

---

# Predicting price of Bitcoin using sentiment

Weiqi Tong  
Mao Guan  
Carlo Provinciali  
David Lee

---

---

# Introduction

---

# The Rise of Bitcoin



# Bitcoin Markets vs Stock Market



- Prices are much more volatile
- No centralized source of information
- Technology is not well understood by the public

---

# Literature

---

---

# Why Sentiment?

Johan Bollen, Huina Mao: “Twitter mood predicts the stock market”

- Sentiment of Tweets predicts movement the Dow Jones industrial average by 3-4 days

Sul, H. K., Dennis, A. R. and Yuan, L. (2017), Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns

- Dissemination speed and visibility of a post can enhance predictions of stock market prices.
-

---

# Why Reddit?

Phillips and Gorse: "'Predicting cryptocurrency price bubbles using social media data and epidemic modelling."

- discussion boards within Reddit do a better job at representing communities that share a common interest

Maurer B, Nelms TC, Swartz L. "When perhaps the real problem is money itself!": the practical materiality of Bitcoin (2013).

- Many users rely on information share in Web community to make decisions about trading cryptocurrencies

Mai et. al: "The Impacts of Social Media on Bitcoin Performance" (April 30, 2016)

- When aggregated at the intraday level, the sentiments on forum messages are more telling indicators of future Bitcoin returns than are Twitter messages
-

---

# Why Socialsent?

Gilbert, C J H E. "Vader: A parsimonious rule-based model for sentiment analysis of social media text" (2014)

- Valence Aware Dictionary for Sentiment Reasoning is a gold-standard sentiment lexicon that is especially attuned to microblog-like contexts. VADER retains (and even improves on) the benefits of traditional sentiment lexicons like LIWC.

Kim Y B, Kim J G, Kim W, et al. (2016): "Predicting fluctuations in cryptocurrency transactions based on user comments and replies"

- Use Vader to predict price movements and number of transactions of cryptocurrencies based on user comments on online cryptocurrency communities such as Bitcoin Talk

Hamilton, William L., et al. (2016) "Inducing domain-specific sentiment lexicons from unlabeled corpora."

- Domain specific lexicon can improve sentiment analysis tasks
-





---

# Dataset

---



---

# /r/BitcoinMarkets

  **negative100percent** 4 points 40 minutes ago



What's the sentiment around here? Optimistic or no?

[permalink](#) [embed](#) [save](#) [report](#) [give gold](#) [reply](#)

  **mikeyvegas17** Bearish 11 points 28 minutes ago



Irrationally Bullish AF

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **jarederaaj** Long-term Holder 1 point 12 minutes ago



Someone should make a frequency chart that measures sentiment for top-level comments. I think it's more split than you're describing here.

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **az9393** Long-term Holder 0 points 17 minutes ago



Oh come on it's not that irrational :) maybe overly optimistic. You can't deny that this thing had plenty of chances to dump further since last visit to 6k but didn't.

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **Simres** 5 points 32 minutes ago

Sentiment is neutral atm but **confused AF**



[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **chrisgilesphoto** Bullish 2 points 39 minutes ago

Cool username.

Optimistic

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **a\_cool\_goddamn\_name** 9 points 36 minutes ago

thanks

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

---

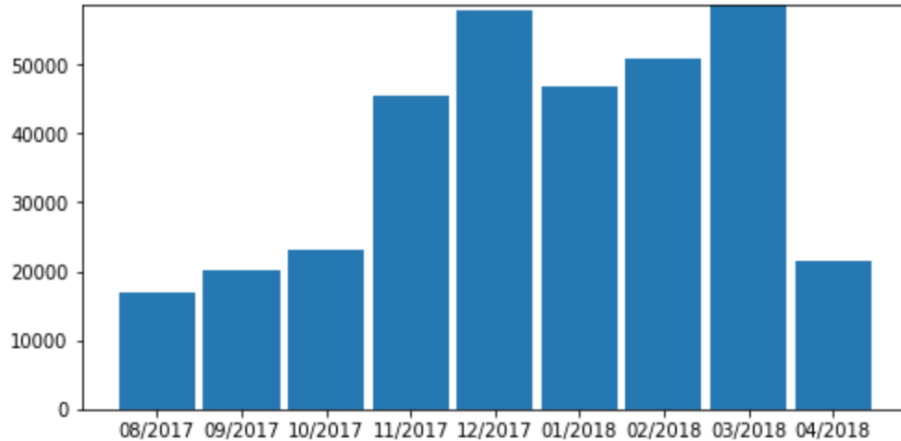
# Data Collection

1. “Daily Discussion” posts from August 2017 to April 2018
  2. Each comment contains:
    - a. Body of the comment
    - b. ID of the comment’s parent
    - c. Date of the daily discussion
    - d. Created at date
    - e. Number of upvotes/downvotes
    - f. Author’s opinion
-

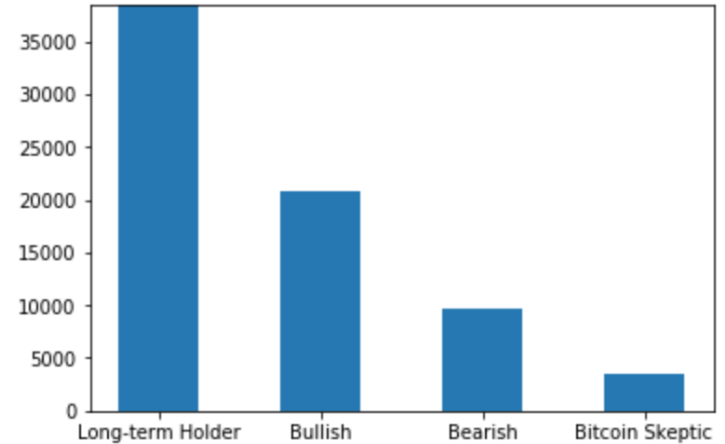
---

# Resulting Dataset

1. Final size of the dataset: **337,190** comments
2. Percentage of comments with flaired authors: **21.5%**



Number of comments by month



Distribution of opinions among flaired authors

---



---

# Feature Engineering

---

---

# Feature Engineering

Original Dataset (each row is each text posted)

- Text body & Datetime
- Sentiment scores (from VADER & SocialSent)
- Other features (parent id, opinion, etc....)



Aggregated Feature Dataset (each row is a timestamp interval)

- Timestamp Interval
  - Sentiment scores averaged during this interval
  - Statistical features
-

---

# Feature Engineering

## Sentiment Features (24 features)

1. Sentiment score (positive, negative, neutral, compound) directly averaged
2. Sentiment score averaged weighted on **number of votes**
3. Sentiment score averaged weighted on **number of comments replied**

TWO Lexicon base: VADER, SocialSent

## Statistical Features (8 features) + Past Price Features (1 feature)

- Proportion of author opinion type (Bullish, Bearish, Long-term, Skeptical, None)
  - Average number of comments, number of words each comment, number of replies
  - Price move during this interval
-



---

# Feature Engineering

  **mikeyvegas17** Bearish 10 points 18 minutes ago\*

That's one hell of a daily candle. That's 6/7 days and 11/15 days of closing up.

I know bitcoin can be irrational, but going up 44% in two weeks seems a bit unsustainable and pullback is in order.

[permalink](#) [embed](#) [save](#) [report](#) [give gold](#) [reply](#)

  **japanese\_\_cat** 1 point 6 minutes ago

What is a rational move by probably still extremely undervalued asset? Which on the other side made millionaires and billionaires from geeks who have no experience with money so they randomly dump it?



Almost anything is possible. The only irrational behavior is here to expect something like moderate moves, which is something that you call rational for no reason.

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **drdixie** 4 points 15 minutes ago

Going down 60% in 4 months could be considered irrational.

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

  **mikeyvegas17** Bearish 1 point 6 minutes ago

Agreed. Also going up 550% the 4 months before that was pretty crazy as well.

9/17/2017 3606.28 3664.81 3445.64 3582.88 12/17/2017 19475.8 20089 18974.1

[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [give gold](#) [reply](#)

---

# Feature Engineering

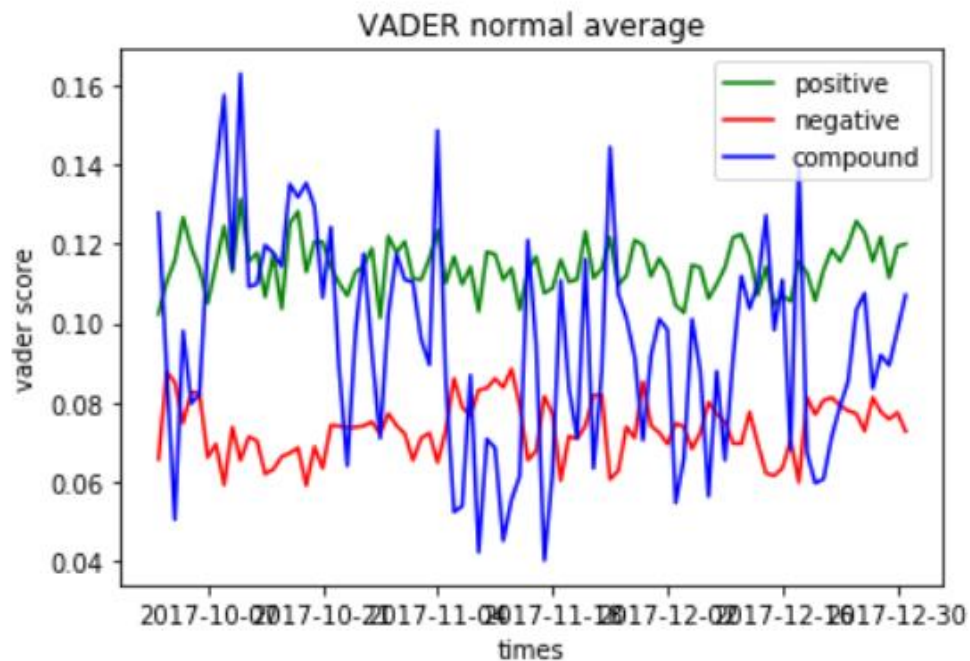
A demo sample of dataset (averaged by 24 hour):

	vader_negative	vader_positive	vader_compound	vader_neutral	vader_votes_compound							
daily_discussion_date					Long-term Holder	Bullish	Bearish	None	Bitcoin Skeptic	num_comments	num_child	num_word
2017-08-09	0.074451	0.101215	0.071324	0.822831								
2017-08-10	0.066139	0.107833	0.093359	0.823644								
2017-08-11	0.074532	0.112947	0.097955	0.812522								
2017-08-12	0.074787	0.113487	0.093948	0.809091	0.184569	0.095310	0.015129	0.701967	0.003026	661	0.912254	193.977307
2017-08-13	0.058828	0.133918	0.141645	0.807245	0.141148	0.090909	0.011962	0.729665	0.026316	418	0.882775	145.076554
2017-08-14	0.064308	0.123766	0.119916	0.811927	0.190118	0.116004	0.009667	0.678840	0.005371	931	0.832438	158.481203
2017-08-15	0.077145	0.112664	0.094633	0.810193	0.158033	0.066725	0.007902	0.764706	0.002634	1139	0.768218	128.840211
2017-08-16	0.080108	0.114494	0.083505	0.805413	0.141066	0.075235	0.023511	0.757053	0.003135	638	0.808777	145.512539
2017-08-17	0.072440	0.107712	0.084955	0.819859	0.173561	0.086331	0.025180	0.705935	0.008993	1112	0.841727	134.631294

---

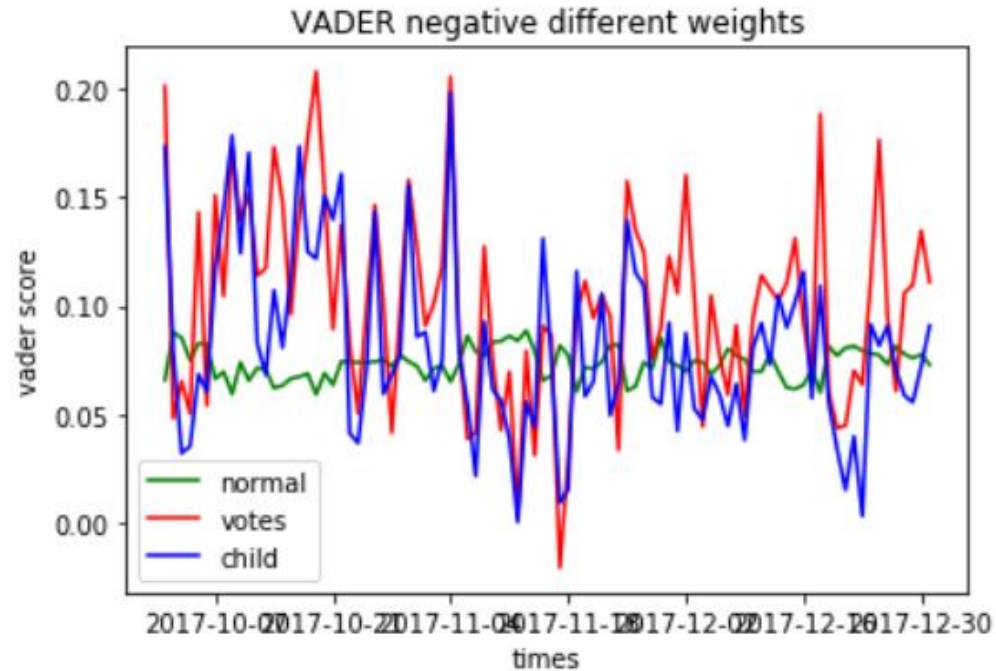
---

# Feature Engineering

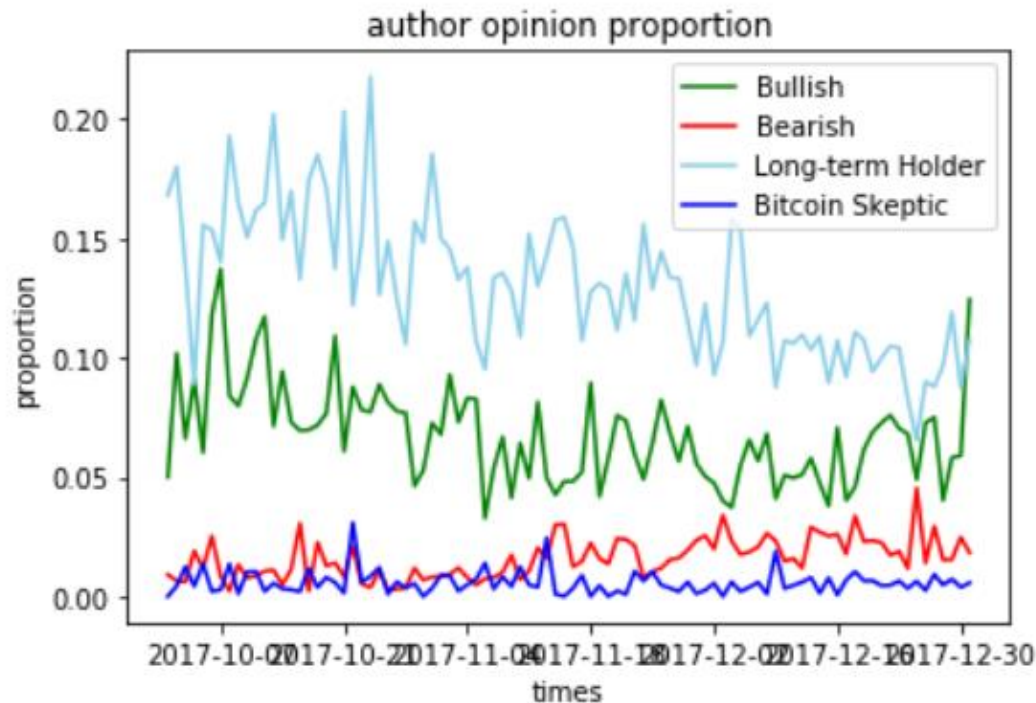


---

# Feature Engineering



# Feature Engineering



---

# Labeling

1. Window Setting: The look back window to calculate the volatility.

1. Label Rule:

Price change positive/negative:

- a. price change beyond one deviation.
- b. price change within one deviation.

17473	17473	2018-03-30 14:00	6743.00	0.002637	6725.22	small rise
17474	17474	2018-03-30 15:00	6820.01	0.011292	6743.00	small rise
17475	17475	2018-03-30 16:00	6860.00	0.005829	6820.01	small rise
17476	17476	2018-03-30 17:00	6916.01	0.008099	6860.00	small rise
17477	17477	2018-03-30 18:00	6788.00	-0.018858	6916.01	big drop
17478	17478	2018-03-30 19:00	6848.01	0.008763	6788.00	small rise
17479	17479	2018-03-30 20:00	6993.72	0.020834	6848.01	big rise
17480	17480	2018-03-30 21:00	7099.01	0.014832	6993.72	small rise
17481	17481	2018-03-30 22:00	7047.87	-0.007256	7099.01	small drop
17482	17482	2018-03-30 23:00	6969.01	-0.011316	7047.87	small drop
17483	17483	2018-03-31 00:00	7020.00	0.007264	6969.01	small rise
17484	17484	2018-03-31 01:00	6920.17	-0.014426	7020.00	small drop
17485	17485	2018-03-31 02:00	6818.26	-0.014947	6920.17	small drop
17486	17486	2018-03-31 03:00	6824.10	0.000856	6818.26	small rise
17487	17487	2018-03-31 04:00	6905.94	0.011851	6824.10	small rise
17488	17488	2018-03-31 05:00	6999.99	0.013436	6905.94	small rise
17489	17489	2018-03-31 06:00	7082.24	0.011614	6999.99	small rise
17490	17490	2018-03-31 07:00	7095.01	0.001800	7082.24	small rise
17491	17491	2018-03-31 08:00	7181.73	0.012075	7095.01	small rise
17492	17492	2018-03-31 09:00	7041.01	-0.019986	7181.73	big drop
17493	17493	2018-03-31 10:00	7068.27	0.003857	7041.01	small rise

---

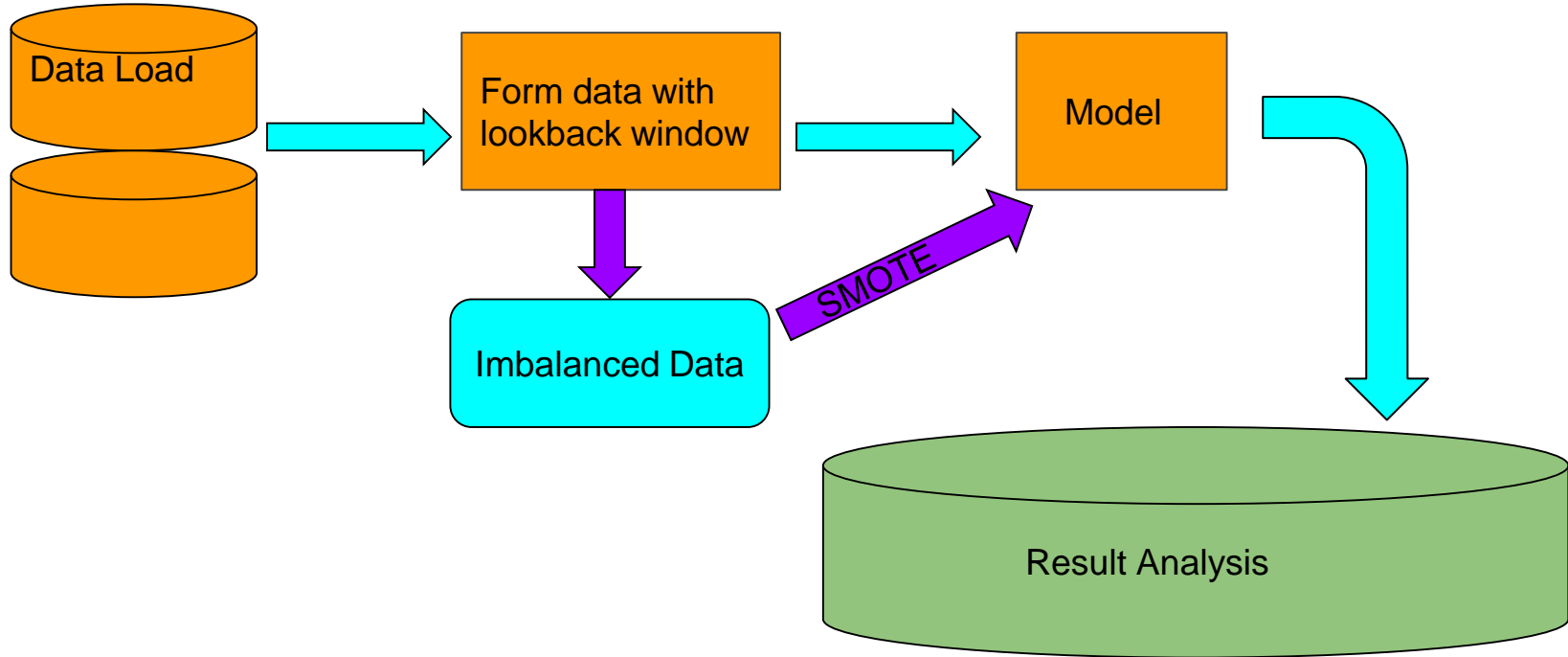
---

# Data Analysis

---

---

# Framework





---

## Parameters:

Train/Test:  
7/3

Window:  
[1,2,5,10,  
24,36,48]

MLP Layers:  
[24,4]

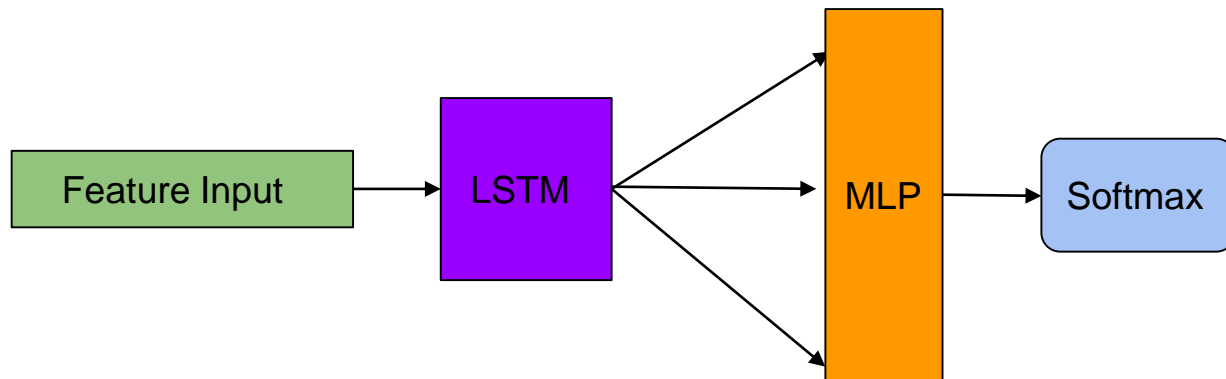
epochs=500,

batch\_size=64

---

# Evaluation Model

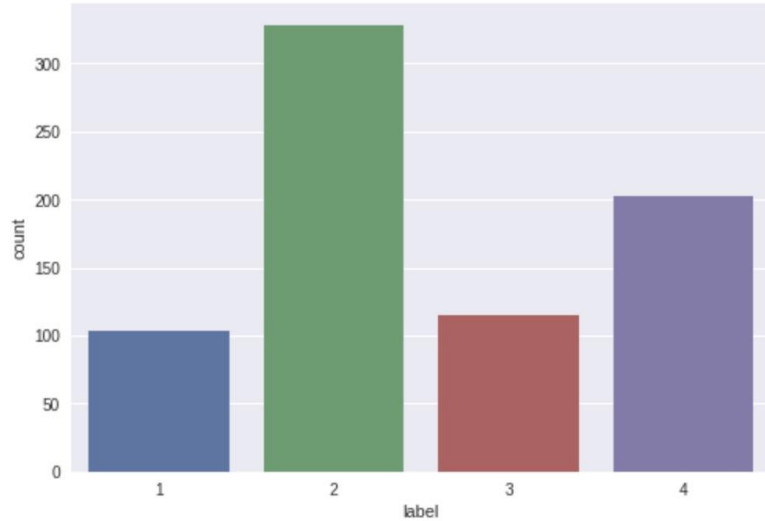
```
model.add(LSTM(128, input_shape = (features.shape[1],features.shape[2]),  
              return_sequences=False,dropout=0.2, recurrent_dropout=0.2 ))  
  
model.add(Dense(24))  
  
model.add(Dense(4))  
model.add(Activation('softmax'))
```



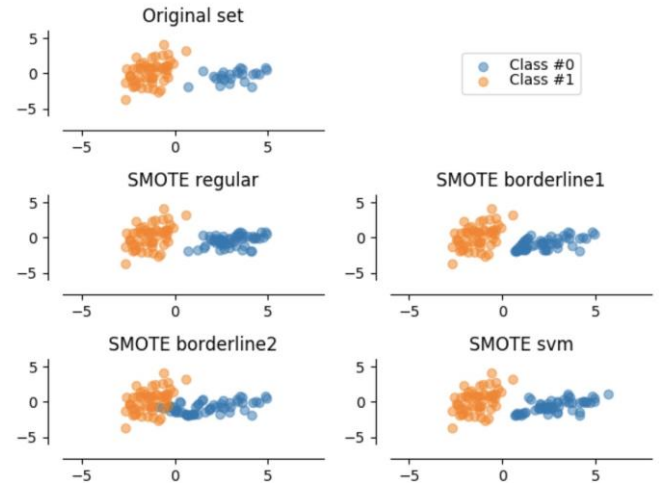
---

# Data Preprocessing

Imbalanced Data:

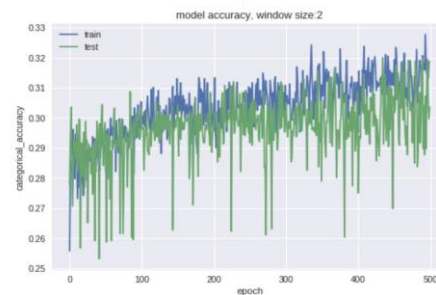
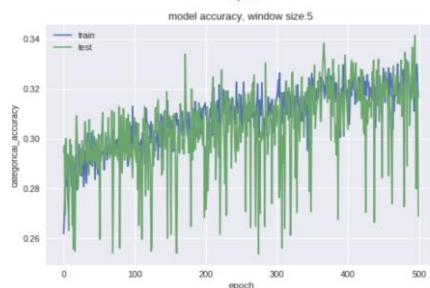
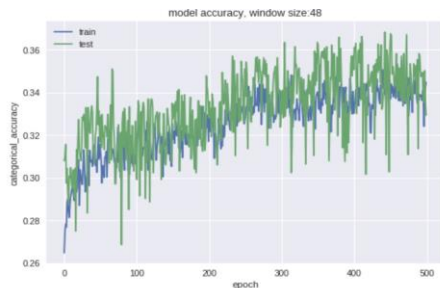
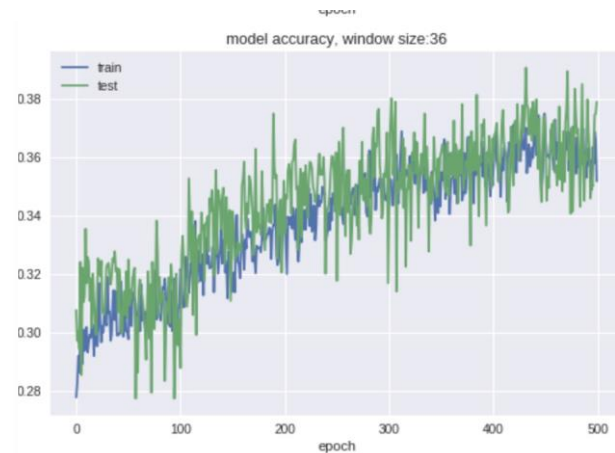
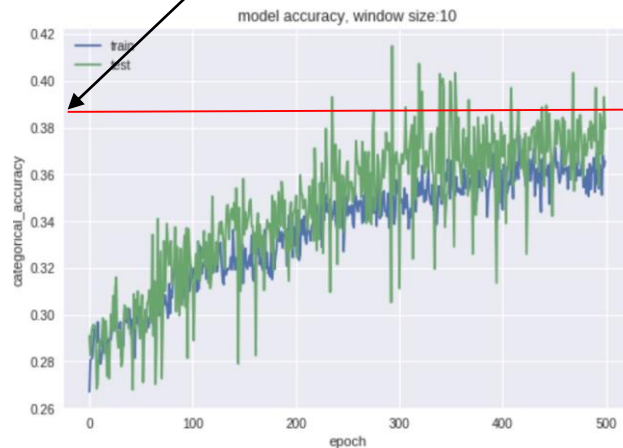


Solution: SMOTE



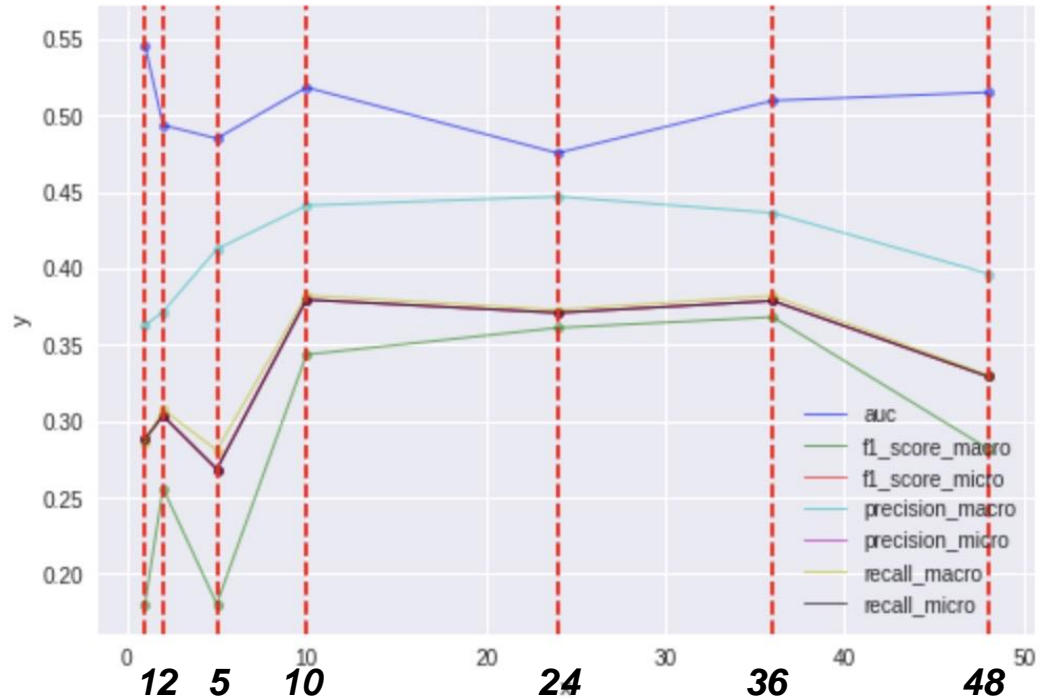
Accuracy:0.38  
Benchmark:0.25

# Result



# Result

1. Sentiment Score are helpful for predicting the price movement.
1. Period 10 (hours) is the best.



---

---

# What to Improve

1. Try sentiment embedding (word2sentiment)
  2. Try different time intervals
  3. More feature engineering (tree of comments replied)
  4. Augment more data period and other sources
-

---

# Q&A

---