hard to get estimate eps, actual eps for every stock on the earning date, which is one of the most important features that sadly I could not scrape. So lack of data is also one aspect.

5 Evaluation

The target is the 1-day return of stock categorized as follows:

- 0 Hold: between -3% and +3%,

- 1 Buy: above +3%,

- -1 Sell: fall below -3%,

Multiple models are compared, including LGBM, KNN and decision tree. The model recommended is Light GBM classifier. Light GBM splits the tree leaf-wise with the best fit whereas other boosting algorithms split the tree depth-wise or level-wise rather than leaf-wise. With decision tree as a baseline, the performance is as follows: And feature importance are as follows:

```
Decision Tree accuracy score: 0.6794871794871795
KNN accuracy score: 0.764102564102564
LightGBM Model accuracy score: 0.7743589743589744
```

Figure 3: baseline Decision Tree vs KNN vs LGBM

	cols	fea_imp
9	svd_3	498
8	svd_2	485
18	sent_TB	479
10	svd_4	473
16	neu_vader	445
7	svd_1	440
24	Negative	436
13	d_c	432
25	Constraining	432
6	Market_cap	403
22	Uncertainty	397
15	neg_vader	386
19	Litigious	385
20	Positive	363
4	Trailing_pe	361
3	Forward_pe	342
11	f_k	330
17	pos_vader	306
12	g_f	301
5	Trailing_eps	275
14	spache	275
23	StrongModal	210
21	WeakModal	165
2	Quarter	148
0	Sector	142
1	Year	91

Figure 4: Feature importance

6 Conclusion

From the feature importance , one can observe that some dimensions in svd serves important role in the model, and neutral sentiment score from `VADER`, and negative words, constraining words from Loughran McDonald-SA-2020 sentiment word list also affect the classification.
Overall, this project demonstrates that some alternative data could be used as features for all kinds of different areas.