Sentiment Analysis for Financial Market Signals

Yuejiao Qiu, Ruoting Shen,
Yuening Wang, Xiana Zhang

GEORGETOWN UNIVERSITY

Agenda

1. Background
2. Dataset
3. Model
    a. BOW
    b. BERT
    c. BERT transfer learning
4. Discussions
5. Improvement

Background

How many of you have heard about WallStreetBets?

- Short Squeeze: Condition that triggers a stock to increase rapidly

GEORGETOWN UNIVERSITY

Twitter

- Influencer
- Language
- Slangs
- Abbreviations

Dataset

Dataset name: Sentiment 140

targets: popularity of the tweet
ids: id of the tweet
date: date of the tweet
flag: query (lyx)
user: user of the tweet
text: content of the tweet

Sentiment	Count
Positive	800,000
Negative	800,000

Source: Kaggle.com

Dataset name: Financial sentiment

sentence: Tweets
sentiment: positive/negative/neutral

Sentiment	Count
Positive	1,852
Neutral	3,130
Negative	860

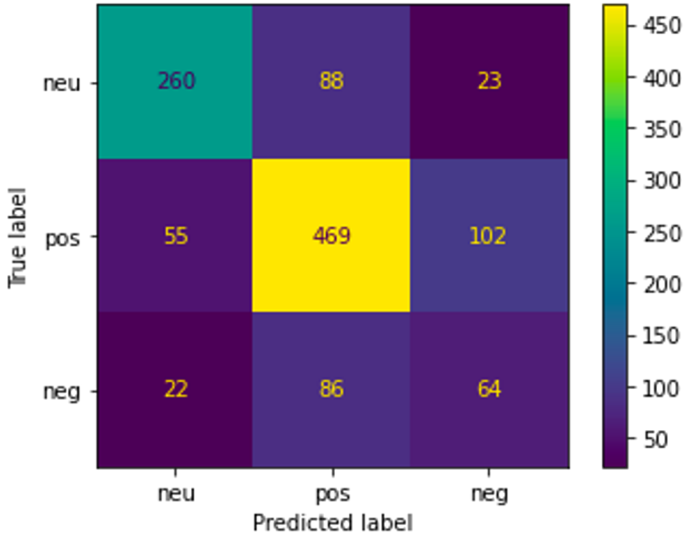
Bag-of-words (BOW)

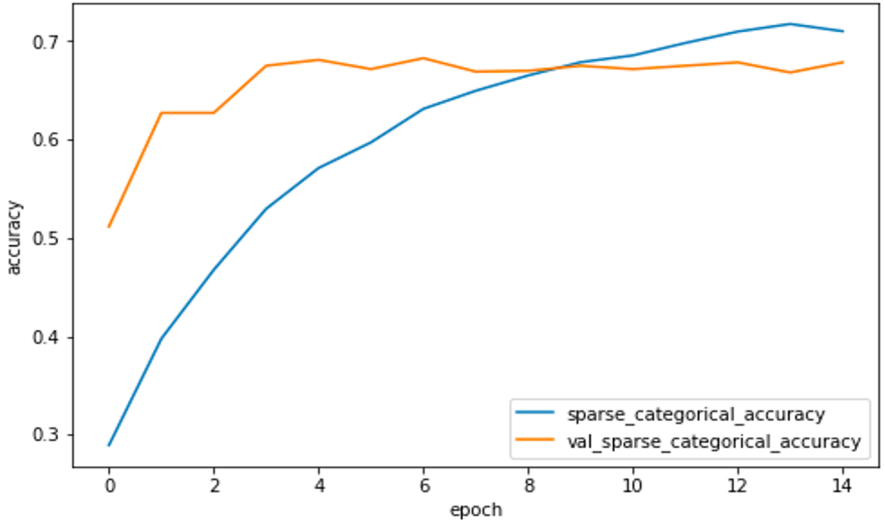
- 4 dense layers
- 3 dropout layers
  - Prevent overfitting
- 772,565 trainable parameters

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 6000)]	0
dense (Dense)	(None, 128)	768128
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 32)	4128
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 8)	264
dropout_2 (Dropout)	(None, 8)	0
dense_3 (Dense)	(None, 5)	45

Model Performance

- F1 score: 0.681
- Accuracy: 67.84%





GEORGETOWN UNIVERSITY

BERT without Fine Tuning

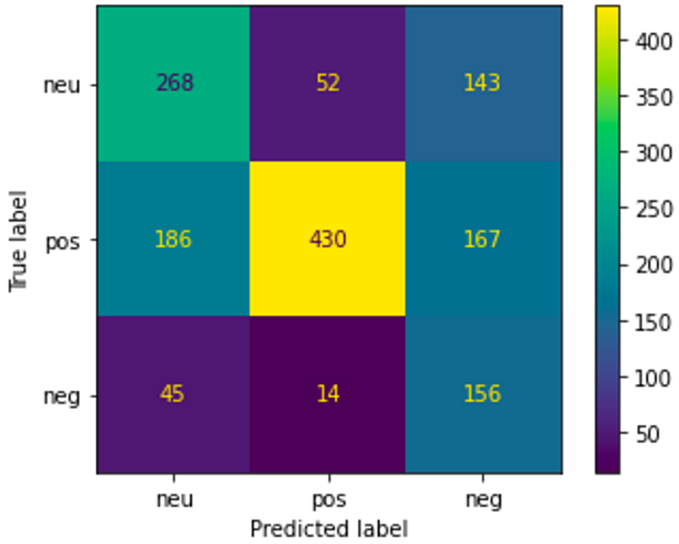
- 4 dense layers
- No dropout layers
- Trainable parameters: 246,917

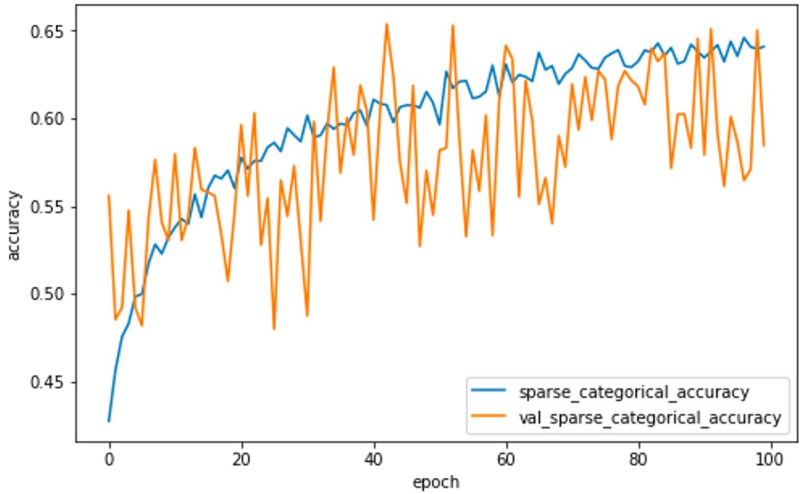
Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
token_type_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (TFBertMainLayer)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	108310272	['input_ids[0][0]', 'token_type_ids[0][0]', 'attention_mask[0][0]']
dense_4 (Dense)	(None, 256)	196864	['bert[1][1]']
dense_5 (Dense)	(None, 128)	32896	['dense_4[0][0]']
dense_6 (Dense)	(None, 128)	16512	['dense_5[0][0]']
dense_7 (Dense)	(None, 5)	645	['dense_6[0][0]']

GEORGETOWN UNIVERSITY

Model Performance

- F1 score: 0.604
- Accuracy: 64.59%





GEORGETOWN UNIVERSITY

BERT with Fine Tuning

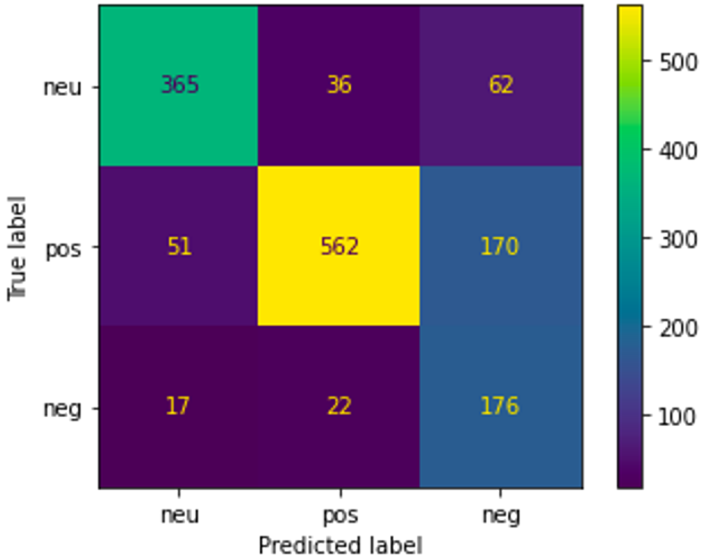
- 4 dense layers
- 3 dropout layers
  - Prevent overfitting
- Trainable parameters: 108,557,189

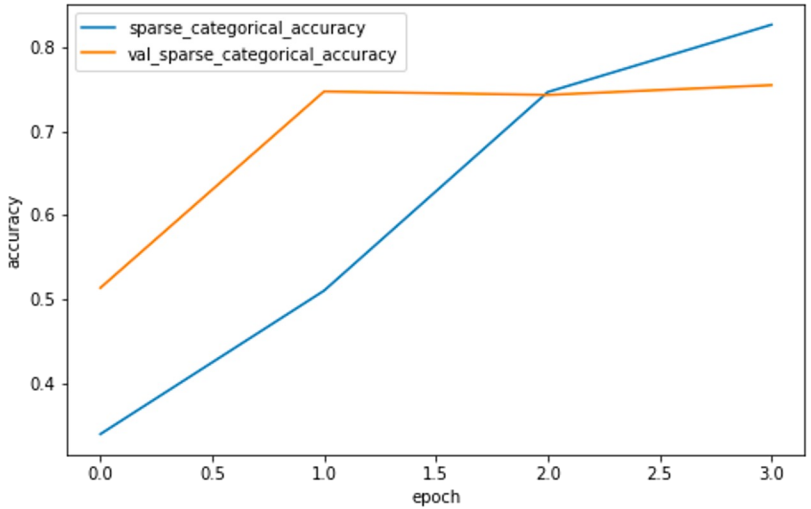
Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
token_type_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
tf_bert_model_1 (TFBertModel)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	108310272	['input_ids[0][0]', 'token_type_ids[0][0]', 'attention_mask[0][0]']
dense_4 (Dense)	(None, 256)	196864	['tf_bert_model_1[0][1]']
dropout_77 (Dropout)	(None, 256)	0	['dense_4[0][0]']
dense_5 (Dense)	(None, 128)	32896	['dropout_77[0][0]']
dropout_78 (Dropout)	(None, 128)	0	['dense_5[0][0]']
dense_6 (Dense)	(None, 128)	16512	['dropout_78[0][0]']
dropout_79 (Dropout)	(None, 128)	0	['dense_6[0][0]']
dense_7 (Dense)	(None, 5)	645	['dropout_79[0][0]']

GEORGETOWN UNIVERSITY

Model Performance

- F1 score: 0.771
- Accuracy: 75.50%





GEORGETOWN UNIVERSITY

BERT Transfer Learning

Prep: Fine-tune BERT with Sentiment140

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
token_type_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (TFBertMainLayer)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	108310272	['input_ids[0][0]', 'token_type_ids[0][0]', 'attention_mask[0][0]']
dense (Dense)	(None, 4)	3076	['bert[0][1]']

GEORGETOWN UNIVERSITY

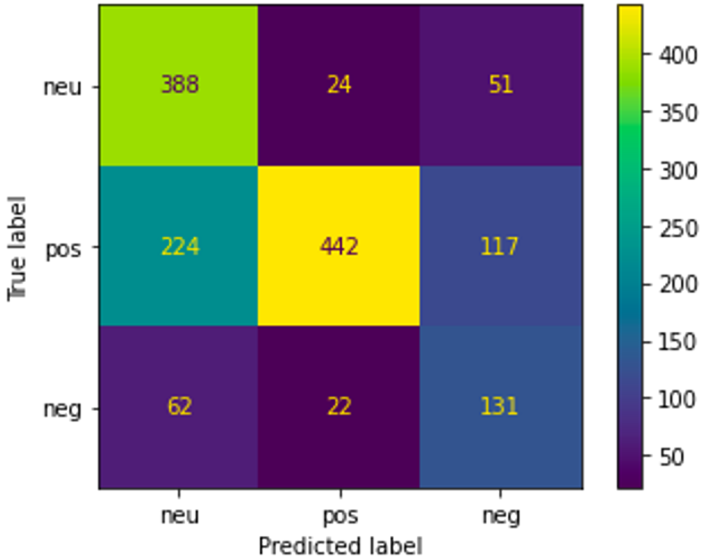
Pre-tuned BERT without Fine Tuning

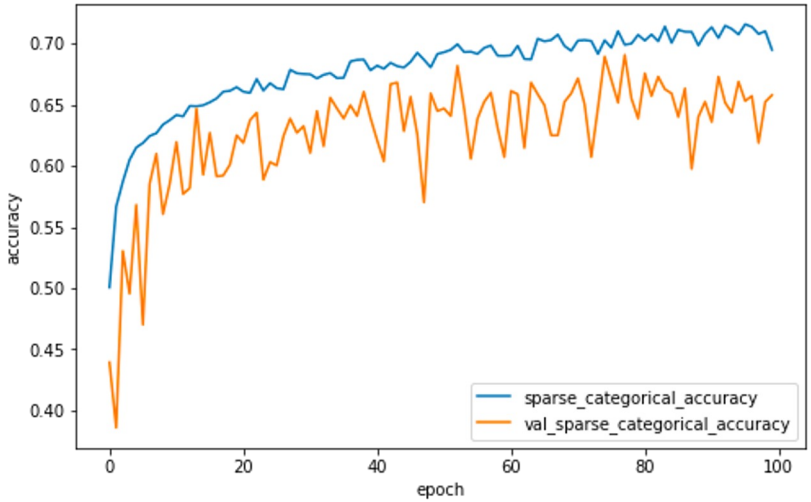
- 4 dense layers
- No dropout layers
- Sentiment140-pretrained BERT layer
- Trainable parameters: 246,917

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
token_type_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (Custom>TFBertMainLayer)	{'pooler_output': (None, 768), 'last_hidden_state ': (None, 100, 768) }	108310272	['input_ids[0][0]', 'token_type_ids[0][0]', 'attention_mask[0][0]']
dense (Dense)	(None, 256)	196864	['bert[0][1]']
dense_1 (Dense)	(None, 128)	32896	['dense[0][0]']
dense_2 (Dense)	(None, 128)	16512	['dense_1[0][0]']
dense_3 (Dense)	(None, 5)	645	['dense_2[0][0]']

Model Performance

- F1 score: 0.664
- Accuracy: 69.06%





GEORGETOWN UNIVERSITY

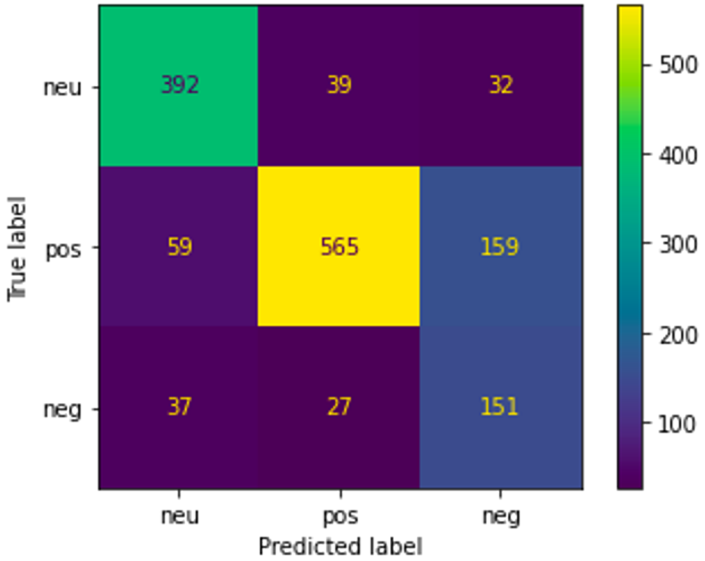
Pre-tuned BERT with Fine Tuning

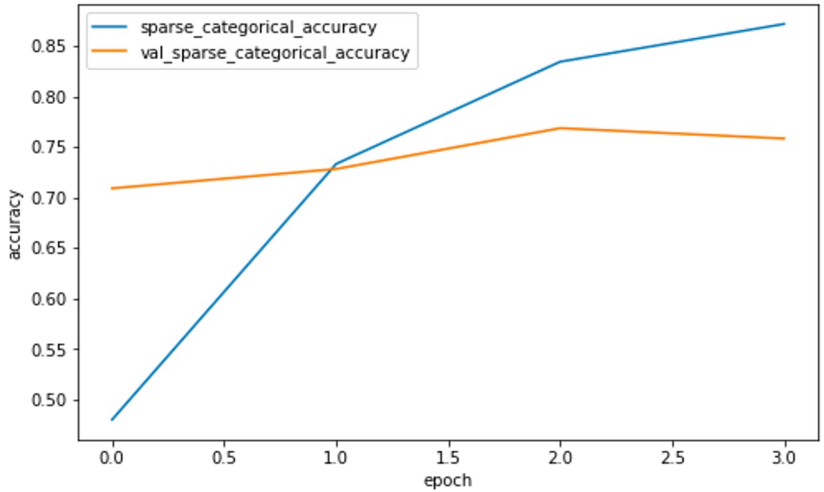
- 4 dense layers
- 3 dropout layer
- Unfreezed Sentiment140-pretrained BERT layer
- Trainable parameters: 108,557,189

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
token_type_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (Custom>TFBertMainLayer)	{'last_hidden_state': (None, 100, 768), 'pooler_output': (None, 768)}	108310272	['input_ids[0][0]', 'token_type_ids[0][0]', 'attention_mask[0][0]']
dense (Dense)	(None, 256)	196864	['bert[0][1]']
dropout (Dropout)	(None, 256)	0	['dense[0][0]']
dense_1 (Dense)	(None, 128)	32896	['dropout[0][0]']
dropout_1 (Dropout)	(None, 128)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 128)	16512	['dropout_1[0][0]']
dropout_2 (Dropout)	(None, 128)	0	['dense_2[0][0]']
dense_3 (Dense)	(None, 5)	645	['dropout_2[0][0]']

Model Performance

- F1 score: 0.775
- Accuracy: 76.52%





GEORGETOWN UNIVERSITY

Model Comparison

Dataset	Training Size	Model	Fine Tune	Hidden Dense	Dropout	#Parameters	#Epochs	Best Accuracy	F1 Score
FinTweets	7512	BOW	N/A	(128, 32, 8)	0.75	772,565	15	67.84%	0.681
FinTweets	7041	BERT	N	(256,128,128)	N/A	246,917	100	64.59%	0.643
FinTweets	7041	BERT	Y	(256,128,128)	0.5	108,557,189	4	75.50%	0.771
Sentiment140	120,000	BERT	Y	None	N/A	108,313,348	1	Pre-tuned BERT	
FinTweets	7041	Pre-tuned BERT	N	(256,128,128)	N/A	246,917	100	69.06%	0.664
FinTweets	7041	Pre-tuned BERT	Y	(256,128,128)	0.5	108,557,189	4	76.52%	0.775

Evaluation with GME and Reddit Data

- Data
  - Time frame: 2021/1/1 - 2022/8/31
  - 94,298 GME-related Reddit Submissions from r/GME, r/wallstreetbets
  - GME historical daily price
- Predictions

Sentiment	Count
Positive	61,656
Neutral	18,765
Negative	13.877

Co-movement with Stock Price

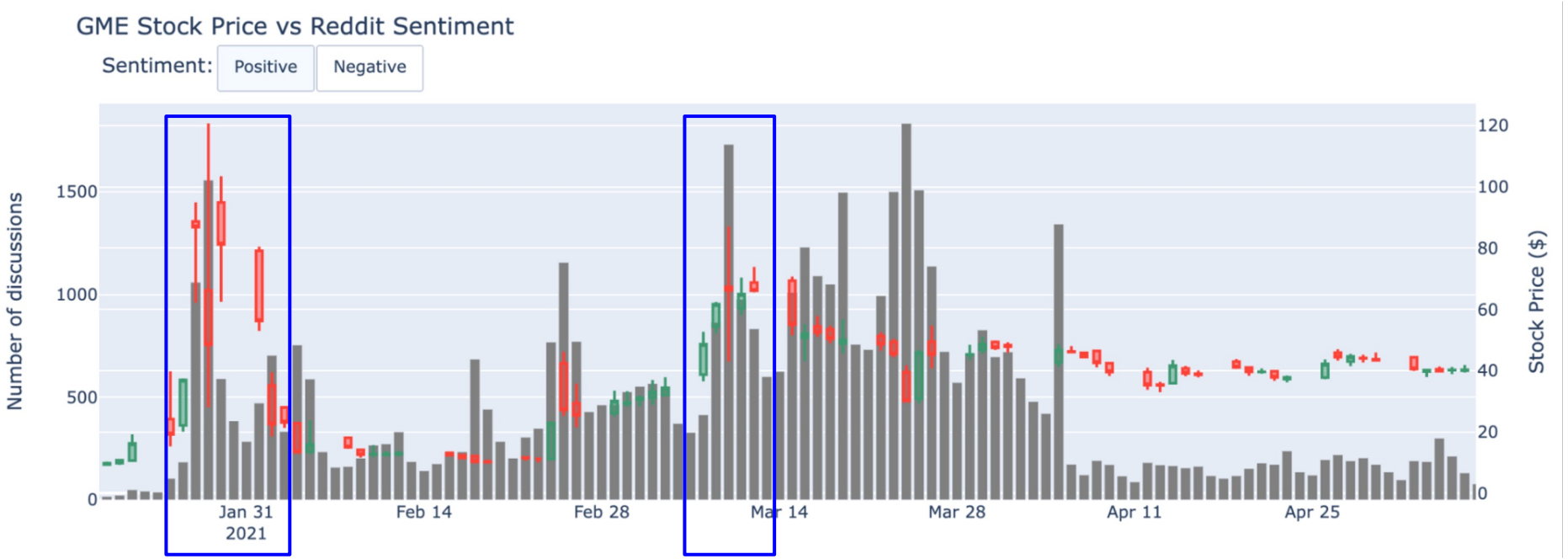


Fig: GME Stock Price vs. Reddit Sentiment

GEORGETOWN UNIVERSITY

Co-movement with Stock Price

Fig: GME Stock Price vs. Reddit Sentiment

GEORGETOWN UNIVERSITY

Correlation Analysis

- Direct correlation

	Open	High	Low	Close	Return
neg_count	0.190421	0.295703	0.034448	0.135550	-0.088591
pos_count	0.208103	0.294619	0.077112	0.167200	-0.049275
%neg	-0.072633	-0.063947	-0.111795	-0.103508	-0.081021
%pos	0.132717	0.102020	0.183161	0.154084	0.012134

- Correlation with smoothed price (3-day average)

	Open_avg	High_avg	Low_avg	Close_avg	Return_avg
neg_count	0.164997	0.243624	0.035612	0.128597	-0.003135
pos_count	0.187143	0.258480	0.072598	0.157566	0.013169
%neg	-0.064149	-0.058372	-0.099738	-0.091140	-0.101963
%pos	0.143393	0.117465	0.186976	0.160782	-0.001114

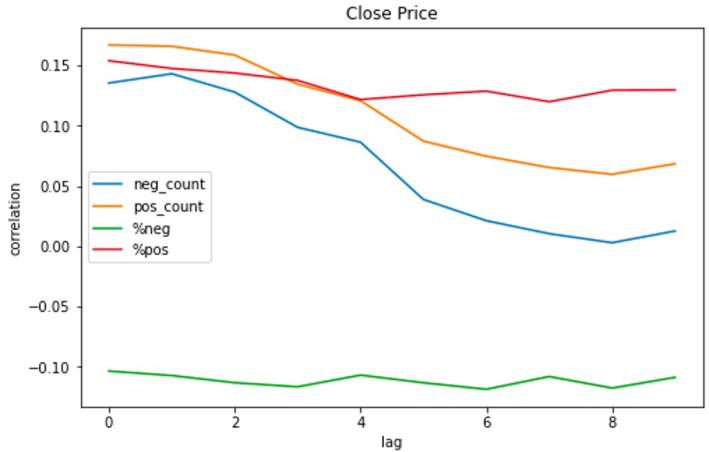
GEORGETOWN UNIVERSITY

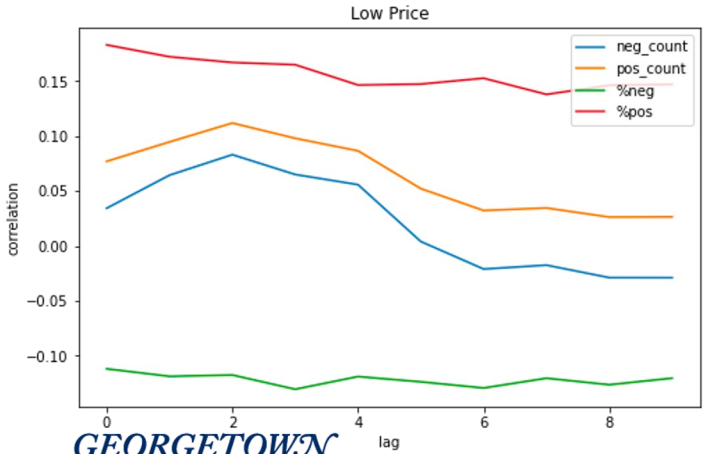
Delayed Time Effect - Time Lagged Cross Correlation

- Lag stock price by 3 days

	Open	High	Low	Close	Return
neg_count	0.112454	0.183123	0.002548	0.069588	-0.071774
pos_count	0.143459	0.206967	0.044253	0.110450	-0.038988
%neg	-0.053496	-0.039359	-0.094069	-0.077686	-0.067056
%pos	0.166088	0.137713	0.203534	0.178016	0.005419

- Time Lagged Cross Correlation





GEORGETOWN UNIVERSITY

Stock Price Movement Prediction

- Sentiment: positive counts
- 6 Technical indicators:
  - MOM (Momentum)
  - RSI (Relative Strength Index),
  - NATR (Normalized Average True Range)
  - CCI (Commodity Channel Index)
  - OBV (On Balance Volume)
  - ADX (Average Directional Movement Index)
- Model Architecture: LSTM
  - Data are transformed into blocks with a time step of 14 days
  - 20 epochs
- Performance
  - Best Accuracy:
    - Train: 69.83%
    - Test: 55.74%

Layer (type)	Output Shape	Param #
lstm_30 (LSTM)	(None, 14, 8)	512
dropout_30 (Dropout)	(None, 14, 8)	0
batch_normalization_30 (BatchNormalization)	(None, 14, 8)	32
lstm_31 (LSTM)	(None, 14, 16)	1600
dropout_31 (Dropout)	(None, 14, 16)	0
batch_normalization_31 (BatchNormalization)	(None, 14, 16)	64
lstm_32 (LSTM)	(None, 8)	800
dropout_32 (Dropout)	(None, 8)	0
batch_normalization_32 (BatchNormalization)	(None, 8)	32
dense_10 (Dense)	(None, 1)	9

GEORGETOWN UNIVERSITY

Conclusion

- Effect of fine tuning
  - Models with unfreezed parameters and fine tuned in the training process yield significant increases in accuracy.
- Effect of pre-training on Sentiment140
  - Pre-tuning on the Sentiment140 Dataset remains beneficial.
  - Especially effective when the training data size is small.
- Predictive power for stock price
  - Limited availability of data
  - Slightly better than random guess
- Possible further study
  - A decaying learning rate for lower layers, which intuitively would contain more general information.

Code links

Please follow click on Github to view the code

Thanks for watching!

Any questions?

GEORGETOWN UNIVERSITY