

Research Article

การพยากรณ์ทิศทางของราคาหุ้นรายวันจากข้อความข่าวภาษาไทย โดยใช้วิธีการประมวลผลภาษาธรรมชาติ

The predictions of a daily stock price direction from the Thai news content by using natural language processing

วิกานดา ผาพันธุ์^{1*} และ อัญชนา พิมพ์สาล¹

Wikanda Phaphan^{1*}, and Aunchana Pimpisal¹

¹ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

¹Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok

*E-mail: wikanda.p@sci.kmutnb.ac.th

Received: 16/04/2020; Revised: 14/05/2020; Accepted: 15/06/2020

บทคัดย่อ

ปัจจัยที่กระทบต่อราคาของหุ้นในตลาดหลักทรัพย์แห่งประเทศไทยนั้นมีอยู่หลายปัจจัย ข่าวสารต่าง ๆ ก็เป็นปัจจัยหนึ่งที่มีผลกระทบต่อราคาของหุ้น ผู้วิจัยจึงเกิดแนวคิดที่จะพยากรณ์ทิศทางของราคาหุ้นรายวันจากข้อความข่าวโดยใช้วิธีการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เพื่อให้นักลงทุนสามารถคาดคะเนทิศทางของราคาหุ้นก่อนที่จะตลาดหลักทรัพย์แห่งประเทศไทยเปิดโดยศึกษาข้อความข่าวจากแหล่งข่าวต่างๆ และใช้การตัดคำ (Tokenizer) จาก Library pythainlp ในโปรแกรมภาษาไพธอน ver.3.7.1 จากนั้นสร้างแบบจำลองโดยใช้ตัวแบบการจำแนก (classification model) เพื่อหาแบบจำลอง (model) และวิธีการตัดคำ (Tokenizer) ที่มีค่าความถูกต้องแม่นยำ (accuracy) สูงสุดเพื่อใช้พยากรณ์ทิศทางของราคาหุ้นรายวัน ซึ่งในงานวิจัยนี้ได้พยากรณ์ทิศทางของราคาหุ้นทั้งหมด 3 วัน คือวันที่ 5, 6 และ 7 กุมภาพันธ์ 2563 โดยสุ่มหุ้นอย่างละ 1 ตัว ด้วยการสุ่มตัวอย่างอย่างง่าย (SRS) จากหุ้น 5 กลุ่ม คือ กลุ่ม ICT กลุ่ม ENER กลุ่ม HEALTH กลุ่ม COMM และ กลุ่ม BANK ผลวิจัยพบว่า กลุ่ม ICT สุ่มได้หุ้นของบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) ตัวแบบ Gradient Boosting Classifier เป็นตัวแบบที่มีความเหมาะสมมากที่สุดและเปรียบเทียบสถานะค่าพยากรณ์กับค่าจริงได้ความถูกต้องร้อยละ 100 ในส่วนของกลุ่ม ENER สุ่มได้หุ้นของบริษัท ไทยออยล์ จำกัด (TOP) และกลุ่ม HEALTH สุ่มได้หุ้นของบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (BH) นั้น ไม่สามารถสรุปตัวแบบที่มีความเหมาะสม

มากที่สุดได้และเปรียบเทียบสถานะค่าพยากรณ์กับค่าจริงได้ความถูกต้องร้อยละ 66.67 ในกลุ่ม COMM สุ่มได้หุ้นของบริษัท ซีพี ออลล์ จำกัด (CPALL) และกลุ่ม BANK คือธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) ตัวแบบ K-Neighbors Classifier เป็นตัวแบบที่ความเหมาะสมมากที่สุดและเปรียบเทียบสถานะค่าพยากรณ์กับค่าจริงได้ความถูกต้องร้อยละ 66.67

คำสำคัญ: วิธีการประมวลผลภาษาธรรมชาติ, ตัวแบบการจำแนก, หุ่น, Pythainlp, การวิเคราะห์ข้อความข่าว

Abstract

Factors affecting a stock price in the Stock Exchange of Thailand have several factors, including the various news. Hence, the concept of the daily stock price direction prediction from the Thai news content using the natural language processing is studied so investors are able to forecast the stock price direction before the Stock Exchange of Thailand operates. We made a study of the Thai news content with the tokenizer in Python version 3.7.1 from library pythainlp and then classification model was used for finding the most accurate values of the model and the tokenizer. This study was carried out the forecast of stock price direction in three days: 5th, 6th, and 7th February, 2020. One stock randomly chosen used simple random sampling from the following five stock groups: the ICT group, the ENER group, the HELTH group, the COMM group, and the BANK group. The results revealed that the stock of Intouch Holdings Company (INTUCH) randomly chosen by the ICT group is an efficient Gradient Boosting Classifier model when it is compared with forecasting and actual values of 100 %, the stocks of Thai Oil Public Company Limited (TOP) and Bumrungrad Hospital (BH), randomly chosen by the ENER group and the HELTH group respectively, are not able to give us efficient models when they are compared with forecasting and actual values of 66.67 %. In addition, the stocks of CP ALL public company limited of COMM group and the stock of Kasikornbank Public Company Limited of the BANK group are efficient KNeighbors Classifier models when they are compared with forecasting and actual values of 66.67 %.

Keywords: natural language processing, classification model, stock, Pythainlp, news analysis

บทนำ

ตลาดหลักทรัพย์เป็นแหล่งซื้อขายแลกเปลี่ยนหลักทรัพย์ระยะยาว ทำหน้าที่เป็นตลาดทุน เพื่อให้บริษัทมหาชน จำกัด ซึ่งถือว่าเป็นตลาดรอง (Secondary Market) สามารถระดมเงินทุนเพิ่มเติมจากสาธารณะได้ โดยหลักทรัพย์ระยะยาวจะประกอบไปด้วยตราสารหนี้ และตราสารทุนซึ่งประกอบไปด้วย หุ้นสามัญ หุ้นบุริมสิทธิ ใบสำคัญแสดงสิทธิแบบต่าง ๆ ใบสำคัญแสดงสิทธิอนุพันธ์ หุ้นกู้ และ หน่วยลงทุน เป็นต้น

ในปัจจุบันการซื้อขายหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทยเป็นที่นิยมและรู้จักกันอย่างกว้างขวาง เนื่องจากเป็นแหล่งระดมเงินทุนที่มีผลตอบแทนสูงกว่าการลงทุนประเภทอื่น เช่น การซื้อที่ดิน การซื้อทองคำแท่งหรือ การฝากเงินกับธนาคารพาณิชย์ เป็นต้น จึงทำให้มีผู้สนใจเข้ามาลงทุนเพื่อสร้างความมั่งคั่งให้กับตนเองเป็นจำนวนมากแต่การลงทุนในตลาดหลักทรัพย์แห่งประเทศไทยมีอ่อนไหวสูง ไม่ว่าจะเป็นเหตุการณ์ใดๆ เกิดขึ้นก็จะส่งผลกระทบต่อสถานะตลาดได้อย่างรวดเร็ว เช่น ปัญหาเศรษฐกิจ ปัญหาการแพร่ระบาดของโรค ปัญหาการเมือง ความมั่นคง และการทหาร เป็นต้น ซึ่งปัญหาต่างๆ เหล่านี้จะส่งผลให้ราคาหลักทรัพย์มีความผันผวนตลอดเวลา ทำให้ผลตอบแทนไม่เป็นไปตามที่นักลงทุนคาดหวัง ดังนั้นการลงทุนในหลักทรัพย์ควรมีหลักเกณฑ์ในการพิจารณาเพื่อเป็นแนวทางในการตัดสินใจของนักลงทุน

ปัจจัยที่กระทบต่อราคาของหุ้นในตลาดหลักทรัพย์แห่งประเทศไทยที่นักลงทุนนิยมพิจารณานั้นมีอยู่หลายปัจจัย เช่น ผลการดำเนินงานของบริษัท จำนวนการซื้อขายหุ้นในแต่ละวัน รวมทั้งข่าวสารต่างๆ ที่มีผลกระทบต่อจิตวิทยาของตลาดหุ้นและราคาของหุ้น อาทิ ข่าวการลงทุนของบริษัท ข่าวการระบาดของโรค ข่าวการชุมนุมทางการเมือง เป็นต้น ด้วยความสำคัญของข่าวสารต่างๆ นี้ ผู้วิจัยจึงเกิดแนวคิดที่จะพยากรณ์ทิศทางของราคาหุ้นรายวันจากข้อความข่าวภาษาไทย โดยใช้วิธีการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เพื่อให้ นักลงทุนสามารถคาดคะเนทิศทางของราคาหุ้นก่อนที่ตลาดหลักทรัพย์แห่งประเทศไทยจะเปิดได้ โดยจะทำการศึกษาข้อความข่าวของหุ้นจากตลาดหลักทรัพย์แห่งประเทศไทย (SET) จำนวน 5 ตัว จากหุ้น 5 กลุ่ม ด้วยโปรแกรมภาษาไพธอน ver.3.7.1 บนโปรแกรม Visual Studio Code และใช้โปรแกรม Power BI ช่วยในการแสดงผล

งานวิจัยที่เกี่ยวข้อง

Li et al. (2014) ศึกษาการเพิ่มประสิทธิภาพการพยากรณ์ราคาหุ้นระหว่างวันของ Hong Kong Stock Exchange (HKEx) โดยรวมข้อมูลจาก 2 แหล่งคือ 1. ข้อความข่าว และ 2. ราคาหุ้นในอดีต ซึ่งข้อมูลทั้งสองแหล่งนี้ มีความอิสระต่อกัน มีผลกระทบต่อราคาหุ้นระหว่างวันมาก และมีคุณสมบัติดังนี้ 1. เวลา กล่าวคือข้อมูลข้อความข่าวและราคาหุ้นในอดีตส่วนมีเรื่องของเวลาเข้ามาเกี่ยวข้อง 2. ความถี่ ในการพยากรณ์ราคาหุ้นระหว่างวันเป็นการพยากรณ์ระยะสั้น หากต้องการความถี่ในการพยากรณ์สูง จะต้องควบคุมข้อมูลให้มีลักษณะดังนี้ 1. มีปริมาณซื้อขายมาก และ 2. มีความแตกต่างระหว่างช่วงเวลา โดยการเตรียมข้อความข่าวที่เป็นภาษาจีนนั้น จะแบ่งกลุ่มโดยใช้ซอฟต์แวร์การแบ่งข้อมูลภาษาจีนซึ่งมีคลังคำศัพท์ทางการเงิน จากนั้นทำการลบคำที่ไม่สำคัญออกไป และให้ค่าน้ำหนักด้วย TFIDF (term frequency - inverse document frequency) แล้วเรียงลำดับเวลาของราคาหุ้น รวมทั้งจัดการข้อความความข่าวและราคาหุ้นตอนตลาด HKEx เปิด คือเวลา 10.00-12.30 น. และ 14.30-16.00น. นอกจากนี้คณะผู้วิจัยได้ทำการระบุหัวข้อข่าวทุกๆ 5, 10, 15, 20, 25 และ 30 นาที รวมทั้งได้เพิ่มตัวชี้วัดการวิเคราะห์หุ้นทางเทคนิคคือ Relative Strength Index (RSI), Raw Stochastic Value (RSV), Williams Index, Bias และ Psychological Line (PYS) จากนั้น ทำการลดความซ้ำซ้อนของข้อมูล (normalization) ก่อนสร้างตัวแบบ ส่วนตัวแบบ (model) ที่ใช้ แบ่งเป็น 3 กลุ่มคือ 1. ตัวแบบที่ฝึกสอนด้วยข้อมูลเพียง

แหล่งเดียว (ข้อความข่าวหรือราคาหุ้นในอดีต) 2. ตัวแบบที่ฝึกสอนด้วยข้อมูลจาก 2 แหล่ง โดยใช้การรวมตัวแปร (feature) อย่างง่าย 3. ตัวแบบที่ฝึกสอนด้วยข้อมูลจาก 2 แหล่งโดยใช้วิธีการรวมกันแบบ Multi-kernel จากการวิจัยพบว่า Multi-Kernel Support Vector Regression (MKSVR) ซึ่งเป็นตัวแบบในกลุ่มที่ 3 เป็นตัวแบบที่มีค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (root mean square error) และค่าเฉลี่ยส่วนเบี่ยงเบนสัมบูรณ์ (mean absolute error) ต่ำสุด

Sadia et al. (2019) ศึกษาการพยากรณ์ราคาหุ้นโดยใช้ข้อมูลราคาหุ้นจากเว็บไซต์ Kaggle และทำการเตรียมข้อมูลโดยการตรวจสอบค่าสูญหาย แปลงข้อมูลเชิงคุณภาพ กำหนดราคาปิดของหุ้นโดยกำหนดประเภทให้ มีค่าเป็น -1 เมื่อราคาปิดของวันนั้นน้อยกว่าราคาปิดของเมื่อวาน และมีค่าเป็น 1 เมื่อราคาปิดของวันนั้นมีค่ามากกว่าราคาปิดของเมื่อวาน รวมถึงการแบ่งข้อมูลฝึกสอน (train) และข้อมูลทดสอบ (test) โดยพิจารณา 2 ตัวแบบคือ 1. Random Forest และ 2. Support Vector Machine และใช้ตาราง Confusion Matrix ในการแสดงค่าความแม่นยำ (accuracy) ผลการวิจัยพบว่าตัวแบบ Random Forest มีค่าความแม่นยำเป็นร้อยละ 80.8 และตัวแบบ Support Vector Machine มีค่าความแม่นยำร้อยละ 78.7

นอกจากนี้ยังมีงานวิจัยอื่นๆ ที่เกี่ยวข้องกับการพยากรณ์ทิศทางของราคาหุ้นโดยใช้ข้อมูลราคาหุ้นในอดีต เช่น งานวิจัยของ Huang et al. (2005) , Kim (2003) และ Tay & Cao (2001) เป็นต้น

วิธีดำเนินการวิจัย

การเก็บรวบรวมข้อมูล

งานวิจัยชิ้นนี้ผู้วิจัยใช้โปรแกรมภาษาไพธอนเป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูลข่าว และทำการสุ่มเลือกหุ้นโดยใช้การสุ่มตัวอย่างอย่างง่าย (simple random sampling : SRS) อย่างละ 1 ตัว จากหุ้น 5 กลุ่ม คือ 1. เทคโนโลยีสารสนเทศและการสื่อสาร (ICT) สุ่มได้หุ้นของบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) 2. พลังงานและสาธารณูปโภค (ENERG) สุ่มได้หุ้นของบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) 3. การแพทย์ (HEALTH) สุ่มได้หุ้นของบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) 4. พาณิชยกรรม (COMM) สุ่มได้หุ้นของบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) 5. ธนาคาร (BANK) สุ่มได้หุ้นของธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) รวมจำนวน 5 ตัว จาก SET100 ซึ่งเป็นดัชนีราคาหุ้นของตลาดหลักทรัพย์แห่งประเทศไทย (the stock exchange of Thailand) ที่ได้รับการคัดเลือก 100 อันดับแรก เก็บรวบรวมข้อมูลตั้งแต่วันที่ 18 พฤษภาคม 2562 ถึง วันที่ 7 กุมภาพันธ์ 2563 โดยรวบรวมข้อความข่าวจากแหล่งข่าวต่างๆ จำนวน 6 แหล่งข่าวคือ 1. สำนักข่าวกรุงเทพธุรกิจ 2. สำนักข่าวหุ้นอินไซด์ 3. สำนักข่าวข่าวหุ้น 4. สำนักข่าวโพสต์ทูเดย์ 5. สำนักข่าว Innnews และ 6. สำนักข่าวหุ้น สมาร์ท จากลิงค์ข้อมูลข่าวของเว็บไซต์ <https://stock.gapfocus.com/> ซึ่งสามารถค้นหาข่าวของหุ้นแต่ละตัวที่สุ่มเลือกได้ เนื่องจากเก็บรวบรวมข้อมูลเป็นการรวบรวมลิงค์ข้อความของข่าวที่มีจำนวนมากจึงใช้ Developer Tools ของ Google Chrome มาช่วยคัดแยกลิงค์ข้อความของข่าวโดยใช้ Library BeautifulSoup ในโปรแกรมภาษาไพธอนช่วยในการเก็บรวบรวม ดังรูปที่ 1 และ 2 จากนั้นทำความสะอาดข้อมูลโดยทำการลบสัญลักษณ์ต่างๆ เช่น \[.*?]\(), ลบตัวเลข และตัวอักษรภาษาอังกฤษออก รวมถึงการลบเว้นวรรคออกทั้งหมด ดังรูปที่ 3

และผู้วิจัยได้เก็บรวบรวมข้อมูลราคาหุ้นรายวันจากเว็บไซต์ <https://finance.yahoo.com/> โดยใช้ Module HistoricalPrices ใน Library yahoofinance ของโปรแกรมภาษาไพธอนช่วยในการเก็บรวบรวม เพื่อกำหนดสถานะของราคาปิดตลาดในแต่ละวันจากสมการ

$$Diff = Closeprice_i - Closeprice_{i-1} \quad (1)$$

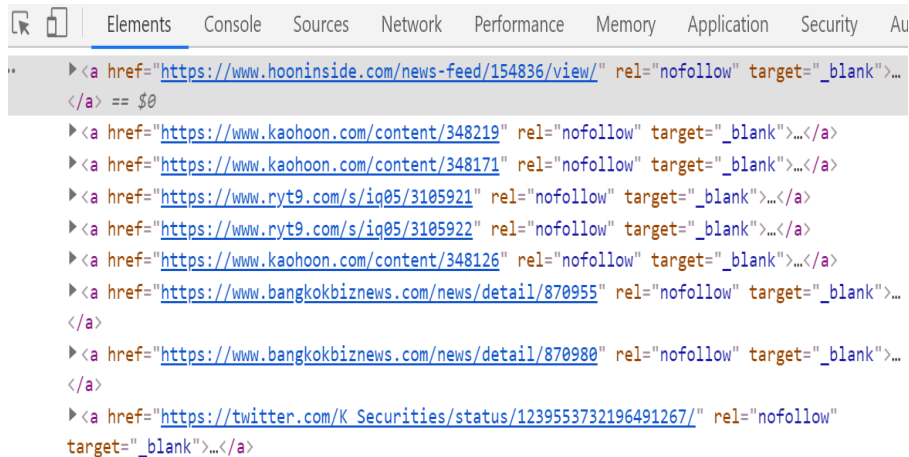
กำหนดให้ $Diff$ คือ ค่าความแตกต่างของราคาหุ้น

$Closeprice_i$ คือ ราคาปิดของหุ้นวันที่ i

$Closeprice_{i-1}$ คือ ราคาปิดของหุ้นวันที่ $i-1$

จากนั้นนำค่า $Diff$ มาหาสถานะ โดยกำหนดให้ $Diff > 0$ หมายถึง ราคาหุ้นมีแนวโน้มเป็นบวก (positive)

$Diff < 0$ หมายถึง ราคาหุ้นมีแนวโน้มเป็นลบ (negative) และ $Diff = 0$ หมายถึง ราคาหุ้นมีแนวโน้มเป็นกลาง (neutral)



```

<a href="https://www.hooninside.com/news-feed/154836/view/" rel="nofollow" target="_blank">...
</a> == $0
<a href="https://www.kaohoon.com/content/348219" rel="nofollow" target="_blank">...</a>
<a href="https://www.kaohoon.com/content/348171" rel="nofollow" target="_blank">...</a>
<a href="https://www.ryt9.com/s/ig05/3105921" rel="nofollow" target="_blank">...</a>
<a href="https://www.ryt9.com/s/ig05/3105922" rel="nofollow" target="_blank">...</a>
<a href="https://www.kaohoon.com/content/348126" rel="nofollow" target="_blank">...</a>
<a href="https://www.bangkokbiznews.com/news/detail/870955" rel="nofollow" target="_blank">...
</a>
<a href="https://www.bangkokbiznews.com/news/detail/870980" rel="nofollow" target="_blank">...
</a>
<a href="https://twitter.com/K_Securities/status/1239553732196491267/" rel="nofollow"
target="_blank">...</a>

```

รูปที่ 1 ลิขัข้อมูลความข่าวของ <https://stock.gapfocus.com> จาก Developer Tools ของ Google Chrome

stock_name	date	Title	link	source
TOP	03 Dec 07:01	หุ้นบีทีเอส-อีลิกซ์'กลดคอพิษ' สันจรจากราคาโลกขยว	https://www.bangkokbiznews.com/news/detail/856814	www.bangkokbiznews.com
TOP	03 Dec 07:01	รัฐเร่งเลขลงทุน ต่อเติม 'ความเชื่อมั่น'	https://www.bangkokbiznews.com/news/detail/856783	www.bangkokbiznews.com
TOP	03 Dec 06:31	เปิดบันทึกถึง 4 ชั่วโมง เจาะเตือนแบบ 3 สัร	https://www.bangkokbiznews.com/news/detail/856823	www.bangkokbiznews.com
TOP	03 Dec 06:01	ไทยออลส์คาดการณ์แนวโน้มสถานการณ์"ราคาน้ำมัน" 2 - 6 ธ.ค. 62 และสรุปสถานการณ์ฯ 25 - 29 พ.ย. 62	https://www.bangkokbiznews.com/news/detail/856752	www.bangkokbiznews.com
TOP	02 Dec 17:32	ราคาน้ำมันดิบเบงโกะไม่ไหวหนัก จากความไม่แน่นอนของการเจรจาการค้าสหรัฐฯ - จีน	https://www.thunhoon.com/215955/39/39/	www.thunhoon.com
TOP	02 Dec 16:02	ธนาคารพาณิชย์ "หวังสื่อคำประกันแบบบล็อกเชน" ในเครือกลุ่ม SCG	https://mgronline.com/stockmarket/detail/9620000115371	mgronline.com
TOP	02 Dec 14:32	ไทยออลส์ ราคาน้ำมันดิบเบงโกะไม่ไหวหนัก จากความไม่แน่นอนของการเจรจาการค้าสหรัฐฯ - จีน วัน/เวลา: 2 ธ.ค.	https://www.hooninside.com/news-feed/137527/view/	www.hooninside.com
TOP	02 Dec 14:31	Daily View - P.M. - บล.กสิกรไทย	https://www.kaohoon.com/content/329763	www.kaohoon.com
TOP	02 Dec 12:02	SET ดูปฏิทินถึง 20 จุด! โบทและฉายจังหวะซ้อน 6 ขึ้นเด่น-เน้นรับความเสี่ยงสูงได้	https://www.kaohoon.com/content/329699	www.kaohoon.com
TOP	02 Dec 12:01	ความคาดหวังโลกคลั่งกำลังการผลิต + PMI ขึ้นเป็นบวก หนุนบรรยากาศการค้าในช่วงสั้น	https://www.bangkokbiznews.com/news/detail/856709	www.bangkokbiznews.com
TOP	02 Dec 12:01	สหรัฐลดน้ำหนักขึ้นสูงเป็นประวัติการณ์ บริษัท ไทยออลส์ จำกัด (มหาชน) รายงานสถานการณ์น้ำมันดิบตลาดปร	https://www.ryt9.com/s/prg/3073321	www.ryt9.com
TOP	02 Dec 11:32	สหรัฐลดน้ำหนักขึ้นสูงเป็นประวัติการณ์ บริษัท ไทยออลส์ จำกัด (มหาชน) รายงานสถานการณ์น้ำมันดิบตลาดปร	https://www.innews.co.th/economy/news_545376/	www.innews.co.th
TOP	02 Dec 11:02	TOP เนย ราคาน้ำมันดิบเบงโกะไม่ไหวหนัก จากความไม่แน่นอนของการเจรจาการค้าสหรัฐฯ - จีน/เวลา: 2 ธ.ค.	https://www.hooninside.com/news-feed/137527/view/	www.hooninside.com
TOP	02 Dec 11:01	Daily Strategy for Investors on December 2, 2019	https://www.kaohoon.com/content/329685	www.kaohoon.com
TOP	02 Dec 09:32	SET ปิดดัชนีลงหวด ขวัญใจดัชนีหนักขึ้น	https://www.thunhoon.com/215873/00/49/	www.thunhoon.com
TOP	02 Dec 08:46	รายงานบทสรุปขยดเชย ประจำวันที่ 29 พฤศจิกายน 2562 2 ธ.ค. 62 07:49 น. -สำนักข่าวอินโฟเควสท์ (IQ) หลั	https://www.ryt9.com/s/ig05/3073154	www.ryt9.com
TOP	02 Dec 08:46	รายงานบทสรุปขยดเชย ประจำวันที่ 29 พ.ย. 2562 2 ธ.ค. 62 07:53 น. -สำนักข่าวอินโฟเควสท์ (IQ) หลั	https://www.ryt9.com/s/ig05/3073156	www.ryt9.com
TOP	02 Dec 08:01	SET ปิดชีพไฟแดงวันก่อน 1580-1620 จุด เป็นจุดซื้อเชิงพื้นฐาน ชู CPF, BCH, STPI หนุนเด่น	https://www.kaohoon.com/content/329624	www.kaohoon.com
TOP	29 Nov 15:31	"ไทยออลส์" คว้ารางวัลดีเด่นสาขาความเป็นเลิศด้านการจัดการทรัพยากรบุคคล จาก THAILAND CORPORATE EX	https://mgronline.com/greeninnovation/detail/9620000114617mgronline.com	mgronline.com
TOP	29 Nov 13:32	ไทยออลส์คว้ารางวัลดีเด่นสาขาความเป็นเลิศด้านการจัดการทรัพยากรบุคคล จาก THAILAND CORPORATE EX	https://www.hooninside.com/news-feed/137354/view/	www.hooninside.com
TOP	29 Nov 13:03	ไทยออลส์ คว้ารางวัลดีเด่นสาขาความเป็นเลิศด้านการจัดการทรัพยากรบุคคล	http://www.efnancethai.com/LastestNews/LatestNewsMain.a	www.efnancethai.com
TOP	29 Nov 12:32	TOP เนย ราคาน้ำมันดิบเบงโกะไม่ไหวหนัก จากความไม่แน่นอนของการเจรจาการค้าสหรัฐฯ - จีน/เวลา: 29 พ.ย. 2562 / 10:46:01	https://www.hooninside.com/news-feed/137336/view/	www.hooninside.com

รูปที่ 2 การเก็บรวบรวมลิงค์ของแหล่งข่าว

Date	Text
5/2/2020	'ภาวะตลาดหุ้นไทยเดือนมกราคมที่ผ่านมาไม่ค่อยสดใสเนื่องจากได้รับปัจจัยกดดันทั้งภาวะภัยแล้งและประเด็นการเมืองใน
4/2/2020	'วิเคราะห์สถานการณ์ราคาน้ำมันกพ', 'ราคาน้ำมันดิบปรับตัวลดลงต่อหลังความต้องการใช้น้ำมันในจีนซบเซาจากการระบาดข
3/2/2020	'หุ้นไทยภาคเข้าปีที่จุดลบจุดหรือ', 'กรุงเทพธุรกิจออนไลน์การซื้อขายหลักทรัพย์ภาคเข้าวันจันทร์ทเวลาณดัชนีปิดอยู่ที่
2/2/2020	'ปัดทบทวนแผนลงทุนปียัดเทรด'พลังงานสะอาด', 'ทิศทางการลงทุนทั่วโลกที่มุ่งสู่การผลิตและการใช้เชื้อเพลิงที่สะ
1/2/2020	'ไออาร์เร่งขยายปั๊มขายดีเซลบี', 'ไออาร์'คาดยอดขายน้ำมันผ่านปั๊มบีโดต่อเนื่องเร่งออกแคมเปญมิดีกระด้นยอดขาย
31/1/2020	'วิเคราะห์สถานการณ์ราคาน้ำมันกพ', 'ราคาน้ำมันดิบปรับตัวลดลงต่อหลังความต้องการใช้น้ำมันในจีนซบเซาจากการระบาดข
30/1/2020	'ปัดทบทวนการค้าโรหมันล้าน', 'ปัดทบทวนการค้าโรหมันล้านบาทเพิ่มขึ้นรับรายได้เพิ่มขึ้นกว่า', 'เมื่อวันทีมคมบริษัทป
29/1/2020	'หุ้นไทยภาคเข้าปีที่จุดลบจุดหรือ', 'กรุงเทพธุรกิจออนไลน์การซื้อขายหลักทรัพย์ภาคเข้าวันพุธมเวลาณดัชนีปิดข
28/1/2020	'พลังงานน้ำมันดีเซลลดสต็อกขึ้นราคาคงเดิม', 'ราคาน้ำมันพุ่งขึ้นดีเซลปรับลดสต็อกขึ้นราคาคงเดิม', 'บมจปตทน้ำมันแ
27/1/2020	'หุ้นไทยภาคเข้าปีที่จุดลบจุดหรือ', 'กรุงเทพธุรกิจออนไลน์การซื้อขายหลักทรัพย์ภาคเข้าวันพฤหัสบดีเวลาณดัชนีปิด
26/1/2020	'พจนททำเรือศรีสุวรรณ'ถกแหลมม้งเรื่องไกรยีนปชสอมนายกเอื้อกลุ่มทุน', 'พนักงนทททก'ศรีสุวรรณ'หนุนสอมนประมุ
25/1/2020	'สนธิรัตน์เร่งหนุนใช้หวังดันราคาสินค้าเกษตร', 'สนธิรัตน์'ประกาศนโยบายพลังงานเดินทางผลักดันยกเลิกกสยสอสม
24/1/2020	'หุ้นไทยภาคเข้าปีที่จุดลบจุดหรือ', 'กรุงเทพธุรกิจออนไลน์การซื้อขายหลักทรัพย์ภาคเข้าวันศุกร์มเวลาณดัชนีปิดข
21/1/2020	'ปัดทบทวนจากปรับขึ้นราคาน้ำมันทุกชนิดลดเว้นขยับสตร', 'ปัดทบทวนจากปรับขึ้นราคาน้ำมันทุกชนิดลดเว้นขยับสตรมีผลพวงนี้'
20/1/2020	'สนธิรัตน์เร่งหนุนใช้หวังดันราคาสินค้าเกษตร', 'สนธิรัตน์'ประกาศนโยบายพลังงานเดินทางผลักดันยกเลิกกสยสอสม
18/1/2020	'หนี้ครัวเรือนระเบิดเวลาเศรษฐกิจปี', 'เทคโนโลยี'กำลังกลายเป็นจุดเริ่มต้นแห่งปัญหาทางการเงินส่วนบุคคลหากไม่มีวินย
17/1/2020	'ท่าเลทองหรือโอกาสทองของใคร', 'แม่โครงการอีอีซีเกิดขึ้นเพียงไม่กี่ปีก็แตกกลายเป็นท่าเลทองที่เกิดการลงทุนอสากร
16/1/2020	'บอทดตทไฟเขียวลงทุนบีเอสแสนบาท', 'บอทดตทไฟเขียวลงทุนบีเอสแสนบาททยอยขยายธุรกิจค้าขายและทอ
15/1/2020	'หุ้นไทยภาคเข้าปีที่จุดลบจุดหรือ', 'กรุงเทพธุรกิจออนไลน์การซื้อขายหลักทรัพย์ภาคเข้าวันพุธมเวลาณดัชนีปิดข
14/1/2020	'ปัจจัยภายนอกหนุน', 'คาดปรับตัวขึ้นปีจุดก่อนจะสั่นแอ่นตัวต่อรับเชิงบวกสหรัฐขึ้นเตรียมลงนามข้อตกลงการค้าเฟสแ

รูปที่ 3 ข้อความข่าวในแต่ละวัน

การเตรียมข้อมูล

ขั้นตอนนี้เป็นขั้นตอนการแปลงข้อมูลที่เก็บรวบรวม (raw data) ได้ให้กลายเป็นข้อมูลที่สามารถนำมาวิเคราะห์ โดยการแปลงข้อมูลนี้จะเป็นกระบวนการที่ใช้เวลามากที่สุด ซึ่งในงานวิจัยนี้มีขั้นตอนดังนี้

1. รวมข้อมูลข้อความข่าวและข้อมูลหุ้นรายวันที่ได้จากขั้นตอนแรก โดยกำหนดข้อมูลสถานะของราคาปิดเป็นตัวแปร y ดังรูปที่ 4

Date	Close	Diff	Close_status	Text
27/11/2019	69.5	-1	negative	'วิเคราะห์สถานการณ์ราคาน้ำมันกพ', 'ราคาน้ำมันดิบปรับตัวลดลงต่อหลังความต้องการใช้น้ำมันในจีนซบเซาจากการระบาดข
28/11/2019	69	-0.5	negative	'อาจหันเหจากข่าวกฎหมายใหม่ของสหรัฐแต่ยังมองโลกโง่เง่า', 'สงครามการค้าไม่แยกไปกว่าที่เป็นอยู่', 'เข้าปีน
29/11/2019	69	0	neutral	'แก๊งค์ไต้หวันที่มีปัจจัยบวกในช่วงเดือนธ', 'บรรยากาศลงทุนช่วงสิ้นปีชะลอตัวตามการรื้อการค้า', 'หลังประธานาธิบดี
2/12/2019	68.25	-0.75	negative	'ความคาดหวังโอบล้อมกำลังการผลิตจีนเป็นบวกหนุนบรรยากาศเชิงบวกไปทั่วสิ้น', 'จีนกดดันสหรัฐปรับลดภาษีลง
3/12/2019	68.5	0.25	positive	'หุ้นบีโอดีเล็ก'กอดคอขึ้นหุ้นแรงจากการค้าโลกงสวย', 'หุ้นกลุ่มบีโอดี'และ'อีเล็กทรอนิกส์'พุ่งกว่าช่วงหนึ่งเดือน
4/12/2019	68	-0.5	negative	'วิเคราะห์สถานการณ์ราคาน้ำมันกพ', 'ราคาน้ำมันดิบปรับตัวลดลงต่อหลังความต้องการใช้น้ำมันในจีนซบเซาจากการระบาดข
6/12/2019	66.75	-1.25	negative	'สำนักข่าวหุ้นอินโอดี'ความเคลื่อนไหวเรื่องคดีทุจริตในโอดี'ข่าวรองกรรมการผู้อำนวยการด้านการตลาดบริษัทหลักทรัพย์
9/12/2019	66.75	0	neutral	'กองทุนหุ้นไทย-ลงทุนที่ทวนปีสุดท้าย', 'ปีสุดท้ายของกองทุนรวมระยะยาวยังเป็นตัวชี้วัดสำคัญในการลดหย
11/12/2019	64.5	-2.25	negative	'ปัดทบทวนจากปรับขึ้นราคาน้ำมันทุกชนิดลดเว้นขยับสตร', 'ปัดทบทวนจากปรับขึ้นราคาน้ำมันทุกชนิดลดเว้นขยับสตรมีผลพวงนี้'
12/12/2019	65	0.5	positive	'ลิแกเนอโลบปฎิรูปปาล์มน้ำมันโครได้ผลประโยชน์', 'ห้องข่าวเศรษฐกิจธุรกิจหุ้นแล้วสักทุกมิติเรื่องเศรษฐกิจการลง
13/12/2019	67.5	2.5	positive	'วิเคราะห์สถานการณ์ราคาน้ำมันกพ', 'ราคาน้ำมันดิบปรับตัวเพิ่มขึ้นเล็กน้อยจากความหวังปริมาณน้ำมันดิบลดจากข้อ
16/12/2019	68	0.5	positive	'รีอีไอเกียรดินาคินตัวเต็งเข้า', 'เวลาช่วงผ่านไปเร็วเหลือเกินเหลือแค่ปีเดียวอีกไม่กี่วันก็จะสิ้นปีแล้วไม่ทราบว่พร
17/12/2019	67.5	-0.5	negative	'กับขงมีอเวร็ดแบงก์กระด้นลงทุน', 'กับขงส่นความร่วมมือกับธนาคารโลกยกระดับการลงทุนเพื่อความยั่งยืน
18/12/2019	69	1.5	positive	'"อรรถพลวิรัตน์"ชิงชัยโอดี', 'คณะกรรมการสรรหาของบอทดตทคณเลือก'ชัยโอดี'ปัดทบทใหม่หลังโชวีรภัยที่
19/12/2019	68.25	-0.75	negative	'โดยมีจุดขายคดขาดทุน', 'คาดการณ์ตลาดหุ้นไทยวันนี', 'เรามีมุมมองและคาดดัชนีแนวโน้มด้านจุดแนวรับจุดเป็นหลา
20/12/2019	68	-0.25	negative	'กลุ่มปัดทงตงบนแสนล้านลยธุรกิจไทยต', 'ห้องข่าวเศรษฐกิจธุรกิจหุ้นแล้วสักทุกมิติเรื่องเศรษฐกิจการลงทนอก
23/12/2019	70	2	positive	'หุ้นไทยแกว่งตัวกรอบแคบจุด', 'โบรคาคาดหุ้นไทยแกว่งตัวจุดแคบได้แรงหนุนสหรัฐขึ้นบรรลจุดลดงบการการค้าเพ
24/12/2019	69.5	-0.5	negative	'คาดแกว่งตัวจุดเนื่องจากภาวะตลาดรอปัจจัยใหม่กระด้นการลงทุน', 'ตลาดหุ้นวานนี้', 'ปรับขึ้นแรงจุดปิดที่ระดับจุด
25/12/2019	69.5	0	neutral	'หุ้นไทยแกว่งตัวแคบกรอบแคบจุด', 'บลกรูมมองหุ้นไทยวันนี'เคลื่อนไหวตามบททดสอบจุดก่อนจะสั่นแอ่น

รูปที่ 4 ข้อมูลสถานะของราคาปิดรวมกับข้อมูลข้อความข่าว

2. ลบประโยคที่ไม่มีผลต่อการพยากรณ์ทิศทางของหุ้นออก เช่น 'มูลค่าการซื้อขายล้านบาทที่บาทเพิ่มขึ้นบาท', 'มูลค่าการซื้อขายล้านบาทที่บาทลดลงบาท', 'ที่มา' เป็นต้น
3. Word Tokenize เป็นการตัดข้อความข่าวแยกออกมาเป็นคำโดยใช้ Library pythainlp ซึ่งในงานวิจัยนี้ใช้ 9 วิธีคือ

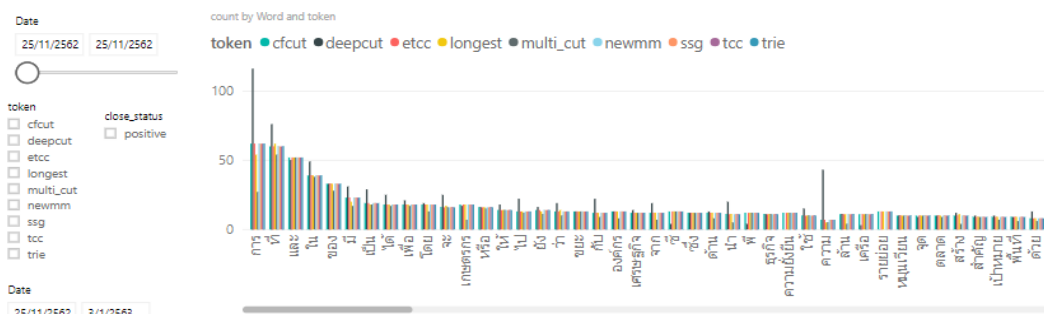
- 1) Cfcut เป็นการแบ่งประโยคภาษาไทยโดยใช้เทคนิคคอนดิชันนอลแรนดอมฟิลด์ (conditional random field) ซึ่งเรียนรู้จากชุดข้อมูลของ TED
- 2) Deepcut เป็นการตัดคำโดยใช้เทคนิค 1D Convolution Neural Network หรือ CNN ซึ่งเป็นโครงข่ายประสาทเทียมชนิดหนึ่งในกลุ่ม Bio-inspired โดยที่ CNN จะจำลองการมองเห็นของมนุษย์ที่มองพื้นที่เป็นที่ย่อยๆ และนำกลุ่มของพื้นที่ย่อยๆ มาผสานกัน
- 3) Longest เป็นการตัดคำโดยวิธีการเทียบคำที่ยาวที่สุด (longest matching) เสนอโดย Sonlertlamvanit (1993) ซึ่งเป็นวิธีที่ใช้พจนานุกรมช่วยในการตัดคำภาษาไทย โดยวิธีนี้จะทำการตรวจสอบหรือแสกนข้อความจากซ้ายไปขวาแล้วนำไปเปรียบเทียบกับพจนานุกรม ถ้าอักษรที่ประกอบกันนั้นไม่สามารถเทียบเป็นคำในพจนานุกรมได้ก็จะทำการลดความยาวลง จนกว่าจะเทียบคำในพจนานุกรมได้แล้วกลับไปยังจุดย้อนกลับ จากนั้นก็เริ่มทำงานที่จุดย้อนกลับอีกครั้งเพื่อทำการตรวจสอบอักษรที่ประกอบกัน (Chaicharoen, 2001) เช่น “ห้วงใยสถานการณ์การระบาดของไวรัสโคโรนา” ข้อความนี้ไม่สามารถเทียบคำพจนานุกรมได้จึงตัดเหลือ “ห้วงใยสถานการณ์การระบาดของไวรัสโคโรนา” ตัดไปเรื่อยๆ จนได้สายอักขระ “ห้วงใย” ซึ่งสามารถเทียบคำในพจนานุกรมได้
- 4) Etc เป็นวิธีการตัดคำโดยใช้เทคนิคการรวมกันของ Forward และ Backward Longest Matching Techniques
- 5) multi_cut เป็นการตัดคำด้วยวิธีการตัดคำเพื่อให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด (maximal matching) เป็นระเบียบวิธีฮิวริสติกส์ (heuristic) วิธีหนึ่งที่ใช้พจนานุกรมเทียบคำภาษาไทย พัฒนาโดย Sonlertlamvanit (1993) ซึ่งได้ดัดแปลงมาจากวิธีการเทียบคำที่ยาวที่สุด เริ่มจากการหาทางเลือกของรูปแบบการตัดคำทั้งหมดที่เป็นไปได้เสียก่อน โดยทำย้อนกลับทีละคำ หลังจากได้คำตอบจากวิธีการเทียบคำที่ยาวที่สุดแล้ว จึงเลือกทางเลือกที่มีจำนวนคำน้อยที่สุด (Chaicharoen, 2001)
- 6) tcc เป็นการตัดคำโดยใช้เทคนิคกลุ่มอักขระภาษาไทย หรือ Thai Character Clusters (TCC) ในภาษาไทยมีลักษณะขององค์ประกอบที่หลากหลายมากกว่าเมื่อเปรียบเทียบกับภาษาอังกฤษ เช่น สระ วรรณยุกต์ เสียงวรรณยุกต์ และอักขระพิเศษ เทคนิค TCC นั้นองค์ประกอบต่างๆ ในวลีจะไม่คลุมเครือและสามารถตั้งกฎขึ้นมาได้ เช่น สระที่อยู่ข้างหน้าและอักขระถัดไปถูกจัดกลุ่มให้อยู่ในกลุ่มเดียวกัน เครื่องหมายวรรณยุกต์จะอยู่บนพยัญชนะเสมอ และไม่สามารถแบ่งออกจากพยัญชนะได้ ส่วนสระที่อยู่ตำแหน่งและตำแหน่งก่อนหน้าจะถูกแบ่งให้อยู่ในกลุ่มเดียวกัน เป็นต้น
- 7) newmm เป็นการตัดคำโดยอ้างอิงจากพจนานุกรมโดยใช้เทคนิค Maximal Matching Algorithm และ Thai Character Cluster (TCC)
- 8) ssg เป็นการตัดคำโดยใช้พยางค์ด้วยเทคนิค Conditional Random Field
- 9) trie เป็นการตัดคำโดยใช้พยางค์ด้วยเทคนิคทรี (trie) ในวิทยาศาสตร์คอมพิวเตอร์ Trie เรียกอีกชื่อว่า Digital Tree หรือ Prefix Tree คือโครงสร้างของข้อมูลต้นไม้ที่เป็นลำดับเพื่อเก็บข้อมูลเป็นแบบไดนามิกหรืออาร์เรย์ที่เชื่อมโยงกัน ใช้กับข้อมูลที่เป็นตัวอักษร การทำงานเป็นแผนผังการค้นหาแบบไบนารี โดยข้อมูลอยู่บนเส้นทางการเดินระหว่าง

จะมีตัวอักษรกำกับอยู่

[illegible]

รูปที่ 5 การตัดคำและสถานะของราคาปิด

4. Word frequency และ Word Index คือ การหาความถี่ของของคำแต่ละคำในแต่ละวัน รวมทั้งการสร้าง Index ให้แต่ละคำด้วย เช่น กำหนดให้คำว่า “การ” มี Index เป็น 1 ,คำว่า “ที่” มี Index เป็น 2 ,และคำว่า “มี” มี Index เป็น 3 เป็นต้น ดังรูปที่ 6 และ 7



รูปที่ 6 ความถี่ของคำในแต่ละวันที่ตัดคำด้วยวิธีการตัดคำต่างๆ ทั้ง 9 วิธี

close_status	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
positive	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	1	1	0
positive	0	0	0	0	0	0	0	0	0	0	0	4	8	0	0	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	2	0	1	0	5	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
neutral	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	2	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0
positive	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	2	0	2	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
neutral	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	1	0	0	0	0
positive	2	0	0	0	0	1	0	2	2	0	0	0	0	13	0	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0
positive	0	0	0	0	0	0	0	0	0	0	0	3	5	0	0	0	0	0	0	0	0	0	0
negative	0	0	0	1	0	0	0	0	0	1	0	0	0	6	0	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	2	0	0	3	0	1	0	11	0	0	0	0	0	0	0	0

รูปที่ 7 สถานะราคาปิดของหุ้นและความถี่ของแต่ละ Word Index

5. TFIDF (term frequency-inverse document frequency) เป็นกระบวนการที่แปลงความถี่ในเอกสาร โดยให้น้ำหนักคำในเอกสารใดๆ เทียบกับคำในเอกสารทั้งหมด ซึ่งสามารถหาได้จากสมการ

$$TFIDF = \text{Term Frequency} \times \text{Inverse Document Frequency} \quad (2)$$

โดยค่า Term Frequency หรือ tf นั้นมีสมการดังนี้

$$tf(w, D) = f_{wD} \quad (3)$$

กำหนดให้ w คือคำที่ต้องการคำนวณ และ D คือ เอกสารที่สนใจ

ส่วนค่า Inverse Document Frequency หรือ idf มีสมการดังนี้

$$idf(t) = 1 + \log \frac{C}{1 + df(t)} \quad (4)$$

กำหนดให้ C คือ จำนวนเอกสารทั้งหมดที่มีอยู่ และ $df(t)$ คือ จำนวนของเอกสารที่มีค่า t ปรากฏอยู่

close_status	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
positive	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0	0	0	0	0	0.02	0.02	0
positive	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0	0	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	0.03	0.01	0	0.03	0	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0.02	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0
neutral	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.01	0	0.03	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.05	0	0	0	0	0	0	0	0
positive	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0.1	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0	0	0	0	0	0
neutral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0.01	0	0	0
positive	0.01	0	0	0	0	0	0.01	0	0.01	0.01	0	0	0	0	0.02	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0	0	0	0	0	0	0	0
positive	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0.02	0
positive	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0.03	0	0	0	0	0	0	0	0
negative	0	0	0	0.02	0	0	0	0	0	0	0.01	0	0	0	0.02	0	0	0	0	0	0	0	0
negative	0	0	0	0	0	0	0	0.01	0	0	0.01	0	0	0	0.02	0	0	0	0	0	0	0	0.01

รูปที่ 8 สถานะราคาปิดของหุ้นและค่า TFIDF เพื่อให้น้ำหนักคำในแต่ละ Word Index

6. Training และ Test Dataset ทำการแบ่งข้อมูลเป็น 2 ส่วนคือข้อมูลฝึกสอน (training dataset) จำนวนร้อยละ 80 ของข้อมูลทั้งหมด และข้อมูลทดสอบ (test dataset) จำนวนร้อยละ 20 ของข้อมูลทั้งหมด

การสร้างแบบจำลอง

เป็นขั้นตอนการวิเคราะห์ข้อมูล ในงานวิจัยนี้ได้ใช้เทคนิคการเรียนรู้ของเครื่อง (machine learning) ประเภทการเรียนรู้แบบมีผู้สอน (supervised learning) ซึ่งใช้แบบจำลองการจำแนก (classification model) ทั้งหมด 7 ตัวแบบคือ 1. K-neighbors Classifier, 2. Logistic Regression, 3. Gradient Boosting Classifier, 4. Random Forest Classifier, 5. AdaBoost Classifier, 6. XGB Classifier และ 7. Support Vector Machine (SVM) โดยตัวแบบ K-neighbors Classifier, Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, AdaBoost Classifier และ Support Vector Machine

(SVM) ใช้ Library sklearn ช่วยในการประมวลผลและตัวแบบ XGB Classifier ใช้ Library xgboost ช่วยในการประมวลผล

การประเมินผล

ในงานวิจัยนี้ใช้ค่าความแม่นยำ (accuracy) เป็นตัววัดประสิทธิภาพของตัวแบบการจำแนก โดยคำนวณจาก Confusion Matrix ซึ่งเป็นตารางค่าพยากรณ์เปรียบเทียบกับข้อมูลจริง และใช้ Library sklearn import cross_val_score ช่วยในการคำนวณ

การใช้งาน

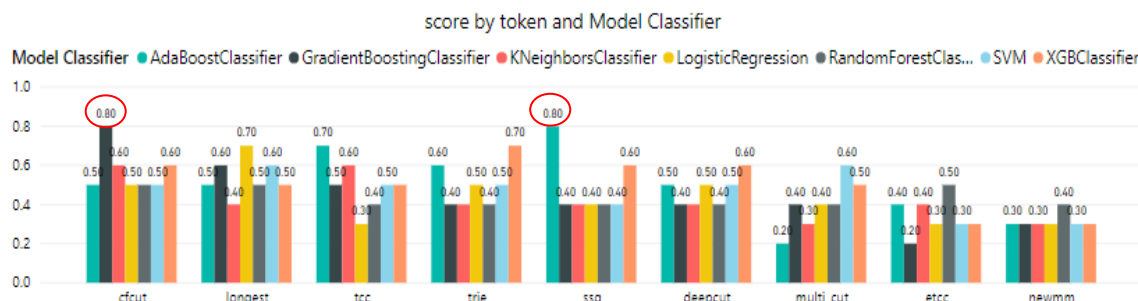
หลังจากเลือกตัวแบบที่มีค่าความแม่นยำสูงที่สุดแล้วจึงนำตัวแบบไปใช้งาน โดยสรุปเป็นขั้นตอนการนำไปใช้งานดังนี้

1. เก็บรวบรวมและเตรียมข้อความข่าวก่อนถึงวันที่ต้องการพยากรณ์ 1 วัน และทำการแบ่งข้อมูลเป็น Training Dataset จำนวน 80% ของข้อมูลทั้งหมดและ Test Dataset จำนวน 20% ของข้อมูลทั้งหมด
2. สร้างแบบจำลองด้วยข้อมูล Training Dataset และใช้ Test Dataset ประเมินประสิทธิภาพของแบบจำลอง (model) เพื่อหาแบบจำลอง และวิธีการตัดคำ (Tokenizer) ที่มีค่าความถูกต้องแม่นยำของการพยากรณ์สูงสุดจากการเรียนรู้ทั้งหมด 5 ครั้ง ซึ่งจะใช้คำสั่ง predict() ใน Library sklearn ทำให้ที่ผลออกมาจะอยู่ในรูปของ ทิศทางบวก (positive), ทิศทางลบ (negative) และเป็นกลาง (neutral)
3. นำตัวแบบที่ได้ในข้อ 2. มาพยากรณ์ทิศทางของราคาหุ้นจากข้อความข่าววันถัดมา

ผลการวิจัย

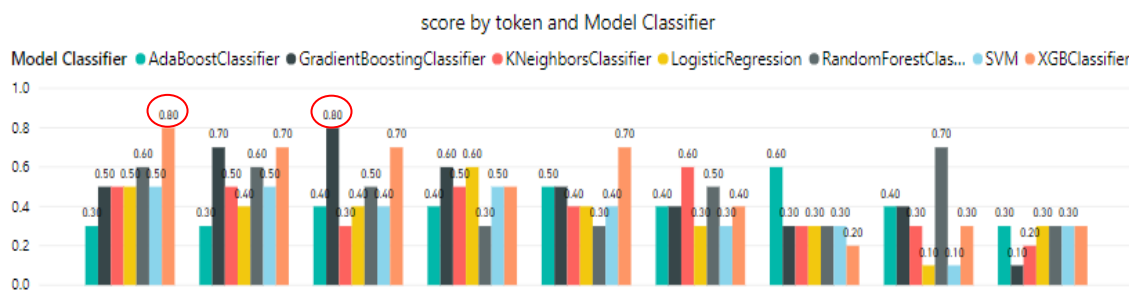
หุ้กลุ่มเทคโนโลยีสารสนเทศและการสื่อสาร (ICT)

ผู้วิจัยได้สุ่มเลือกหุ้กลุ่มเทคโนโลยีสารสนเทศและการสื่อสาร (ICT) 1 ตัว คือ หุ้ของบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) ซึ่งมีธุรกิจด้านการลงทุนในธุรกิจโทรคมนาคม สื่อ และเทคโนโลยี ในการวิจัยครั้งนี้ได้ทำการพยากรณ์ทิศทางของราคาหุ้นจากข้อความข่าวจำนวน 3 วัน ซึ่งมีวันที่ 5, 6 และ 7 กุมภาพันธ์ 2563 โดยแบ่งข้อความสำหรับฝึกสอน (train) และทดสอบตัวแบบ (test) เพื่อหาตัวแบบและการตัดคำที่มีค่าความแม่นยำ (accuracy) สูงที่สุดแบบสุ่ม (หมายความว่าถ้าค่าความแม่นยำสูงที่สุดมีหลายตัวแบบ โปรแกรมจะเลือกตัวแบบที่เหมาะสมโดยวิธีการสุ่ม) เพื่อนำมาพยากรณ์ทิศทางของราคาหุ้นบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) โดยพยากรณ์ซ้ำทั้งหมด 5 ครั้งในแต่ละวัน และเลือกตัวแบบที่ดีที่สุดจาก 5 ครั้งนี้ ได้ผลดังรูปที่ 9, 10 และ 11



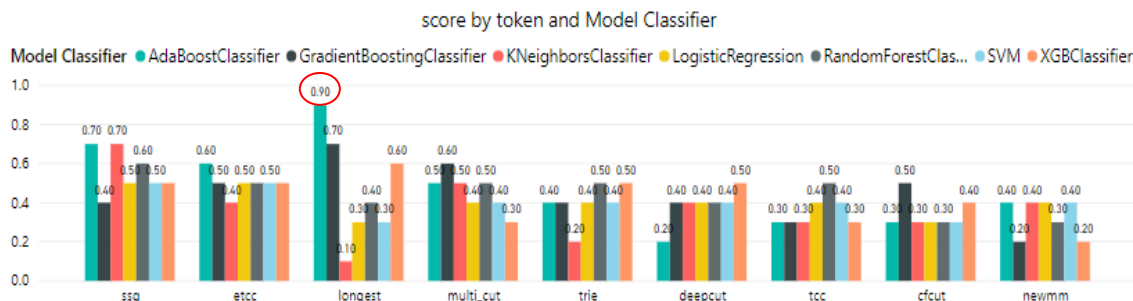
รูปที่ 9 ค่าความแม่นยำ (accuracy) ของหุ่นบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 5 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 4 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 1

จากรูปที่ 9 จะเห็นได้ว่าการทำซ้ำครั้งที่ 1 (จากทั้งหมด 5 ครั้ง) ตัวแบบที่มีค่าความแม่นยำสูงที่สุด (0.8) มี 2 ตัวแบบคือตัวแบบ Gradient Boosting Classifier (วิธีการตัดคำ cfcut) และตัวแบบ AdaBoost Classifier (วิธีการตัดคำ ssg) โปรแกรมจะทำการเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ซึ่งในการประมวลผลครั้งนี้เลือกตัวแบบ Gradient Boosting Classifier เป็นตัวแบบที่ดีที่สุด โดยวิธีการตัดคำคือ cfcut



รูปที่ 10 ค่าความแม่นยำ (accuracy) ของหุ่นบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 6 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 5 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 4

จากรูปที่ 10 จะเห็นได้ว่าการทำซ้ำครั้งที่ 4 ตัวแบบที่มีค่าความแม่นยำสูงที่สุด (0.8) มี 2 ตัวแบบคือตัวแบบ XGB Classifier (วิธีการตัดคำ etcc) และตัวแบบ Gradient Boosting Classifier (วิธีการตัดคำ trie) โปรแกรมจะทำการเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ซึ่งในการประมวลผลครั้งนี้เลือกตัวแบบ Gradient Boosting Classifier และวิธีการตัดคำคือ trie



รูปที่ 11 ค่าความแม่นยำ (accuracy) ของหุ่นบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 7 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 6 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 4

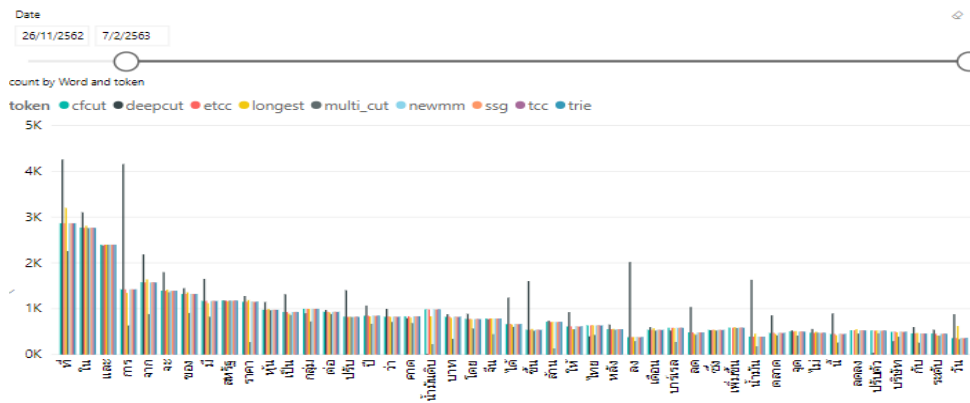
ตารางที่ 1 การพยากรณ์ทิศทางของราคาหุ้นของบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH)

ค่าความแม่นยำ	ตัวแบบการจำแนก	การตัดคำ	หุ้น	วันที่พยากรณ์	ทำซ้ำครั้งที่	ค่าพยากรณ์	ค่าจริง
0.8	GradientBoostingClassifier	cfcut	INTUCH	5/2/2020	1	positive	positive
0.8	GradientBoostingClassifier	trie	INTUCH	6/2/2020	4	negative	negative
0.9	AdaBoostClassifier	longest	INTUCH	7/2/2020	4	negative	negative

ผลการพยากรณ์ทิศทางของราคาหุ้นของบริษัท อินทัช โฮลดิ้งส์ จำกัด (INTUCH) นั้น เมื่อเปรียบเทียบกับข้อมูลจริงของวันที่ 5, 6 และ 7 กุมภาพันธ์ 2563 พบว่ามีความถูกต้องร้อยละ 100 ตามที่ได้แสดงไว้ในตารางที่ 1

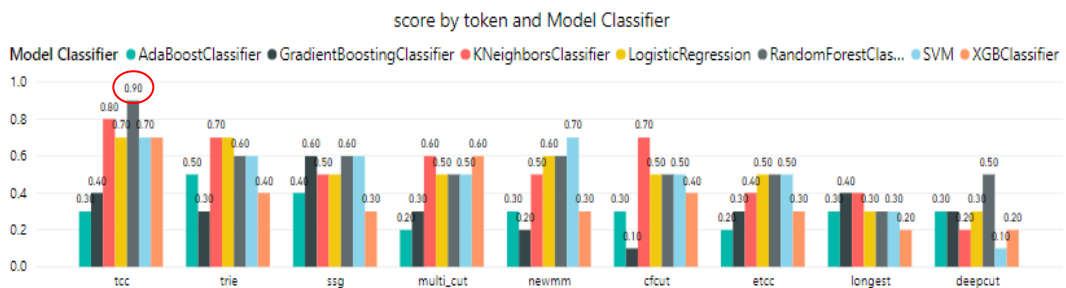
ห้่นกลุ่มพลังงานและสาธารณูปโภค (ENERG)

ผู้วิจัยได้ทำการสุ่มเลือกหุ้นกลุ่มพลังงานและสาธารณูปโภค (ENERG) 1 ตัว คือ หุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) ซึ่งเป็นผู้ประกอบการกลั่นและจำหน่ายน้ำมันปิโตรเลียมที่ใหญ่ที่สุดในประเทศไทย และมีโรงงานที่มีประสิทธิภาพสูงในภูมิภาคเอเชียแปซิฟิก มีธุรกิจที่หลากหลาย เช่น ธุรกิจการกลั่นน้ำมัน ธุรกิจปิโตรเคมีและธุรกิจน้ำมันหล่อลื่นพื้นฐาน รวมทั้งธุรกิจธุรกิจไฟฟ้า ธุรกิจสารทำละลาย เป็นต้น จากการตัดคำด้วย Library pythainlp ซึ่งมี 9 วิธี คือ 'cfcut', 'deepcut', 'etcc', 'longest', 'multi_cut', 'newmm', 'ssg', 'tcc' และ 'trie' ความถี่ของคำตั้งแต่วันที่ 26 พฤศจิกายน 2562 จนถึง วันที่ 7 กุมภาพันธ์ 2563 แสดงดังรูปที่ 12

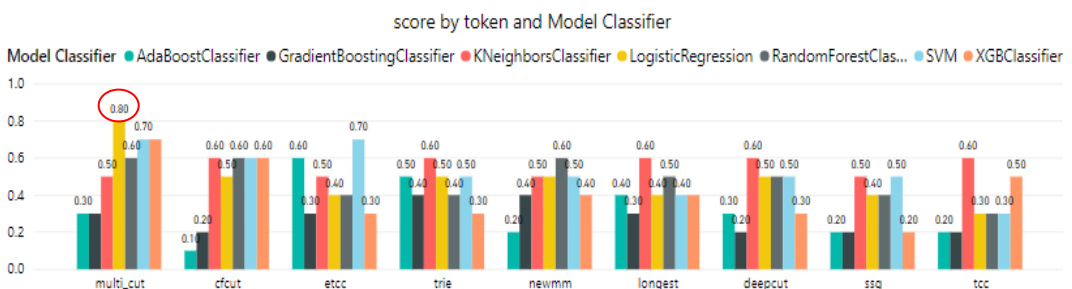


รูปที่ 12 ความถี่ของคำของหุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP)

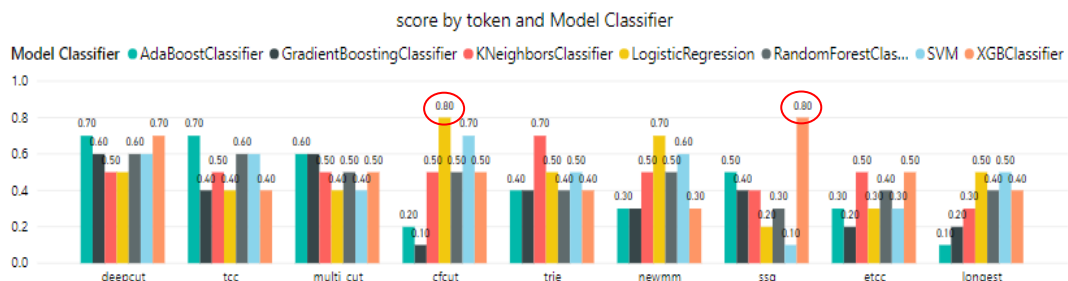
นำข้อมูลมาใช้ในการสร้างตัวแบบการเรียนรู้เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 5, 6 และ 7 กุมภาพันธ์ 2563 โดยพิจารณาจากค่าความแม่นยำของตัวแบบการจำแนกและวิธีการตัดคำโดย Library pythainlp แสดงผลดังรูปที่ 13, 14 และ 15



รูปที่ 13 ค่าความแม่นยำของหุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 5 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 26 พฤศจิกายน 2562 จนถึง วันที่ 4 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 2



รูปที่ 14 ค่าความแม่นยำของหุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 6 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 26 พฤศจิกายน 2562 จนถึง วันที่ 5 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 1



รูปที่ 15 ค่าความแม่นยำของหุ่นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 7 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 26 พฤศจิกายน 2562 จนถึง วันที่ 6 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 3

จากรูปที่ 15 จะเห็นได้ว่าการทำซ้ำครั้งที่ 3 ตัวแบบที่มีค่าความแม่นยำสูงสุด (0.8) มี 2 ตัวแบบคือตัวแบบ Logistic Regression (วิธีการตัดคำ cfcut) และตัวแบบ XGB Classifier (วิธีการตัดคำ ssg) โปรแกรมจะทำการเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ซึ่งในการประมวลผลครั้งนี้เลือกตัวแบบ XGB Classifier และวิธีการตัดคำคือ ssg

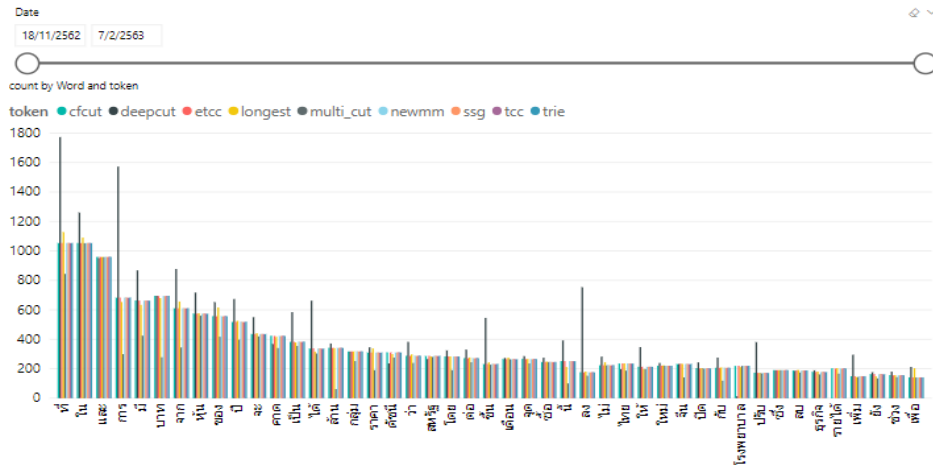
ตารางที่ 2 การพยากรณ์ทิศทางของราคาหุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP)

ค่าความแม่นยำ	ตัวแบบการจำแนก	การตัดคำ	หุ้น	วันที่พยากรณ์	ทำซ้ำครั้งที่	ค่าพยากรณ์	ค่าจริง
0.9	RandomForestClassifier	tcc	TOP	5/2/2020	2	negative	positive
0.8	LogisticRegression	multi_cut	TOP	6/2/2020	1	positive	positive
0.8	XGBClassifier	ssg	TOP	7/2/2020	3	negative	negative

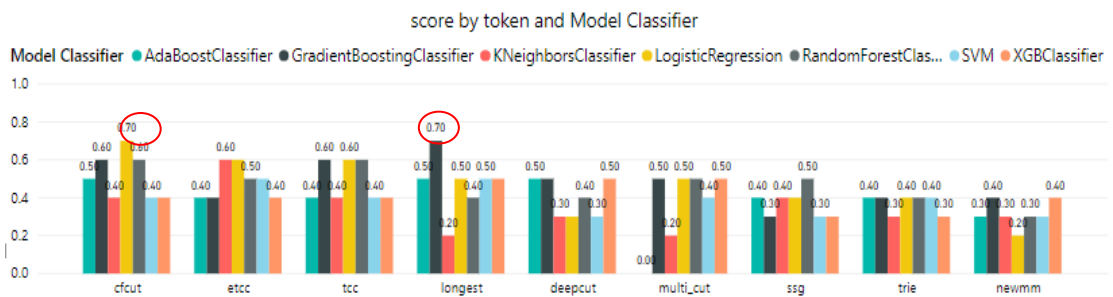
ค่าพยากรณ์ทิศทางของราคาหุ้นบริษัท ไทยออยล์ จำกัด (มหาชน) (TOP) นั้น เมื่อเปรียบเทียบกับข้อมูลจริงของวันที่ 5, 6, 7 กุมภาพันธ์ 2563 มีความถูกต้องร้อยละ 66.67 ตามที่ได้แสดงไว้ในตารางที่ 2

หุ่นกลุ่มการแพทย์ (HEALTH)

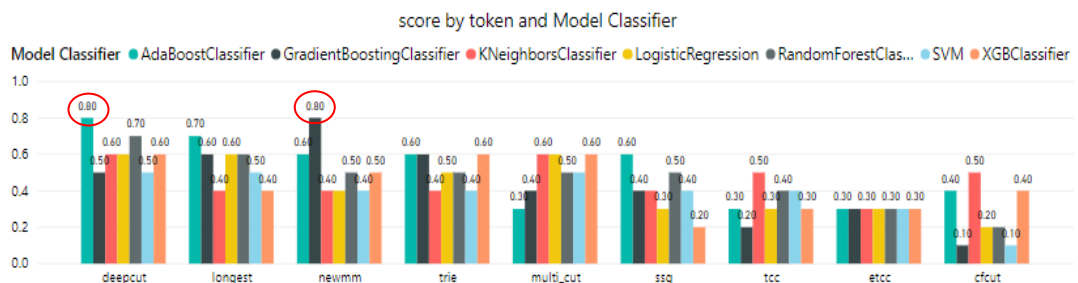
ทางผู้วิจัยได้ทำการสุ่มเลือกหุ่นกลุ่มการแพทย์ (HEALTH) 1 ตัว คือบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) ซึ่งเป็นธุรกิจโรงพยาบาลเอกชนที่ให้บริการทางการแพทย์ทั้งผู้ป่วยในและผู้ป่วยนอก รวมทั้งผู้ป่วยต่างชาติ อีกทั้งยังมีการลงทุนในธุรกิจทางการแพทย์ทั้งในและต่างประเทศ จากการตัดคำด้วย Library pythainlp แสดงความถี่ของคำตั้งแต่วันที่ 18 พฤศจิกายน 2562 จนถึง วันที่ 7 กุมภาพันธ์ 2563 ดังรูปที่ 16 และผลการพยากรณ์ทิศทางของราคาหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) แสดงดังรูปที่ 17, 18 และ 19



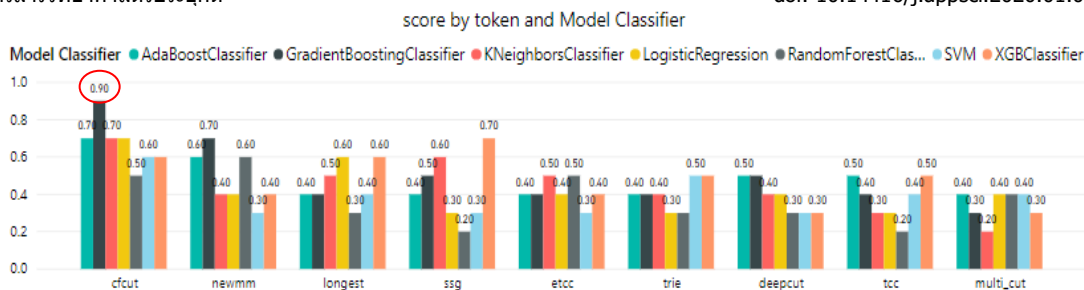
รูปที่ 16 ความถี่ของคำของหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH)



รูปที่ 17 ค่าความแม่นยำของหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 5 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 18 พฤศจิกายน 2562 จนถึง วันที่ 4 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 1



รูปที่ 18 ค่าความแม่นยำของหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 6 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 18 พฤศจิกายน 2562 จนถึง วันที่ 5 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 4



รูปที่ 19 ค่าความแม่นยำของหุ่นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) เพื่อพยากรณ์ทิศทางของราคาหุ้น วันที่ 7 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 18 พฤศจิกายน 2562 จนถึง วันที่ 6 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 1

จากรูปที่ 17 จะเห็นได้ว่าการทำซ้ำครั้งที่ 1 ตัวแบบที่มีค่าความแม่นยำสูงที่สุด (0.7) มี 2 ตัวแบบคือตัวแบบ Logistic Regression (วิธีการตัดคำ cfcut) และตัวแบบ Gradient Boosting Classifier (วิธีการตัดคำ longest) โปรแกรมจะทำการเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ซึ่งในการประมวลผลครั้งนี้เลือกตัวแบบ Logistic Regression และวิธีการตัดคำคือ cfcut

จากรูปที่ 18 จะเห็นได้ว่าการทำซ้ำครั้งที่ 4 ตัวแบบที่มีค่าความแม่นยำสูงที่สุด (0.8) มี 2 ตัวแบบคือตัวแบบ AdaBoost Classifier (วิธีการตัดคำ deepcut) และตัวแบบ Gradient Boosting Classifier (วิธีการตัดคำ newmm) โปรแกรมจะทำการเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ซึ่งในการประมวลผลครั้งนี้เลือกตัวแบบ AdaBoost Classifier และวิธีการตัดคำคือ deepcut

ตารางที่ 3 การพยากรณ์ทิศทางของราคาหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH)

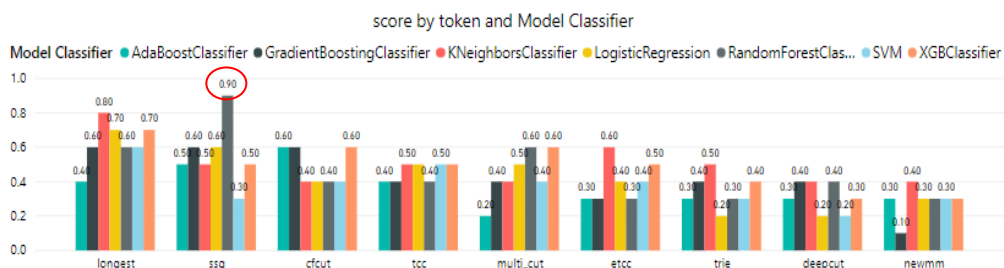
ค่าความแม่นยำ	ตัวแบบการจำแนก	การตัดคำ	หุ้น	วันที่พยากรณ์	ทำซ้ำครั้งที่	ค่าพยากรณ์	ค่าจริง
0.7	LogisticRegression	cfcut	BH	5/2/2020	1	negative	negative
0.8	AdaBoostClassifier	deepcut	BH	6/2/2020	4	negative	negative
0.9	GradientBoostingClassifier	cfcut	BH	7/2/2020	1	negative	positive

ผลการพยากรณ์ทิศทางของราคาหุ้นบริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด (มหาชน) (BH) นั้น เมื่อเปรียบเทียบกับข้อมูลจริงในวันที่ 5 , 6 และ 7 กุมภาพันธ์ 2563 พบว่ามีความถูกต้องร้อยละ 66.67 ตามที่ได้แสดงไว้ในตารางที่ 3

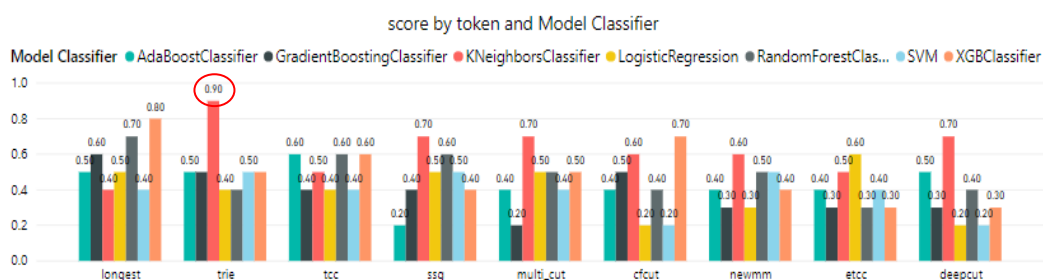
หุ่นกลุ่มพาณิชย์ (COMM)

ทางผู้วิจัยได้ทำการสุ่มเลือกหุ่นกลุ่มพาณิชย์ (COMM) 1 ตัว คือ บริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) ซึ่งเป็นธุรกิจร้านค้าปลีกภายใต้เครื่องหมายการค้า 7-Eleven ในประเทศไทย และลงทุนในธุรกิจผลิตและจำหน่ายอาหาร

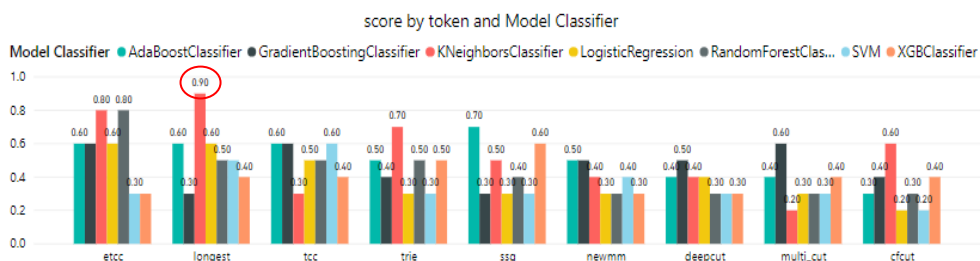
สำเร็จรูปและเบเกอร์รี่ ตัวแทนรับชำระเงินค่าสินค้าและบริการ ผลการเรียนรู้ข้อมูลเพื่อพยากรณ์ทิศทางของราคาหุ้นบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) แสดงดังรูปที่ 20, 21 และ 22



รูปที่ 20 ค่าความแม่นยำของหุ้นบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 5 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 4 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 1



รูปที่ 21 ค่าความแม่นยำของหุ้นบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 6 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 5 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 2



รูปที่ 22 ค่าความแม่นยำของหุ้นบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 7 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 25 พฤศจิกายน 2562 จนถึง วันที่ 6 กุมภาพันธ์ 2563 และเลือกตัวแบบในการทำซ้ำครั้งที่ 3

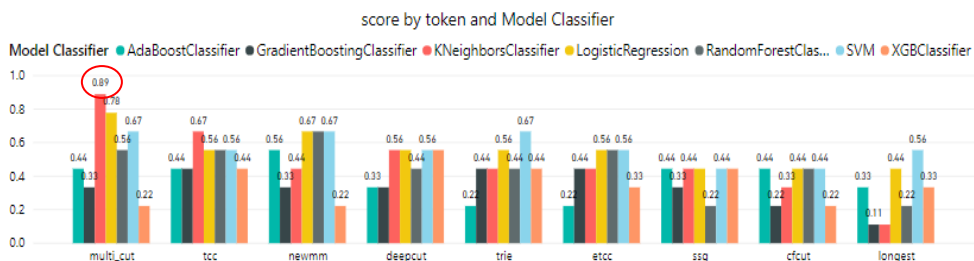
ตารางที่ 4 การพยากรณ์ทิศทางของราคาหุ้นของบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL)

ค่าความแม่นยำ	ตัวแบบการจำแนก	การตัดคำ	หุ้น	วันที่พยากรณ์	ทำซ้ำครั้งที่	ค่าพยากรณ์	ค่าจริง
0.9	RandomForestClassifier	ssg	CPALL	5/2/2020	1	positive	positive
0.9	KNeighborsClassifier	trie	CPALL	6/2/2020	2	negative	positive
0.9	KNeighborsClassifier	longest	CPALL	7/2/2020	3	positive	positive

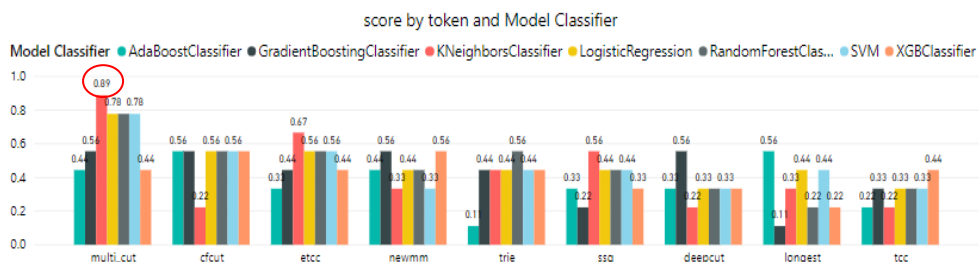
ผลการพยากรณ์ทิศทางของราคาหุ้นบริษัท ซีพี ออลล์ จำกัด (มหาชน) (CPALL) นั้น เมื่อเปรียบเทียบกับข้อมูลจริงของวันที่ 5, 6, 7 กุมภาพันธ์ 2563 มีความถูกต้องร้อยละ 66.67 ตามที่ได้แสดงไว้ในตารางที่ 4

หุ้นกลุ่มธนาคาร (BANK)

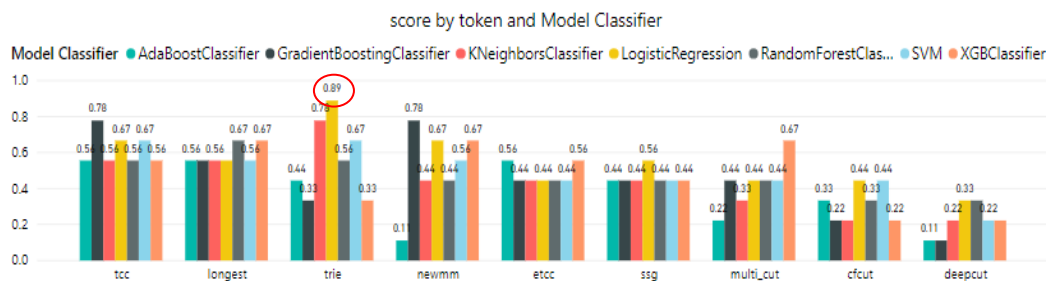
ผู้วิจัยได้ทำการสุ่มเลือกหุ้นกลุ่มธนาคาร (BANK) 1 ตัว คือ ธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) ซึ่งประกอบธุรกิจธนาคารพาณิชย์ ธุรกิจหลักทรัพย์และธุรกิจที่ได้รับอนุญาตไว้ในพระราชบัญญัติธุรกิจสถาบันการเงินฯ และพระราชบัญญัติหลักทรัพย์และตลาดหลักทรัพย์ฯ และประกาศที่เกี่ยวข้อง ณ วันที่ 31 ธันวาคม 2561 ผลการเรียนรู้ข้อมูลเพื่อพยากรณ์ทิศทางของราคาหุ้นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) แสดงดังรูปที่ 23, 24 และ 25



รูปที่ 23 ค่าความแม่นยำของหุ้นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 5 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 28 พฤศจิกายน 2562 จนถึง วันที่ 4 กุมภาพันธ์ 2563 และเลือก ตัวแบบในการทำซ้ำครั้งที่ 1



รูปที่ 24 ค่าความแม่นยำของหุ้นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) เพื่อพยากรณ์ทิศทางของราคาหุ้นวันที่ 6 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 28 พฤศจิกายน 2562 จนถึง วันที่ 5 กุมภาพันธ์ 2563 และเลือก ตัวแบบในการทำซ้ำครั้งที่ 1



รูปที่ 25 ค่าความแม่นยำของหุ่นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) เพื่อพยากรณ์ทิศทางของหุ้นราคา วันที่ 7 กุมภาพันธ์ 2563 โดยใช้ข้อมูลข่าวตั้งแต่วันที่ 28 พฤศจิกายน 2562 จนถึง วันที่ 6 กุมภาพันธ์ 2563 และเลือก ตัวแบบในการทำซ้ำครั้งที่ 2

ตารางที่ 5 การพยากรณ์ทิศทางของราคาหุ้นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK)

ค่าความแม่นยำ	ตัวแบบการจำแนก	การตัดคำ	หุ้น	วันที่พยากรณ์	ทำซ้ำครั้งที่	ค่าพยากรณ์	ค่าจริง
0.888889	KNeighborsClassifier	multi_cut	KBANK	5/2/2020	1	positive	positive
0.888889	KNeighborsClassifier	multi_cut	KBANK	6/2/2020	1	positive	positive
0.888889	LogisticRegression	trie	KBANK	7/2/2020	2	negative	positive

การพยากรณ์ทิศทางของราคาหุ้นธนาคารกสิกรไทย จำกัด (มหาชน) (KBANK) นั้น เมื่อเปรียบเทียบกับข้อมูลจริงของวันที่ 5, 6, 7 กุมภาพันธ์ 2563 พบว่ามีความถูกต้องร้อยละ 66.67 ตามที่ได้แสดงไว้ในตารางที่ 5

วิจารณ์ผลการวิจัย

ในงานวิจัยชิ้นนี้ ผู้วิจัยใช้ Library pythainlp ช่วยในการประมวลผล และใช้การตัดคำ 9 วิธี คือ 'cfcut', 'deepcut', 'etcc', 'longest', 'multi_cut', 'newmm', 'ssg', 'tcc' และ 'trie' จากนั้นแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลฝึกสอน (training dataset) 80% ของข้อมูลทั้งหมด และข้อมูลทดสอบ (test dataset) 20% ของข้อมูลทั้งหมด ใช้ข้อมูลฝึกสอนในการสร้างตัวแบบ โดยแบ่งข้อมูลและสร้างตัวแบบซ้ำทั้งหมด 5 ครั้ง ในแต่ละวัน จากตารางที่ 1-5 จะเห็นว่าตัวแบบการจำแนกและวิธีการตัดคำที่เหมาะสมในแต่ละประเภทหุ้นไม่ได้มีความสัมพันธ์กัน เนื่องจากวิธีการตัดคำเป็นอัลกอริทึม (algorithm) ที่ใช้ในการตัดคำในภาษาไทยเพื่อให้ได้ความหมายที่ถูกต้อง ซึ่งอยู่ในขั้นตอนของการเตรียมข้อมูลก่อนนำไปสร้างตัวแบบ โดยคำในข้อความข่าวจะแตกต่างกันไปในแต่ละกลุ่มของหุ้น อีกทั้งการแบ่งข้อมูลออกเป็นข้อมูลฝึกสอนและข้อมูลทดสอบเป็นการแบ่งโดยใช้การสุ่มตัวอย่างอย่างง่ายแบบไม่ใส่คืน ดังนั้นการตัดคำที่เหมาะสมก็จะแตกต่างกันไปในแต่ละครั้งที่ทำซ้ำ นอกจากนี้หากเกิดกรณีที่ตัวแบบมีค่าความแม่นยำเท่ากันผู้วิจัยจะเลือกตัวแบบที่ดีที่สุดโดยสุ่ม ด้วยเหตุผลทั้งหมดที่กล่าวมาข้างต้นนี้ส่งผลให้แบบจำลองการจำแนกที่เหมาะสมกับการพยากรณ์ทิศทางของราคาหุ้นบางครั้งจะมีความแตกต่างกันในแต่ละครั้งที่ประมวลผล

สรุปผลการวิจัย

ในงานวิจัยนี้ได้ทำการสุ่มเลือกหุ้นอย่างละ 1 ตัว จากหุ้น 5 ประเภท ทำการรวบรวมข้อมูลข้อความข่าวของหุ้นแต่ละตัวในแต่ละวันและสถานะของราคาหุ้น ณ ปัจจุบันเทียบกับราคาหุ้นของวันก่อนหน้า โดยใช้วิธีการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ซึ่งทำการตัดคำโดยใช้วิธีการตัดคำ 9 วิธี ใน Library pythainlp คือ 'cfcut', 'deepcut', 'etcc', 'longest', 'multi_cut', 'newmm', 'ssg', 'tcc' และ 'trie' แล้วให้น้ำหนักคำโดยใช้เทคนิค TFIDF (term frequency-inverse document frequency) ทำการเรียนรู้ข้อมูลฝึกสอน (training dataset) 80% ของข้อมูลทั้งหมด และข้อมูลทดสอบ (test dataset) 20% ของข้อมูลทั้งหมด ใช้แบบจำลองการจำแนก (classification model) ทั้งหมด 7 แบบจำลอง คือ K-neighbors Classifier, Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, AdaBoost Classifier, XGB Classifier และ Support Vector Machine (SVM) เพื่อหาตัวแบบที่มีค่าความแม่นยำ (accuracy) สูงสุด เพื่อทำการพยากรณ์ทิศทางของราคาหุ้นจากข้อความข่าว ในวันที่ 5, 6 และ 7 กุมภาพันธ์ 2563 ของหุ้นทั้งหมด 5 กลุ่ม โดยแสดงในส่วนของร้อยละความถูกต้องเพื่อเป็นการเปรียบเทียบทิศทางของราคาหุ้นของค่าพยากรณ์และค่าจริง ดังตารางที่ 6

จากตารางที่ 6 จะเห็นว่าตัวแบบการจำแนกที่เหมาะสมนั้นแตกต่างกันในแต่ละกลุ่มของหุ้นและบางครั้งไม่สามารถสรุปได้ว่าตัวแบบใดเป็นตัวแบบที่ดีที่สุด ดังนั้นในการนำไปใช้งานจริงผู้วิจัยจึงเขียนโปรแกรมภาษาไพธอนให้สามารถประมวลผลโดยเลือกตัวแบบที่ดีที่สุดจากทั้งหมด 7 ตัวแบบและเลือกการตัดคำที่ดีที่สุดจากทั้งหมด 9 วิธี เพื่อใช้ในการพยากรณ์ทิศทางของราคาหุ้นจากข้อความข่าวในแต่ละวัน

ตารางที่ 6 สรุปผลการวิจัยของหุ้นแต่ละกลุ่ม

ประเภทของหุ้น	ชื่อหุ้น	ชื่อบริษัท	แบบจำลองที่มีประสิทธิภาพสูงสุด		ร้อยละความถูกต้อง
			ตัวแบบการจำแนก	การตัดคำ	
เทคโนโลยีสารสนเทศและการสื่อสาร (ICT)	INTUCH	บริษัท อินทัช โฮลดิ้งส์ จำกัด	Gradient Boosting Classifier	ไม่สามารถสรุปได้	100
พลังงานและสาธารณูปโภค(ENERG)	TOP	บริษัท ไทยออยล์ จำกัด	ไม่สามารถสรุปได้	ไม่สามารถสรุปได้	66.67
การแพทย์ (HEALTH)	BH	บริษัท โรงพยาบาลบำรุงราษฎร์ จำกัด	ไม่สามารถสรุปได้	cfcut	66.67
พาณิชย์ (COMM)	CPALL	บริษัท ซีพี ออลล์ จำกัด	K-Neighbors Classifier	ไม่สามารถสรุปได้	66.67
ธนาคาร (BANK)	KBANK	ธนาคารกสิกรไทย จำกัด	K-Neighbors Classifier	multi_cut	66.67

ข้อเสนอแนะ

1. เนื่องจากภาษาไทยเป็นภาษาที่มีความซับซ้อนดังนั้นจึงควรมีนักภาษาศาสตร์เข้ามาช่วยพิจารณาในการแยกคำ การจัดกลุ่มคำและการให้ความหมาย
2. ควรมีการปรับปรุงและพัฒนาแบบจำลองเพื่อให้แบบจำลองสามารถทำงานแบบ real-time monitoring เพื่อเป็นเครื่องมือช่วยประกอบการตัดสินใจลงทุนที่มีประสิทธิภาพ
3. ควรมีการหาค่าพารามิเตอร์ที่เหมาะสมในแต่ละตัวแบบ ซึ่งในงานวิจัยชิ้นนี้ใช้ค่า Default ของโปรแกรม ภาษาไพธอนทุกตัวแบบ
4. เนื่องจากเวลาในการวิจัยมีจำกัดจึงไม่ได้มีการคัดคำที่ไม่มีผลเกี่ยวข้องกับการพยากรณ์ออก เช่น คำว่า “การ” “บาท” “ที่” และ “ความ” เป็นต้น ดังนั้นในอนาคตควรศึกษาเพิ่มเติมในเรื่องนี้

เอกสารอ้างอิง

- Chaicharoen, N. (2001). *Computerized integrated word segmentation and part-of-speech tagging of Thai* (M.Sc. thesis). Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok. (in Thai)
- Huang, W., Nakamori, Y. & Wang, S-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10), 2513–2522.
- Kim, K-J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1), 307–319.
- Li, X., Huang, X., Deng, X., & Zhu, S. (2014). Enhancing Quantitative Intra-day Stock Return Prediction by Integrating both Market News and Stock Prices Information. *ScienceDirect*, 142, 228-238.
<https://doi.org/10.1016/j.neucom.2014.04.043>
- Sadia, K., Sharma, A., Paul, A., Padhi, S., & Sanyal, S. (2019). Stock Market Prediction Using Machine Learning Algorithms. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(4), 25-31.
- Sonlertlamvanit, W. (1993). *Thai word wrapping in the translation system*. Bangkok, Thailand: National Electronics and Computer Technology Center. (in Thai)
- Tay, F.E.H. & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, 29(4), 309–317.