ผลงานวิจัยดีเด่นด้านตลาดทุน

Stock Market Prediction Using Deep Learning Based Model with Textual Representation and Technical Indicators

โดย คุณธนวัฒน์ ชิวหวรรณ
อาจารย์ที่ปรึกษา: ผศ. ดร. พีรพล เวทีกูล
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

วันที่ 14 กันยายน 2563

Outline

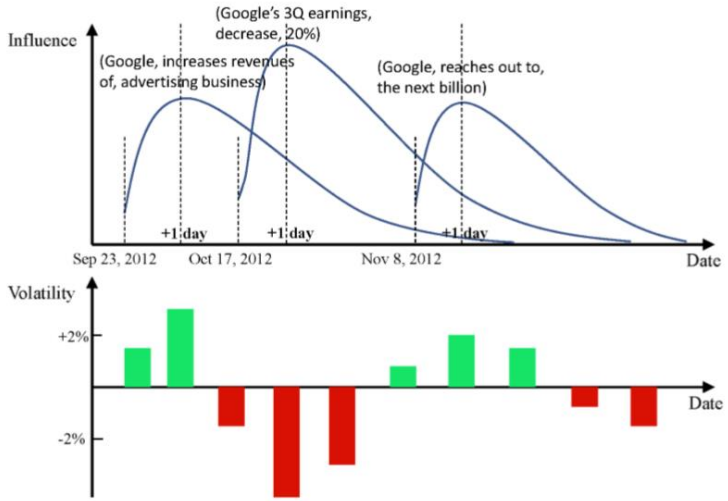
- **Introduction**
- **Research problem and objective**
- **Methodology**
  - Data
  - Textual representation
  - Deep learning model
- **Experimental setup**
- **Experimental result**
- **Discussion**
- **Model explainability**
- **Conclusion**

Introduction

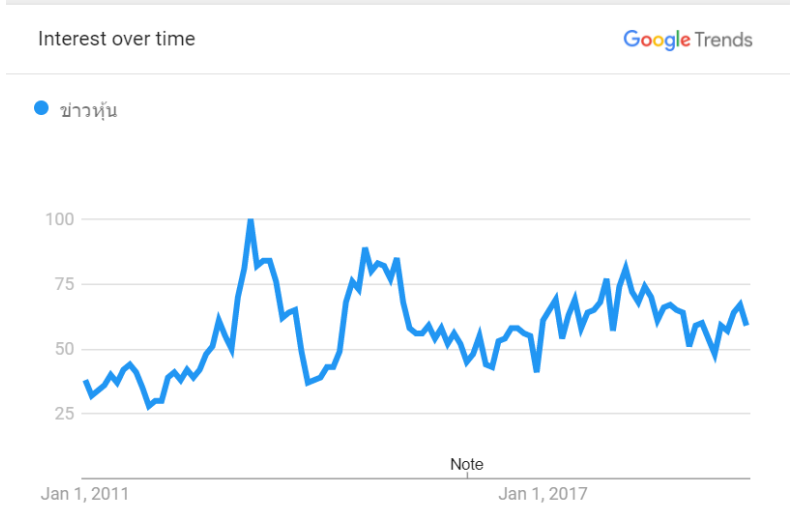
How can we utilize both numeric and textual data for market prediction?

- **Digital Era Challenges: Amount of Data, Variety of Data**
- **How to leverage both textual and numerical information into model?**
- **Deep learning technique: NLP, Event embedding, Attention, BERT**
- **Thai market challenges**

SET index 2011-2019



News headline effect, Ding et al 2015

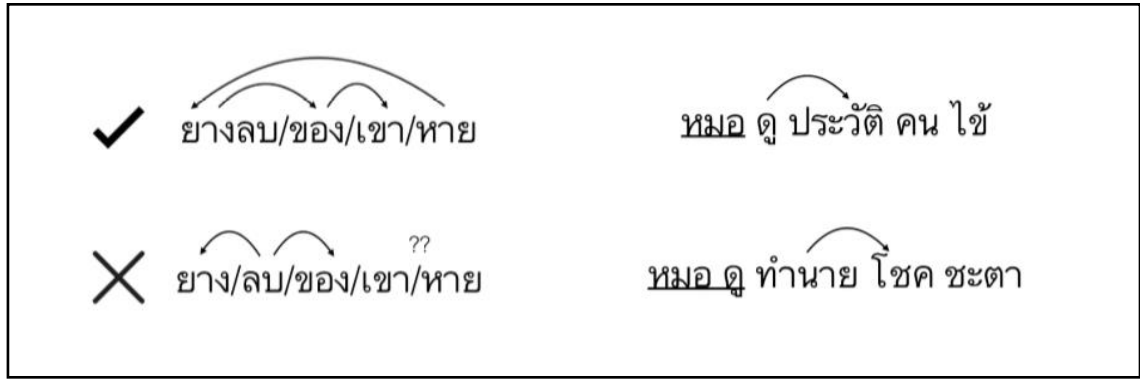


Google trends search term "ข่าวหุ้น" 2011-2019

Research Problem

Purpose of our study

- **Research Objectives:**
  - Introduce deep learning as a tool analyze both textual and numerical data to set market index
  - Optimize and research on deep learning predictive models for the Thailand market
  - Comparative analysis of deep learning performance on textual and numeric data
- **Model Objective: Predict next day SET index returns**
- **Deep learning Objective:**
  - Deep learning architecture experiments
  - Tackle both textual and numerical features input
  - Explore textual representation approach
  - Model explainablity

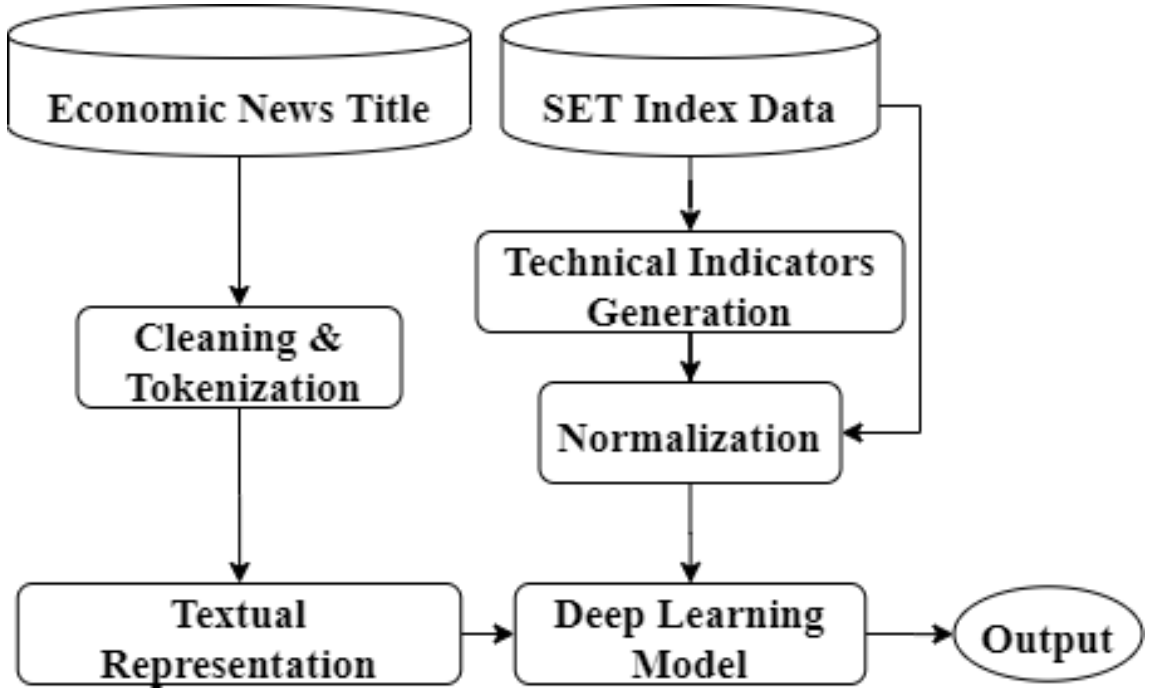


Tokenization and word ambiguities challenges in Thai

An integrated model for word segmentation, part-of-speech tagging, and transition-based dependency parsing, Kwankajornkiet C. 2016

Methodology
Proposed frameworks

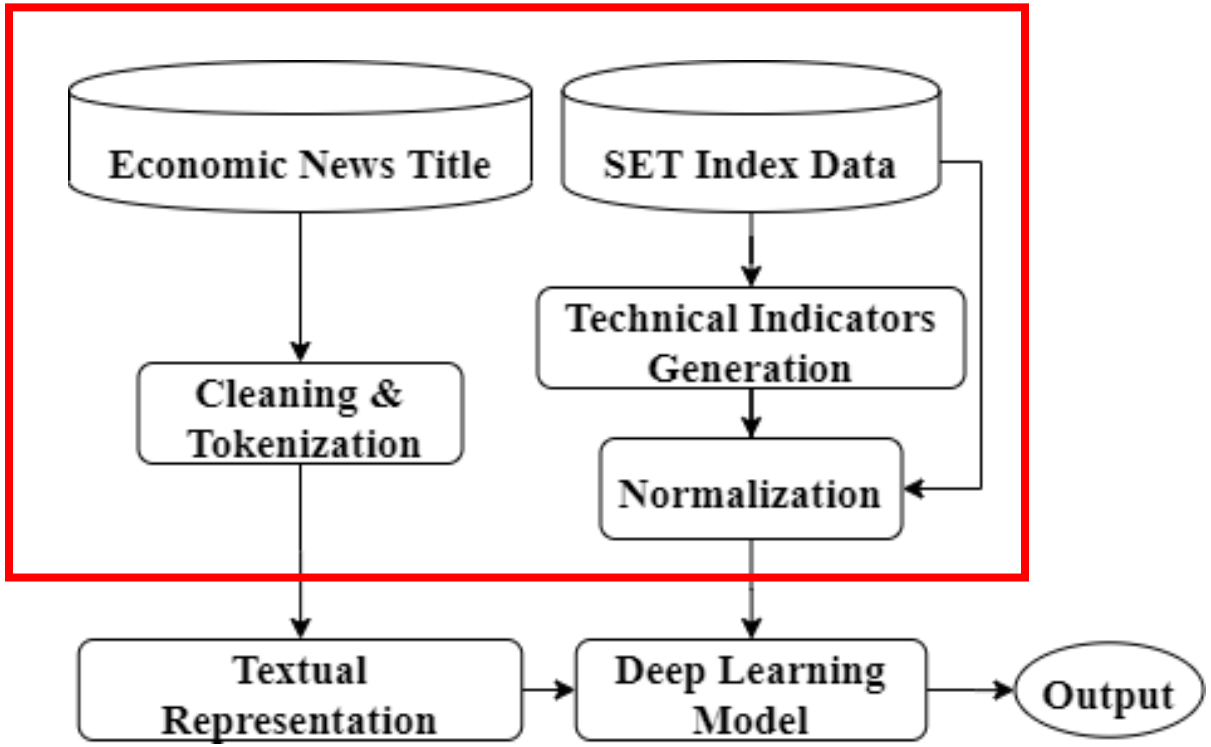
1. **News title data (Textual data)**
   - Cleaning & Tokenization
   - Textual Representation
2. **Set index data (Numerical data)**
   - Technical indicator generation
   - Normalization
3. **Deep learning model**
   - Output trend tomorrow returns



Methodology

Proposed frameworks

1. **News title data (Textual data)**
   - Cleaning & Tokenization
   - Textual Representation
2. **Set index data (Numerical data)**
   - Technical indicator generation
   - Normalization
3. **Deep learning model**
   - Output trend tomorrow returns



Data

Numerical & textual input features

- **SET market index 2008 to 2019 ~ 2,931 trading days**
- **Thai economic news headlines from various online source ~ 885K news**
- **73 numeric timeseries (Price info +Technical indicators)**

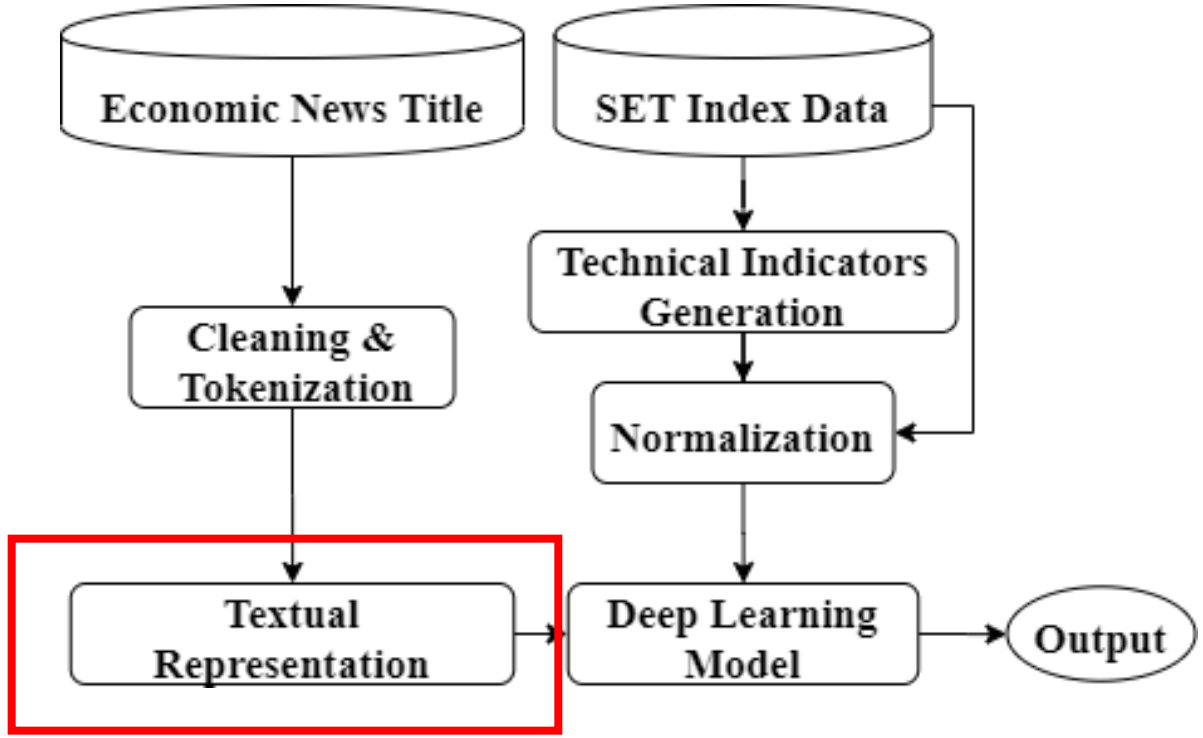
	Data period	Jan-2008 to Dec-2017	Jan-2009 to Dec-2018	Jan-2010 to Dec-2019
News headline records	Training	600,220	631,296	604,764
	Validating	75,425	73,923	72,686
	Testing	73,923	72,686	62,756
Trading days	Training	1,952	1,949	1,950
	Validating	244	244	247
	Testing	244	247	244

Dataset	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1	Training								Validate	Testing		
2		Training								Validate	Testing	
3			Training								Validate	Testing

news_date	header
2009-04-16	คอลัมน์ชีพจรโลกธุรกิจ:ศุลกากรเปิดบริการ16-17เม.ย.
2013-10-23	คอลัมน์จับประเด็น:ขยายเวลาลดภาษีสรรพสามิตดีเซล...
2012-11-14	คอลัมน์ธุรกิจโลก:รมว.ยุ่นยอมรับศก.หดตัว
2016-08-22	กรุงศรีปรับทัพมุ่งสู่ยุค'ดิจิทัล'
2019-01-10	หุ้นการแพทย์ดิ่งระนาว
2012-10-26	TKจิ๋วแต่แจ๋ว..ราชาสินเชื่อมอเตอร์ไซค์
2011-05-31	คลังหวั่นงบลงทุนเบิกจ่ายล่าช้า
2010-10-04	คอลัมน์Moneyweek:เงินบาททะยานทดสอบแข็งค่ารอบใหม่
2020-01-17	KTAMปันผลกองอสังหาฯมูลค่ารวมกว่า500ล้านบาท
2019-05-13	คำพิพากษาศาลฎีกาคดีผู้สมัครถือหุ้นสื่อ

Methodology
Proposed frameworks

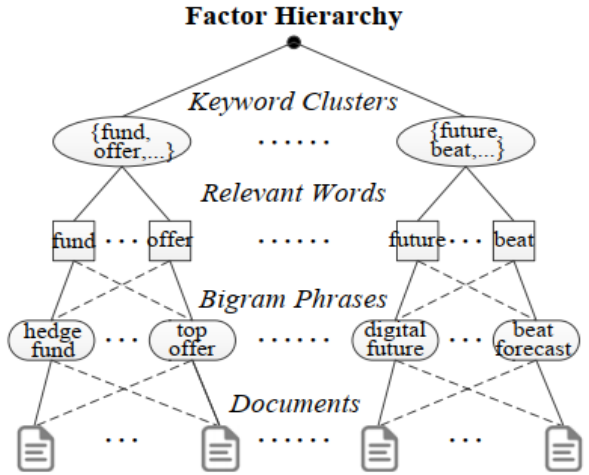
1. **News title data (Textual data)**
   - Cleaning & Tokenization
   - Textual Representation
2. **Set index data (Numerical data)**
   - Technical indicator generation
   - Normalization
3. **Deep learning model**
   - Output trend tomorrow returns



Textual Representation

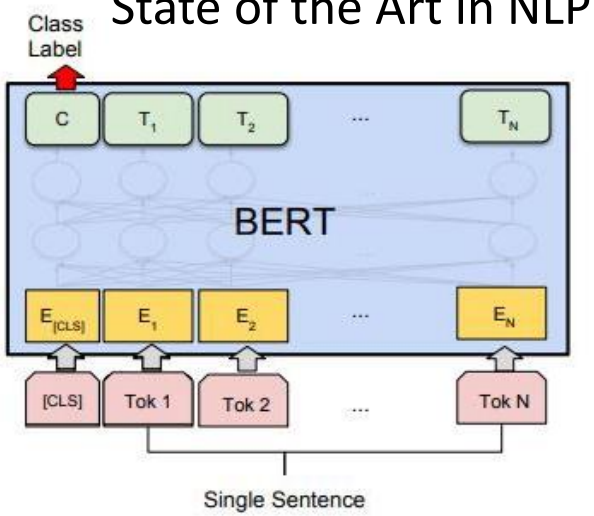
Hierarchical neural network & BERT

1. **Hierarchical Neural Network Approach**
   - Shi, L. et al, DeepClue (2018)
   - Required tokenization
   - End to end model (backpropagate + weights updated during model training)
2. **BERT Aggregated Embedding approach**
   - Devlin et al. (2018)
   - Best SOTA from Google
   - Unsupervised tokenization
   - Proposed a static embedding (BERT's weights not updated)



Shi, L., Teng, Z., Wang, L., Zhang, Y., and Binder, A.: 'DeepClue: visual interpretation of text-based deep stock prediction'

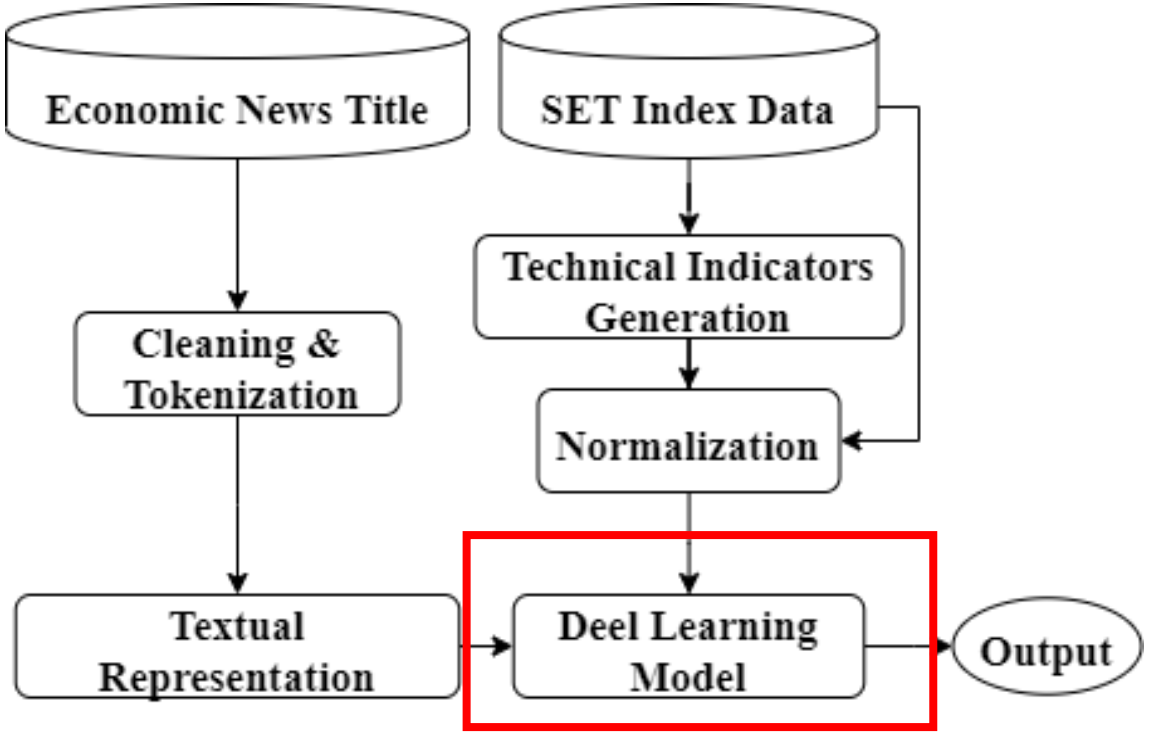
State of the Art in NLP



Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: 'Bert: Pre-training of deep bidirectional transformers for language understanding'

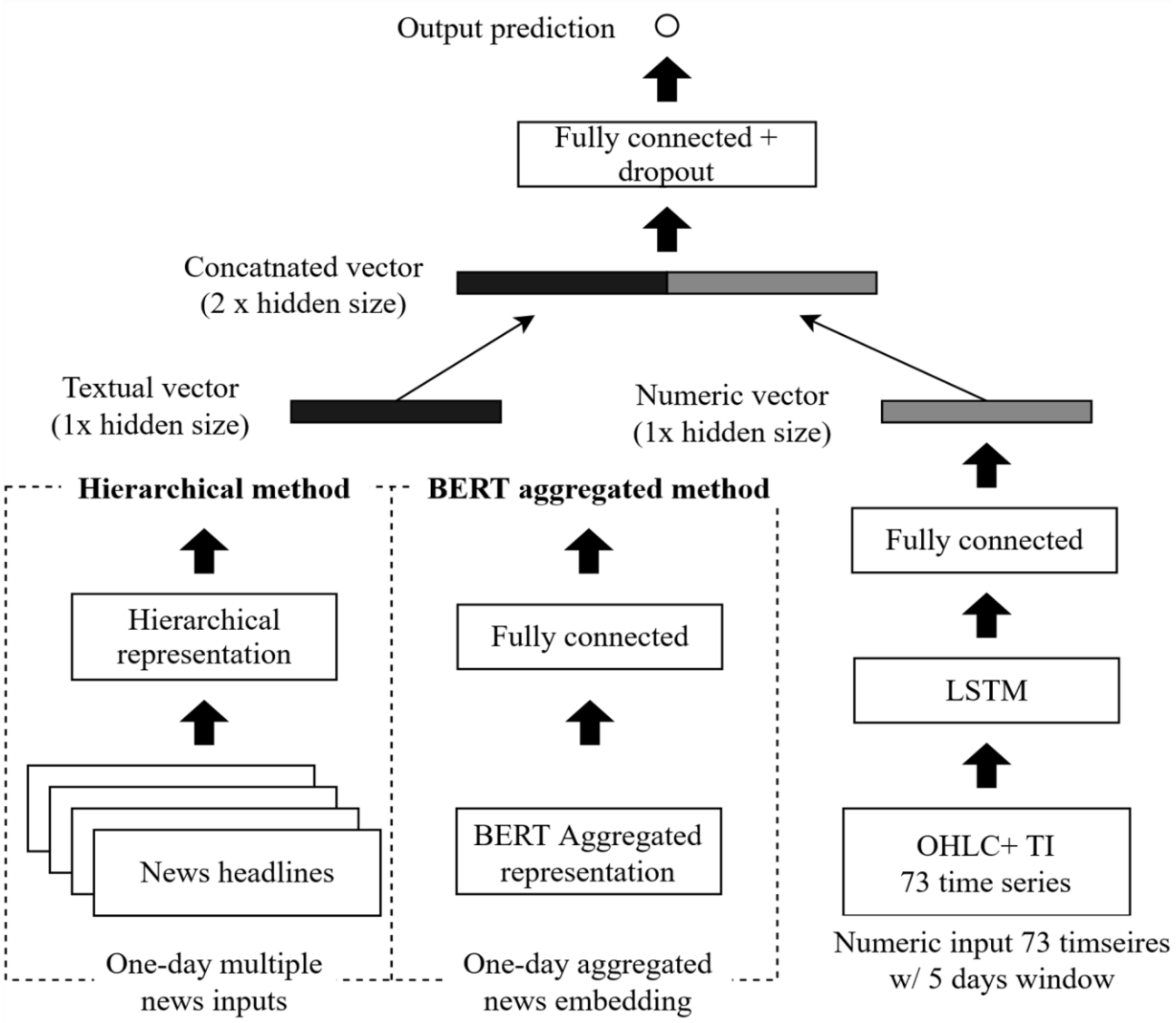
Methodology
Proposed frameworks

1. **News title data (Textual data)**
   - Cleaning & Tokenization
   - Textual Representation
2. **Set index data (Numerical data)**
   - Technical indicator generation
   - Normalization
3. **Deep learning model**
   - Output trend tomorrow returns



Deep Learning Model

Integration between numeric and textual features



Full model integrated model

- **73 Numeric timeseries inputs (Price + TIs)**
- **Textual Representation vectors**
  - Hierarchical

    or
  - BERT
- **Concatenated vector**
- **Fully connected prediction layer**

Experimental Setup

Evaluation metrics and baselines

Metrics:

1. RMSE (root mean square error)
2. Market prediction accuracy
3. Hit profit (proposed metric, count profit when predict correct direction)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{t=n}(\hat{y}_t - y_t)^2}{n}} \qquad Accuracy = \frac{\sum_{t=1}^{t=n}(h_t)}{n} \qquad Hit\ profit = \sum_{t=1}^{t=n}(2h_t - 1)\, y_t$$

$$h_t = \begin{cases} 1 & \text{if } \hat{y}_t \text{ and } y_t \text{ are both positive or negative,} \\ 0 & \text{other wise,} \end{cases}$$

Baselines:

1. SET INDEX (buy/hold)
2. RANDOM (1000 simulations)
3. LSTM (OHLC)
4. LSTM (OHLC +TI)
5. BERT NEWS
6. BERT NEWS +LSTM TI
7. HIERARCHY NEWS
8. HIERARCHY NEWS +LSTM TI

Experimental Results

Three-year average performances (model on best validate loss)

Model	RMSE	Accuracy	Hit profit
SET INDEX	-	-	-
RANDOM	1.536%	50.1%	0.0%
LSTM(OHLC)	0.587%	49.2%	-5.8%
LSTM(OHLC+TI)	**0.586%**	**53.4%**	4.6%
BERT NEWS	0.589%	52.5%	0.3%
BERT NEWS + LSTM TI	0.589%	49.7%	5.8%
HIERARCHY NEWS	0.587%	51.2%	3.2%
HIERARCHY NEWS + LSTM TI	0.590%	51.3%	**10.1%**

RMSE: indifferent results 0.58-0.59%.

Hit Profit: HIERARCHY NEWS + LSTM TI at 10.1%

Accuracy: LSTM(OHLC+TI) at 53.4%.

HIERARCHY NEWS + LSTM TI ranks the third at 51.3%.

Discussion

Effect of features types (numeric or textual)

Input type	Model	RMSE	Accuracy	Hit profit	Hit profit diff LSTM TI	Hit profit diff Textual model
Numeric only	LSTM(OHLC)	0.59%	49.20%	-5.80%	-10.30%	-
	LSTM(OHLC+TI)	0.59%	53.40%	4.60%	-	-
Textual only	BERT NEWS	0.59%	52.50%	0.30%	-4.30%	-
	HIERARCHY NEWS	0.59%	51.20%	3.20%	-1.40%	-
Numeric + Textual	BERT NEWS + LSTM TI	0.59%	49.70%	5.80%	1.20%	5.50%
	HIERARCHY NEWS + LSTM TI	0.59%	51.30%	**10.10%**	5.60%	6.90%

Combination of textual features and 73 timeseries show best performance

Discussion

Textual representation approach

Textual Representation	Model	RMSE	Accuracy	Hit profit	Hit profit diff BERT
BERT	BERT NEWS	0.589%	52.5%	0.3%	-
	BERT NEWS + LSTM TI	0.589%	49.7%	5.8%	-
HIERARCHY	HIERARCHY NEWS	0.587%	51.2%	3.2%	+3.0%
	HIERARCHY NEWS + LSTM TI	0.590%	51.3%	**10.1%**	+4.4%

Hierarchical outperforms BERT approach

Explainability
Integrated gradient

Integrated gradient, Attribution method

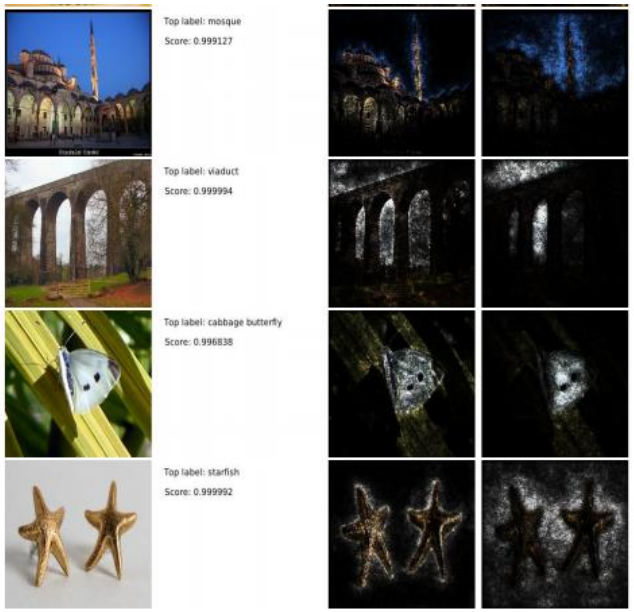


Figure 2. **Comparing integrated gradients with gradients at the image.** Left-to-right: original input image, label and softmax score for the highest scoring class, visualization of integrated gradients, visualization of gradients*image. Notice that the visualizations obtained from integrated gradients are better at reflecting distinctive features of the image.

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Figure 4. **Attributions from question classification model.** Term color indicates attribution strength—Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.

Sundararajan, M., Taly, A., and Yan, Q., 2017. Axiomatic attribution for deep networks.

Explainability

Importance score at word/sentence level

Top 20 Positive	Importance Score	Top 20 negative	Importance Score
%	4,153	-	(7,620)
เพิ่ม	3,285	ไม่	(2,708)
หุ้น	2,698	ปี	(1,960)
ซื้อ	2,627	นี้	(1,392)
ลด	1,972	รับ	(1,360)
ไทย	1,531	ขึ้น	(1,254)
ขาย	1,452	โต	(1,227)
ราคา	1,393	รัฐ	(808)
ส่งออก	1,265	อี	(680)
!	1,217	มี	(605)
ใหม่	1,170	ยัง	(585)
กำไร	1,104	คน	(479)
ปรับ	1,076	3	(470)
เปิด	1,041	แรก	(438)
ลงทุน	1,014	สอบ	(437)
ธุรกิจ	949	รายได้	(423)
พุ่ง	937	มา	(423)
ตลาด	933	หมื่น	(415)
แบงก์	920	2	(411)
เร่ง	871	เข้า	(387)

Sentences Score	News Date	Word Importance
-0.33	2016-12-30 642916	ท่องเที่ยว โต เกิน เป้า สร้าง รายได้ ทะลุ 2.51 ล้าน ล้าน บาท ต่างชาติ
-0.28	2016-12-30 642920	กลยุทธ์ การลงทุน ใน ปี 2017
-0.14	2016-12-30 642883	ตลาดหุ้น คึกคัก ปิด บวก 13 จุด
+0.15	2019-01-12 791364	สรรพากร ริด ภาษี ทะลุเป้า 7 % ดึง เทคโนโลยี เพิ่มประสิทธิภาพ
+0.30	2019-01-12 791407	แจก โบนัส ทำ เชื่อมั่น ครัวเรือน ขยับ
+0.18	2019-01-12 791418	เพิ่ม ลงทุน ทองคำ 20 %
-0.18	2017-04-06 662549	ผวา ทรัมป์ ขึ้น ภาษี 15 %
+0.23	2017-04-06 662639	บุริรัมย์ เพิ่ม มูลค่า สินค้า เกษตร จับมือ จ. ระนอง เชื่อมโยง
+0.26	2018-05-21 744116	ปั้น ร้าน ธง ฟ้า เป็น ร้าน ขายของ ถูก จีบ ยู นิ ลี เวอร์ - สห พัฒน์ ผลิต
+0.48	2017-09-16 695459	พา ที ไข ก๊อก ซีอีโอ นก แอร์ หุ้น พุ่ง 8 %
-0.63	2017-09-16 695510	ปตท. ขึ้น NGV รอบ แรก 16 ก.ย. นี้ 27 สตางค์
-0.24	2017-09-16 695443	จับคู่ ค้า อาเซียน โชว์ สินค้า หนาน หนิง
-0.19	2017-08-08 687644	เดียว ทุบ เดียว ดัน
+0.51	2017-08-08 687653	TISCO หุ้น เด่น สุด ของ กลุ่ม ธนาคาร กำไร เพิ่ม พัน ล้าน
+0.45	2017-08-08 687552	โบรก ชี้ หุ้น ไทย ไป ต่อ ตัวเลข เศรษฐกิจ ฟื้น แนะ ซื้อ TTA เป้า 13 บ.
+0.70	2018-03-10 730498	สหรัฐ เซ็น เก็บภาษี เหล็ก 11 ชาติ ลงนาม ซี พี ที พี พี

Conclusion

Conclusion:

- Improve the performance of the SET index prediction using textual representation and numerical time-series as inputs to the deep learning model
- Hierarchical neural network structure represent textual information better than BERT embedding method
- Interpretability framework could be added to visualize and debug model prediction

Room for improvement

- Extend deep learning to intra-day data (Implementation & resource challenges)
- BERT embedding without aggregation but train model directly (resource challenges)
- Explore deep learning with trading strategy-oriented objectives
- Quantitative measures for model interpretability

Acknowledgement

Thank you

We would like to express our appreciation to all the parties who support us during this research. Big thank you to all involved

The Stock Exchange of Thailand (SET)

Capital Market Research Institute (CMRI), all officers, and committees.

The Datamind Laboratory, Department of Computer Engineering, Chula

- Asst. Prof. Peerapon Vateekul, Ph.D.
- Tanawat Chiewhawan

Website development

- Kunat Pipatanakul

Web

SET index prediction model

URL https://deep-insight-trader.web.app/

Web

News headline visualization

Thank you for your attention

Q&A

Reference and resource

Published paper:

https://dl.acm.org/doi/proceedings/10.1145/3411174

Research grants:

https://www.set.or.th/th/setresearch/grant/research_grant.html

ผลงานวิจัยดีเด่นด้านตลาดทุน

Stock Market Prediction Using Deep Learning Based Model with Textual Representation and Technical Indicators

โดย คุณธนวัฒน์ ชิวหวรรณ
อาจารย์ที่ปรึกษา: ผศ. ดร. พีรพล เวทีกูล
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

วันที่ 14 กันยายน 2563


CMRI
Capital Market Research Institute