การศึกษาการจำแนกข่าวหุ้นภาษาไทย ด้วยการใช้คู่คำเป็นคุณลักษณะ

THAI STOCK NEWS SENTIMENT CLASSIFICATION USING WORDPAIR FEATURES

อภินันท์ จัตตุพันธ์

APINAN CHATTUPAN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-IT-M-001-013

การศึกษาการจำแนกข่าวหุ้นภาษาไทย ด้วยการใช้คู่คำเป็นคุณลักษณะ

THAI STOCK NEWS SENTIMENT CLASSIFICATION USING WORDPAIR FEATURES

อภินันท์ จัตตุพันธ์

APINAN CHATTUPAN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2558

KMITL-2015-IT-M-001-013

THAI STOCK NEWS SENTIMENT CLASSIFICATION USING WORDPAIR FEATURES

APINAN CHATTUPAN

A THESIS SUBMITTED IN FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2015

KMITL-2015-IT-M-001-013

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การศึกษาการจำแนกข่าวหุ้นภาษาไทย ด้วยการใช้คู่คำเป็นคุณลักษณะ

Thai stock news sentiment classification using wordpair features

นักศึกษา นายอภินันท์ จัตตุพันธ์

รหัสประจำตัว 57606151

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.พรฤดี เนติโสภากุล

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
รองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช	
ดร.เทพชัย ทรัพย์นิธิ	
รองศาสตราจารย์ ดร.พรฤดี เนติโสภากุล	
ผู้ช่วยศาสตราจารย์ ดร.กิตติ์สุชาต พสุภา	
รองศาสตราจารย์ ดร.อาริต ธรรมโน	

วัน / เดือน/ ปี ที่สอบ วันอังคารที่ 15 ธันวาคม 2558 เวลา 14.30 น. เป็นต้นไป

สถานที่สอบ ณ ห้อง 333 ชั้น 3 คณะเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

(รองศาสตราจารย์ ดร.จันทร์บูรณ์ สถิตวิริยวงศ์)

คณบดีคณะเทคโนโลยีสารสนเทศ

วันที่ 21 **เดือน** ธันวาคม **พ.ศ.** 2558

หัวข้อวิทยานิพนธ์	การศึกษาการจำแนกข่าวหุ้นภาษาไทย ด้วยการใช้คู่คำเป็นคุณลักษณะ
นักศึกษา	นายอภินันท์ จัตตุพันธ์
รหัสประจำตัว	57606151
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
พ.ศ.	2558
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รองศาสตราจารย์ ดร. พรฤดี เนติโสภากุล

บทคัดย่อ

การวิเคราะห์ข่าวหุ้นภาษาไทย (Stock news analysis) เป็นการวิเคราะห์ที่ได้รับความนิยมนอกเหนือจากการวิเคราะห์ดัชนีหรือข้อมูลสถิติเป็นหลัก สืบเนื่องจากข่าวหุ้นที่ประกาศจากโบรคเกอร์ มีลักษณะรูปแบบการเขียนที่แตกต่างกันออกไป อีกทั้งประเภทของข่าวมีความหลากหลาย ตัวอย่างเช่น รายงานพื้นฐาน หุ้นมีข่าว และ ปัจจัยที่มีผลต่อตลาด เป็นต้น

วิทยานิพนธ์นี้ได้นำเสนอการวิเคราะห์ความสัมพันธ์ระหว่างข่าวหุ้น และข้อมูลทางสถิติเกี่ยวกับตลาดหุ้น โดยการสร้างกราฟจำลองแสดงความสัมพันธ์เมื่อมีการประกาศของข่าวออกมา รูปแบบกราฟจะมีการปรับขึ้น หรือลดลงจากค่าปกติ และมีการทดสอบสมมติฐานเพื่อยืนยันความสอดคล้องของข้อมูลเหล่านี้

นอกจากนี้เราได้เก็บรวบรวมคุณลักษณะใหม่ ซึ่งใช้ในการจำแนกข่าวหุ้นซึ่งมีลักษณะเป็นข้อความ โดยคุณลักษณะใหม่มีชื่อเรียกว่า 'คู่คำ' สำหรับการจำแนกข่าวหุ้นออกเป็น ข่าวบวก ข่าวลบ และข่าวที่เป็นกลาง สำหรับการจำแนกข่าวหุ้นที่เกิดขึ้นภายในอนาคต

ผลของการทดลองแสดงให้เห็นถึงประสิทธิภาพของการใช้คู่คำเป็นคุณลักษณะในการจำแนกข่าวหุ้น ซึ่งเมื่อเปรียบเทียบประสิทธิภาพกับการใช้คำที่อยู่ภายในประโยคเป็นคุณลักษณะ จะให้ประสิทธิภาพที่ใกล้เคียงกันแต่การใช้คู่คำเพื่อเป็นตัวแทนของประโยค จะมีปริมาณการใช้คุณลักษณะหรือขนาดที่น้อยกว่าการใช้คำที่อยู่ภายในประโยคเป็นคุณลักษณะเป็นอย่างมาก

Thesis Title	Thai stock news sentiment classification using wordpair features
Student	Mr. Apinan Chattupan
Student ID.	57606151
Degree	Master of Science
Program	Information Technology
Major	Information Science
Academic Year	2015
Thesis Advisor	Associate Professor Dr. Ponrudee Netisopakul

ABSTRACT

Thai stock news is daily issued from many stock brokers. Thai stock news is an important source of information for stock traders to make a decision on stock trading. However, a usual Thai stock news such as basic report, stock's news or market research report has a long message and sometimes not easily to interpret or conclude.

This research investigates the relationship of stock news provided by brokers and stock trading statistics, by creating graph to visualize the time-based relationship of stock news issue date and price surge/drop or volume surge/drop on the same timeframe. Then we do hypothesis testing to confirm the significant of the findings.

Moreover, we assume that 'features' that can be used for classifying the news must be presented as text in the news. We propose to construct a set of these 'texts' to be used as features that are called 'wordpairs' in order to classify Thai stock news into three sentiments: positive (+), negative (-) and neutral (0) classes, using known sentiment news as a training set and unseen news as a testing set.

The experiments of wordpair and individual word model give similarly results, but the wordpair model are used features less than the individual word model.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยความช่วยเหลือจากรองศาสตราจารย์ ดร. พรฤดี เนติโสภากุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ทุ่มเทและเสียสละเวลาอันมีค่าในการให้คำปรึกษาหัวข้อวิจัยตั้งแต่การวางแนวทางการวิจัย ให้คำแนะนำในการทำงานในแต่ละขั้นตอน การตรวจสอบความถูกต้องและให้แนวทางในการแก้ไขปัญหา รวมไปถึงการร่วมเขียนต้นฉบับบทความวิจัยและการแก้ไขบทความวิจัยในฉบับสุดท้าย ขอบคุณ นายรัฐวุฒิ เลิศสุขศักดา และเพื่อนร่วมงานภายในห้องวิจัยการจัดการความรู้และวิศวกรรมความรู้ คณะเทคโนโลยีสารสนเทศสำหรับความช่วยเหลือและข้อแนะนำตลอดจนการรับฟังการนำเสนอผลงานภายในห้องวิจัย ขอขอบคุณคณะผู้ตรวจและผู้จัดงานประชุมวิชาการที่ให้ข้อคิดเห็น และการปรับแก้ไขบทความ รวมทั้งการช่วยเหลือและประสานงานในการนำเสนอผลงานทางวิชาการ

ขอกราบขอบพระคุณคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังที่ได้สนับสนุนทุนการศึกษาและค่าใช้จ่ายในการศึกษาเล่าเรียนในระดับบัณฑิตศึกษา นอกจากนี้ยังได้สนับสนุนงบประมาณในการตีพิมพ์บทความทางวิชาการในวารสารการประชุมวิชาการ ให้สำเร็จลุล่วงเป็นอย่างดี

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา ที่ให้การสนับสนุนด้านการศึกษาจนสามารถสำเร็จลุล่วงเป็นอย่างดี ขอกราบขอบพระคุณคณะครูและอาจารย์ทุกท่านที่ได้มอบความรู้ แนวคิดในการดำเนินชีวิตและการใช้ชีวิตอยู่ร่วมกันในสังคม ตั้งแต่การเรียนในระดับปริญญาตรี ตลอดจนสำเร็จการศึกษาในระดับปริญญาโท

อภินันท์ จัตตุพันธ์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของวิทยานิพนธ์

การวิเคราะห์ข้อความ (Text analysis) ด้วยการจำแนกข้อมูล (Classification) เป็นการวิเคราะห์ประเภทหนึ่งที่ได้รับความนิยมเพื่อให้เครื่อง (Machine) หรือคอมพิวเตอร์สามารถเรียนรู้ เข้าใจ และสามารถวิเคราะห์ข้อความที่เกิดขึ้นใหม่ได้ในอนาคต สำหรับข่าวหุ้นก็จัดอยู่ในข้อความที่สามารถนำมาวิเคราะห์ได้นอกเหนือจากข้อมูลทางสถิติ เนื่องจากข่าวหุ้นมีการประกาศจากโบรคเกอร์ หนังสือพิมพ์ หรือเว็บไซต์ที่เกี่ยวข้องกันการลงทุน ในทุกวันทำการที่มีการลงทุน ข่าวหุ้นหรือผลลัพธ์ที่ได้จากการวิเคราะห์จะก่อให้เกิดประโยชน์แก่นักลงทุนในการระบุทิศทางของตลาด อารมณ์ หรือแม้กระทั่งทราบความเคลื่อนไหวแบบเฉพาะเจาะจงของหุ้นตัวที่นักลงทุนให้ความสนใจได้

ปัญหาที่สำคัญสำหรับการวิเคราะห์ข่าวหุ้นคือ การอ่าน การทำความเข้าใจเนื้อหาภายในข่าวที่มีการประกาศออกมา นักลงทุนแต่ละคนสามารถวิเคราะห์อารมณ์และทิศทางของตลาดหุ้นได้ถูกต้องมากน้อยแค่ไหน สาเหตุเนื่องมาจาก มุมมองและการตีความหมายของแต่ละบุคคลมีความแตกต่างกันไปตามประสบการณ์ ดังนั้นการใช้เครื่องหรือคอมพิวเตอร์ช่วยในการแยกแยะหรือการจำแนกข้อมูล จึงเข้ามีบทบาท และเป็นเหตุผลสำคัญในการวิเคราะห์ข่าวหุ้น สำหรับคำถามที่ตามมาคือ เราจะรู้ได้อย่างไรว่าข่าวหุ้นที่ประกาศจากโบรคเกอร์มีผลต่อหุ้น และอีกหนึ่งคำถามต่อมาคือ เราสามารถที่จะใช้ข้อมูลอะไรในการวิเคราะห์ข่าวหุ้นได้บ้าง จึงทำให้เกิดสมมติฐานในงานวิจัยนี้ขึ้นมาเพื่อนำไปสู่กระบวนการการทดลองและหาข้อสรุปผล

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อวิจัยและพัฒนาตัวแทนของข้อความรูปแบบใหม่สำหรับการวิเคราะห์ข่าวหุ้นภาษาไทย
2. เพื่อศึกษาความถูกต้องและความสอดคล้อง ระหว่างข่าวหุ้นภาษาไทยและข้อมูลทางสถิติของตลาดหุ้นไทย ประกอบด้วยการเปลี่ยนราคาหุ้นและปริมาณการซื้อขายหุ้น
3. เพื่อวัดประสิทธิภาพความน่าเชื่อถือของตัวแทนข้อความรูปแบบใหม่ เปรียบเทียบกับการวิเคราะห์ด้วยตัวแทนข้อความรูปแบบเดิม

1.3 แนวคิดที่ใช้ในการวิจัย

แนวคิดที่ใช้ในการวิจัยเกิดจากงานวิจัยก่อนหน้า (Chattupan and Netisopakul, 2014) กล่าวถึงการวิเคราะห์ข่าวหุ้นในภาษาไทยด้วยการใช้คำ (Individual word) ที่อยู่ภายในข่าวหุ้นเป็นตัวแทนหรือคุณลักษณะในการจำแนกข้อมูล (Classification) ออกเป็นข่าวที่มีอารมณ์เป็นบวก (+ / 1) อารมณ์เป็นลบ (- / -1) และอารมณ์เป็นกลาง (0) ต่อตลาดหุ้น ซึ่งภายในงานวิจัยก่อนหน้านี้มีการตรวจสอบความสอดคล้องหรือความสัมพันธ์ระหว่างข่าวหุ้น และข้อมูลทางสถิติของตลาดหุ้นไทยด้วยการใช้ค่าสหสัมพันธ์ (Correlation coefficient) เป็นตัวประเมิน และพบว่าหุ้นบางกลุ่มธุรกิจ เช่น หมวดธุรกิจปิโตรเคมี (Petrochemical) หรือหมวดธุรกิจพลังงาน (Energy) มีค่าสหสัมพันธ์เป็นไปในทิศทางเดียวกันระหว่างการเปลี่ยนแปลงราคาหุ้น และอารมณ์ของข่าวหุ้น แต่ภายในงานวิจัยพบข้อเสียที่สำคัญหลังจากขั้นตอนการหาค่าสหสัมพันธ์คือ ผลของการจำแนกข้อมูลด้วยคำให้ประสิทธิภาพที่ไม่ค่อยสูงนัก ส่งผลต่อความผิดพลาดหากนำไปใช้เป็นข้อมูลชุดการสอนให้เครื่องหรือคอมพิวเตอร์ ดังนั้นการทดลองเพื่อหาตัวแทนของข้อมูลรูปแบบอื่นสำหรับใช้ในการวิเคราะห์ข่าวหุ้นภาษาไทยจึงเป็นแนวความคิดที่เกิดขึ้น เพื่อเปรียบเทียบประสิทธิภาพและความถูกต้องระหว่างการใช้คำเป็นคุณลักษณะในงานวิจัยเดิมและคุณลักษณะใหม่ที่ใช้เป็นตัวแทนข้อมูลประกอบกับวิธีการประเมินความสอดคล้องในระดับที่มีความซับซ้อนมากขึ้นระหว่างข่าวหุ้นภาษาไทยและข้อมูลทางสถิติของตลาดหุ้นไทยเพื่อให้การวิเคราะห์มีความน่าเชื่อถือมากขึ้น

1.4 คำถามงานวิจัย (Research questions)

1. ภายในข่าวหุ้นภาษาไทย การกล่าวถึงหุ้น (Stock symbol) ตัวใดตัวหนึ่งส่งผลต่อการเปลี่ยนแปลงของราคาและปริมาณการซื้อขายของหุ้นตัวใดตัวหนึ่งที่ถูกกล่าวถึงหรือไม่
2. ตัวแทนของข่าวหุ้นรูปแบบใหม่ที่ถูกใช้เป็นคุณลักษณะ เมื่อเปรียบเทียบประสิทธิภาพกับการใช้คำ (Individual word) เป็นตัวแทนของข่าวหุ้นจะมีประสิทธิภาพเป็นอย่างไร
3. รูปแบบ (Format), วิธีการเขียนข่าว (Writing style) หรือโครงสร้างของข่าว (Structure) จะส่งผลต่อประสิทธิภาพการจำแนกข้อมูลหรือไม่

1.5 ขอบเขตของการวิจัย

1. กำหนดการใช้ข่าวหุ้นภาษาไทยเป็นข้อมูลชุดการสอน (Training set) จากโบรคเกอร์ บล. บัวหลวง ข้อมูลระหว่างวันที่ 4 เมษายน 2014 ถึงวันที่ 27 พฤษภาคม 2015
2. กำหนดการใช้ข่าวหุ้นภาษาไทยเป็นข้อมูลชุดทดสอบ (Testing set) จากโบรคเกอร์อื่น ได้แก่ บล. ธนชาต, บล. กรุงศรีและโบรคเกอร์อื่น ข้อมูลระหว่างวันที่ 10 มีนาคม 2015 ถึงวันที่ 3 กรกฎาคม 2015
3. การเปรียบเทียบระหว่างข่าวหุ้นและข้อมูลทางสถิติ ประกอบด้วยการเปลี่ยนแปลงราคา (Price changes) และปริมาณการซื้อขายหุ้น (Trading volume) จะใช้ข้อมูลสถิติจาก ดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม (Industry) และดัชนีหมวดธุรกิจ (Sector)

1.6 ขั้นตอนการวิจัย

1. ศึกษารูปแบบของข่าวหุ้นจากงานวิจัยก่อนหน้า เพื่อปรับเปลี่ยนรูปแบบตัวแทนของข้อมูลที่จะถูกนำมาใช้เป็นคุณลักษณะในการจำแนกข้อมูล
2. ตรวจสอบความสอดคล้องระหว่างข้อมูลข่าวหุ้นและข้อมูลทางสถิติของตลาดหุ้นไทย
3. ทดลองสกัดรูปแบบ (Feature extracting) และเก็บรวบรวมคุณลักษณะ (Feature collecting) ให้ได้ปริมาณในระดับหนึ่งสำหรับทำการทดลองเบื้องต้น
4. สร้างโมเดลการเรียนรู้จากชุดการสอน (Training model) โดยการทดลองใช้คุณลักษณะรูปแบบใหม่
5. เปรียบเทียบผลการทดลองระหว่างโมเดลการเรียนรู้ที่ใช้คุณลักษณะรูปแบบใหม่ และโมเดลการเรียนรู้ที่ใช้คำเป็นคุณลักษณะ
6. ทำซ้ำระหว่างขั้นตอนที่ 3 - 5 เพื่อวัดผลกับสภาพแวดล้อมหรือการปรับตั้งค่าในรูปแบบอื่น

1.7 นิยามศัพท์

1. คำ (Individual word) หมายถึงพยางค์หรือตัวของคำหนึ่งคำที่ออกเสียงเพียงครั้งเดียว อาจมีความหมายที่สมบูรณ์หรือมีความหมายที่ไม่สมบูรณ์ก็ได้
2. คู่คำ (Wordpair) หมายถึงคำสองคำที่มีความหมายสมบูรณ์หรือมีความหมายเพิ่มมากขึ้นเมื่อใช้ประกอบกัน ทำหน้าที่ในการขยายความเสริมกันและกัน อยู่ในรูปแบบของคีย์เวิร์ดตามด้วยคำที่มีขั้ว

1.8 การมีส่วนร่วมต่องานวิจัยในสาขาวิชา (Contribution)

1. งานวิจัยนี้ได้คิดค้นการเก็บรวบรวมคู่คำ (Wordpair) เพื่อใช้เป็นตัวแทนของข่าวหรือคุณลักษณะในการจำแนกอารมณ์ (Sentiment classification) ของข่าวหุ้น และการนำคู่คำไปประยุกต์ใช้จริงภายในข่าวหุ้นภาษาไทย
2. ทำการทดลองเพื่อตรวจสอบประสิทธิภาพของคู่คำ และเปรียบเทียบประสิทธิภาพกับการวิเคราะห์ในรูปแบบอื่น

1.9 ผลงานที่ได้รับการตีพิมพ์

ผลงานตีพิมพ์ตลอดการทำวิยานิพนธ์นี้ มีทั้งหมด 2 บทความ (สามารถดูได้ที่ภาคผนวก ข)

- Apinan Chattupan, and Ponrudee Netisopakul. "Stock sentiment analysis model using data mining (In Thai)." *In Knowledge and Smart Technology (KST), 2014. Proceeding of 6th National Conference on.* Chonburi, Thailand.
- Apinan Chattupan, and Ponrudee Netisopakul. "Thai Stock News Sentiment Classification using Wordpair Features." *In Proceeding of 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29.* pp. 188-195. Shanghai, China. Oct. 30 – Nov. 1, 2015.

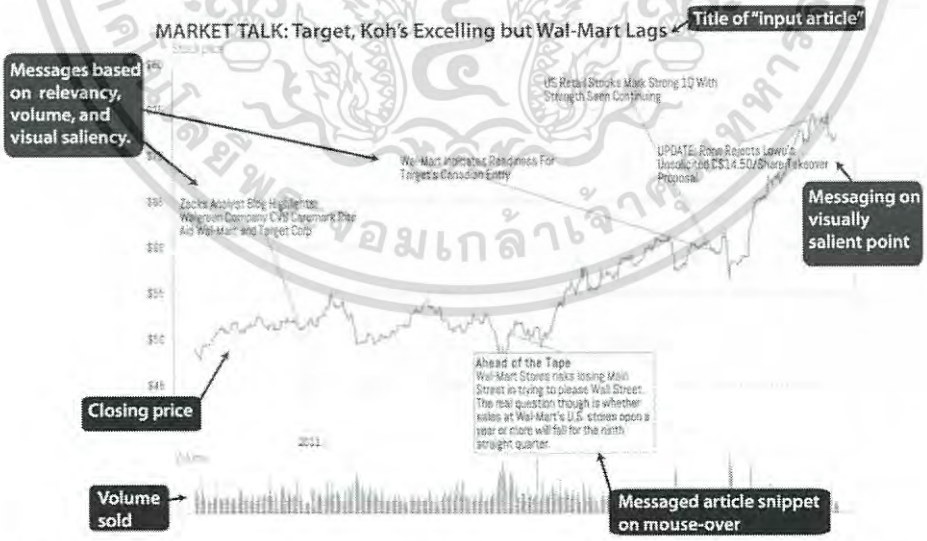
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้กล่าวถึงทฤษฎีหรืองานวิจัยที่มีความใกล้เคียงหรือเกี่ยวข้องกับการทดลองในเรื่องของการวิเคราะห์ข่าวหุ้นหรือข้อความในภาษาไทย โดยแยกประเภทของเทคโนโลยีที่ใช้ในงานวิจัยออกเป็น 3 หัวข้อหลักประกอบด้วย เทคโนโลยีการสร้างกราฟจำลอง (Graph visualization) การทดสอบสมมติฐานด้วยสถิติ (Hypothesis testing) และงานวิจัยการทำเหมืองข้อมูลด้านการประมวลผลทางภาษา (Text mining) กล่าวถึงโครงสร้างของภาษาและงานวิจัยในด้านของการทำเหมืองข้อความ การจำแนกข้อความที่มีความเกี่ยวข้องหรือความใกล้เคียงกันเรื่องของหุ้น

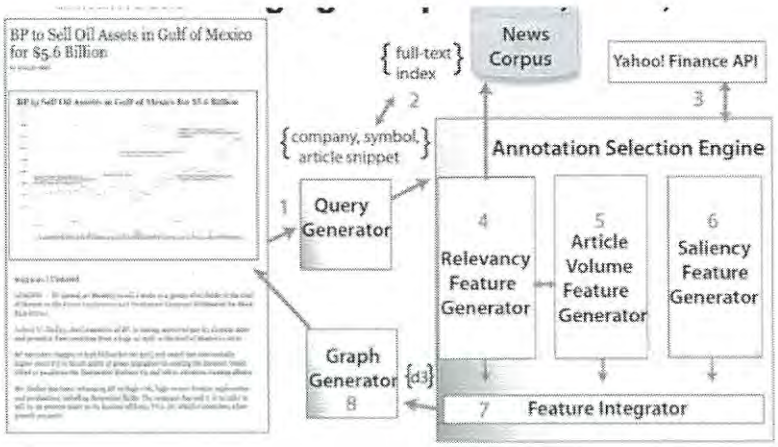
2.1 เทคโนโลยีการสร้างกราฟจำลอง (Graph visualization)

การจำลองข้อมูล (Data visualization) คือ การแสดงผลรูปแบบหนึ่งซึ่งใช้แทนคำพูด โดยการใช้ภาพเพื่อแสดงข้อมูลในเชิงปริมาณที่วัดได้ให้มีความน่าสนใจ และสามารถเห็นภาพได้อย่างชัดเจน เช่น ตัวเลข แผนภูมิ กราฟ เป็นต้น (กิรดา กลีบมาลัย, 2015)

Hullman, et al. (2013) ได้แสดงการใช้กราฟจำลองในการวิเคราะห์ตลาดหุ้น โดยข้อมูลที่ใช้ในการวิเคราะห์จะเป็นบริบทของข่าวหุ้น แสดงตัวอย่างการใช้กราฟจำลองในภาพที่ 2.1



ภาพที่ 2.1 แสดงตัวอย่างการใช้กราฟจำลองในการวิเคราะห์ตลาดหุ้น (Hullman, et al., 2013)



ภาพที่ 2.2 การประมวลผลข่าวและข้อมูลทางการเงินเพื่อสร้างกราฟจำลอง (Hullman, et al., 2013)

จากภาพที่ 2.2 Hullman, et al. (2013) ยังแสดงให้เห็นถึงโครงสร้างการทำงานของระบบที่สร้างกราฟจำลอง โดยใช้ข้อมูลจากบริบทของข่าว และข้อมูลทางการเงิน มาเป็นคุณลักษณะหรือข้อมูลที่สำคัญในการสร้างกราฟ

2.2 การทดสอบสมมติฐานด้วยสถิติเครื่องหมาย-อันดับของวิลคอกซัน (Wilcoxon signed-rank test)

การทดสอบสมมติฐานด้วยสถิติเครื่องหมาย-อันดับของวิลคอกซัน เป็นสถิติที่ไม่ใช้พารามิเตอร์สำหรับการทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ยประชากร เหมาะสมกับกลุ่มตัวอย่างซึ่งไม่สามารถใช้การทดสอบด้วยค่า Z ได้ (อัจฉริยา ปราบอริพ่าย, 2008)

สำหรับข้อตกลงเบื้องต้นของการทดสอบด้วยวิลคอกซัน ประกอบด้วย

1) ตัวอย่างของข้อมูลจะต้องได้มาด้วยการสุ่ม
2) ข้อมูลที่ศึกษาเป็นแบบต่อเนื่อง
3) การแจกแจงของประชากรเป็นแบบสมมาตรตรงค่าเฉลี่ย μ
4) สเกลของข้อมูลเป็นแบบช่วง

สำหรับสมมติฐานของวิลคอกซันสามารถทดสอบได้ 3 รูปแบบ โดยเขียนรูปแบบของสมมติฐานได้ดังนี้

$H_0: x_i = \mu_0 \quad H_1: x_i \neq \mu_0 \quad$ หรือ

$H_0: x_i \geq \mu_0 \quad H_1: x_i < \mu_0 \quad$ หรือ

$H_0: x_i \leq \mu_0 \quad H_1: x_i > \mu_0$

จากรูปแบบของการทดสอบสมมติฐานของวิลคอกซัน สามารถเขียนสรุปกฎการตัดสินใจออกได้เป็น 3 รูปแบบ คือ

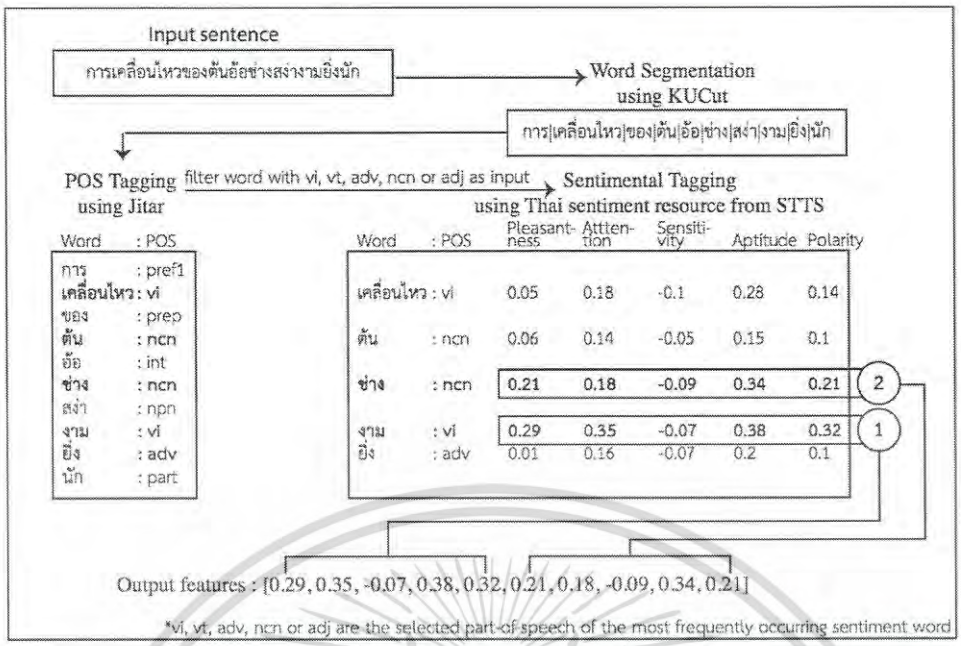
1) สมมติฐานแย้ง $H_1: x_i \neq \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤติ
2) สมมติฐานแย้ง $H_1: x_i < \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T^+ ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤติ
3) สมมติฐานแย้ง $H_1: x_i > \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤติ

2.3 การทำเหมืองข้อมูลด้านการประมวลผลทางภาษา (Text mining)

รูปแบบของการประมวลผลข้อความภาษาไทย Taboada, et al. (2011) ได้กล่าวภายในบทความ Lexicon-Based Methods for Sentiment Analysis ถึงการวิเคราะห์ความคิด (Opinion) ซึ่งเป็นข้อความรูปแบบหนึ่ง โดยใช้การวิเคราะห์ในระดับของคำ (Word), วลี (Phase), ประโยค (Sentence) และการจัดหมวดหมู่ของข้อความที่เกี่ยวข้องกันให้อยู่ในรูปแบบพจนานุกรม (Dictionary) โดยการวิเคราะห์จะมุ่งเน้นการวิเคราะห์ข้อความออกเป็นเชิงบวกหรือลบ (Sentiment analysis) ด้วยการใช้เครื่องมือวัดระดับความรู้สึกของคำเรียกว่า The Semantic Orientation Calculator (SO-CAL) สร้างขึ้นจากสมมติฐานแรกคือ คำแต่ละคำจะมีขั้วของตัวเอง และสมมติฐานที่สองคือ ความหมายของคำสามารถแปลงเป็นตัวเลขเพื่อใช้ในการคำนวณได้

2.3.1 การวิเคราะห์ข้อความด้วยการระบุชนิดของคำ (Part of speech)

ในการวิเคราะห์ความรู้สึกของข้อความส่วนใหญ่จะมุ่งเน้นไปที่ชนิดของคำ (Part of speech) เป็นหลัก ภายในงานวิจัย (Lertsuksakda, et al., 2014) และ (Lertsuksakda, et al., 2015) มีการใช้ชนิดของคำ ประกอบด้วย คำนาม (Noun), คำกริยา (Verb), คำคุณศัพท์ (Adjective) และคำวิเศษณ์ (Adverb) ประกอบการวิเคราะห์หลักของระบบเพื่อคำนวณระดับขั้วของคำ สำหรับตัวอย่างการใช้ชนิดของคำประกอบการวิเคราะห์แสดงในภาพที่ 2.3



ภาพที่ 2.3 แสดงตัวอย่างการใช้ชนิดของคำประกอบการวิเคราะห์ (Lertsuksakda, et al., 2015)

นอกจากนี้ (Taboada, et al., 2011) ได้ให้ความสำคัญกับการวิเคราะห์คำคุณศัพท์ โดยให้เหตุผลว่าคำคุณศัพท์เป็นคำพื้นฐานภายในข้อความที่แสดงขั้วของคำได้ชัดเจน ส่วนคำนาม, คำกริยา และคำวิเศษณ์ ก็สามารถที่จะมีขั้วได้ จากตัวอย่างของประโยค 'The young man strolled purposefully through his neighborhood.' ซึ่งมีขั้วหรืออารมณ์ของประโยคเป็นบวก ภายในประโยคจะเห็นได้ว่าคำกริยา 'strolled', คำคุณศัพท์ 'purposefully' และคำนาม 'neighborhood' ต่างก็มีขั้วภายในตัวเองแทบทั้งสิ้น คือขั้วบวก แต่มีข้อสังเกตคือ คำนาม 'young man' กลับเป็นคำที่ไม่แสดงขั้วออกมา นอกจากนี้ยังรวมไปถึงคำกริยาอีกด้วย จากตัวอย่างของประโยค 'The teacher inspired her students to pursue their dreams.' เปรียบเทียบกับประโยค 'This movie was inspired by true events' คำกริยา 'inspired' จากทั้งสองประโยคให้ขั้วที่แตกต่างกัน โดยประโยคแรกจะแสดงขั้วบวก และขั้วเป็นกลางสำหรับประโยคที่สอง สำหรับตัวอย่างคำอื่นที่มีขั้วจากการคำนวณค่าภายในงานวิจัย (Taboada, et al., 2011) แสดงในตารางที่ 2.1

ตารางที่ 2.1 แสดงตัวอย่างคำนาม, คำกริยา และคำวิเศษณ์ พร้อมกับระดับของขั้วจากการคำนวณ (Taboada, et al., 2011)

คำ	ระดับของขั้ว
Determination (Noun)	1
Relish (V)	4
Delay (Noun and Verb)	-1
Excruciatingly (Adjective)	-5
Purposefully (Adjective)	2

สำหรับงานวิจัยฉบับอื่น (Chattupan and Netisopakul, 2014), (Mittermayer, 2004) และ (Wu, et al., 2014) ได้กล่าวถึงการวิเคราะห์อารมณ์ในหัวข้อของหุ้น โดย Chattupan and Netisopakul (2014) ทำการวิเคราะห์ข่าวหุ้นโดยการใช้คำภายในข่าวเป็นคุณลักษณะในการจำแนกออกเป็นข่าวด้านบวก ข่าวด้านลบ และข่าวเป็นกลาง ส่วน Mittermayer (2004) มีการใช้กระเป๋าคำ (Bag of word), ชื่อเฉพาะ (Named Entities) และ วลีของคำนาม (Noun Phrase) ในการวิเคราะห์ข้อความ นอกจากนี้ Wang (2012) กล่าวถึงคำศัพท์ที่มีค่าอารมณ์ส่วนใหญ่ภายในเว็บไซต์ประเภทไมโครบล็อก (Microblog) ว่าส่วนใหญ่จะเกิดหลังคำคุณศัพท์ เป็นหลัก

2.3.2 การวิเคราะห์ข้อความด้วยการสกัดตัวแทนของข้อความ

Thongtep and Theeramunkong (2010), Sutheebanjard and Premchaiswadi (2010) และ Lertcheva and Aroonmanakun (2009) ได้กล่าวถึงวิธีการสกัดรูปแบบที่เป็นตัวแทนของข้อความเพื่อใช้ในการวิเคราะห์ โดยมุ่งเน้นที่การสกัดชื่อบุคคล (Person name), ชื่อองค์กร (Organizational), ชื่อผลิตภัณฑ์ (Product name) หรือชื่อเฉพาะอื่น ๆ แสดงตัวอย่างในภาพที่ 2.4 นอกจากนี้ภายในบทความ (Vichayakitti and Jaruskulchai, 2005) ได้ทำการวิเคราะห์ข้อความภาษาไทย โดยใช้การแท็กเพียงชั่วคราว (Temporal Tagged) บทความ (Yang, et al., 2010) ได้ทำการวิเคราะห์ความสัมพันธ์ระหว่างหัวข้อและตัวแทนของข้อความภายใต้หัวข้อเพื่อตรวจสอบความสัมพันธ์ที่จะนำมาเป็นขั้ว และบทความ (Fan, et al., 2008) ที่ใช้การหาความสัมพันธ์ของข้อความทั้งหมดเปรียบเทียบกับบางส่วนของข้อความ หรือแม้กระทั่งการวิเคราะห์ด้วยอีโมติคอน (Emoticon) ภายในบทความ (Boia, et al., 2013) โดยสาระสำคัญของการวิเคราะห์ด้วยตัวแทนของประโยค มีความสำคัญเพื่อหาตัวแทนของข้อความที่มีความยาวมาก และส่งผลต่อความซับซ้อนของการวิเคราะห์ข้อความ ให้สามารถวิเคราะห์รูปแบบได้ง่ายขึ้น ลดเวลาในการประมวลผลข้อความลงและเป็นการเพิ่มประสิทธิภาพความถูกต้องในการวิเคราะห์

คนร้ายคือ นายไก่ อยู่บ้านเลขที่

(Space)(Person name:นายไก่)(Space)

ก่อนหน้านั้น นายไก่ได้แจ้งว่า

(Space)(Person name:นายไก่)(Text:ได้แจ้งว่า)

ภาพที่ 2.4 แสดงตัวอย่างการสกัดชื่อบุคคลและรูปแบบของบริบทรอบข้าง (Thongtep and Theeramunkong, 2010)

2.3.3 การสร้างพจนานุกรมเพื่อใช้เก็บรวบรวมคำในการวิเคราะห์

การสร้างพจนานุกรม (Dictionary) สำหรับการวิเคราะห์ข้อความเกิดขึ้นเพื่อเก็บรวบรวมตัวแทนของข้อความที่สกัดออกจากข้อความด้วยวิธีการต่าง ๆ บทความ (Mizumoto, et al., 2012) กล่าวถึงการสร้างพจนานุกรมกึ่งดูแล (Semi-supervise) เพื่อเก็บรวบรวมคำศัพท์ที่มีขั้วคำ หลังจากที่เก็บรวบรวมคู่คำไปบางส่วน จะมีการจัดการอัตโนมัติสำหรับขั้วคำใหม่ที่ไม่รู้จักขั้ว ตัวอย่างของขั้วคำที่เก็บรวบรวมภายในพจนานุกรม และขั้วคำที่ถูกเพิ่มโดยอัตโนมัติแสดงในภาพที่ 2.5

SEEDS IN THE EXPERIMENT.

positive words	negative words
wide	small
high	weakness
steady	stable
individual	low
favourable	down
relief	decrease
grow	negative
better	reduct
continue	degeneraton
achievement	

THE WORDS INCLUDED TO THE DICTIONARY.

positive words	negative words
wide	small
expansion	ambiguous
investment	collapse
screening	stalemate
forward	hard
great	abolition
high	financing

ภาพที่ 2.5 ขั้วคำที่เก็บรวบรวมภายในพจนานุกรม และขั้วคำที่ถูกเพิ่มโดยอัตโนมัติ (Mizumoto, et al., 2012)

แต่การสร้างพจนานุกรมมีข้อเสียบางประการ โดย Wu, et al., (2011) กล่าวถึงการแปลพจนานุกรมจากภาษาอังกฤษเป็นภาษาจีน ซึ่งพบปัญหาความแตกต่างทางวัฒนธรรมในแต่ละภาษา ส่งผลให้การแปลมีข้อผิดพลาดเกิดขึ้น

การทำเหมืองข้อความและการจำแนกข้อมูล Schumaker and Chen (2009) และ Fung, et al. (2002) ได้กล่าวในบทความถึงการทำนายแนวโน้มหุ้นโดยการทำเหมืองข้อมูลหรือเหมืองข้อความโดยการคาดการณ์ราคาหุ้นหลังจากที่มีการประกาศข่าวออกมา โดยใช้โมเดลซัพพอร์ตเวกเตอร์ (Support Vector Machine) นอกจากนี้ยังมีการศึกษาประสิทธิภาพของตัวแทนข้อความ เปรียบเทียบตัวแทนทั้ง 3 รูปแบบประกอบด้วย กระเป๋าคำ (Bag of word), วลีของคำนาม (Noun phrase) และ ชื่อเฉพาะ (Named Entities) เหมือนกับ (Mittermayer,2004) และผลลัพธ์ที่ได้คือการใช้ชื่อเฉพาะ (Named Entities) มีความเหมาะสมมากกว่ารูปแบบอื่น

2.3.4 โมเดลการจำแนกข้อมูล

ต้นไม้ตัดสินใจ (Decision Tree) คือโครงสร้างของข้อมูลที่มีลักษณะเป็นลำดับชั้น (Hierarchy) ประกอบด้วย ราก (Root), กิ่ง (Branch) และใบ (Leaf) โดยหลักการของต้นไม้ตัดสินใจจะใช้การพิจารณาความแตกต่างกันภายในเอนโทรปี (Information entropy) เพื่อแบ่งกลุ่มข้อมูลไปสู่กลุ่มย่อย ๆ สำหรับวิธีการเลือกราก (Root) จะใช้การคำนวณเอนโทรปีเพื่อหาค่าสูงสุดหรือค่าเกน (Gain) มาสร้างเป็นราก และคำนวณค่าการแบ่งแยก (Split information) เปรียบเทียบอัตราส่วนเพื่อคำนวณอัตราส่วนเกน (Gain ratio) ด้วยสมการ

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (1)$$

ทำการคำนวณสมการ (1) ซ้ำหลังจากที่ได้รากเรียบร้อยแล้ว เพื่อหาโหนดต่อไปที่มีน้ำหนักรองลงมาจากโหนดราก

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) คือการหาระนาบในการตัดสินใจ โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้ง 2 กลุ่มให้มากที่สุด ภายในบทความ (Lertsuksakda, et al., 2015) แสดงถึงการวิเคราะห์อารมณ์จากนิทานด้วยการใช้โมเดลซัพพอร์ตเวกเตอร์แมชชีน โดยการทำนายผลออกเป็นอารมณ์ บวก (Positive), ลบ (Negative) และเป็นกลาง (Neutral) สำหรับการวิเคราะห์จะใช้การจับคู่แบบ 1 ต่อทั้งหมด (All) ในแต่ละคลาส ตัวอย่างเช่น คลาสบวก (Positive) ต่อคลาสลบ (Negative) และคลาสเป็นกลาง (Neutral) เพื่อหา

ผลลัพธ์ของข้อมูล มิติของข้อมูลสามารถเขียนแทนด้วย (X_i, y_i) และมีสมการการคำนวณการตัดสินใจคือ

$$W \times X + b > 0 \qquad ; y_i = 1 \tag{2}$$

$$W \times X + b < 0 \qquad ; y_i = -1 \tag{3}$$

โดยที่ w หมายถึงค่าน้ำหนัก และ b คือค่าความเอนเอียง หลังจากที่ได้เส้นการตัดสินใจหรือเส้นแบ่ง (Hyperplane) เรียบร้อยแล้วจะดูว่าเส้นแบ่งใดที่เหมาะสมที่สุดในการจัดกลุ่มของข้อมูล โดยดูจากระยะห่างของเส้นแบ่งถึงเส้นตรงที่ผ่านข้อมูลที่ใกล้ที่สุดและขนานกับเส้นแบ่งของทั้ง 2 กลุ่ม โดยระยะห่างนี้เรียกว่า Margin (จิรา แก้วสุวรรณ์, 2006)

2.3.5 ตัวชี้วัดการประเมินผล

การประเมินความถูกต้องของโมเดลการจำแนกข้อมูลจะใช้การประเมินผลจากสมการคำนวณค่า Precision, Recall และ F-measure ในการประเมิน โดยในบทความ (Gao, et al., 2010) ได้แสดงสมการการคำนวณ Precision ดังสมการ

$$Precision = TP \div (TP + TN) \tag{4}$$

สมการการคำนวณ Recall ดังสมการ

$$Recall = TP \div (TP + FP) \tag{5}$$

สมการการคำนวณ F-measure ดังสมการ

$$F\text{-}measure = (2 \times Precision \times Recall) \div (Precision + Recall) \tag{6}$$

2.3.6 การประยุกต์ใช้ในงานที่ใกล้เคียง

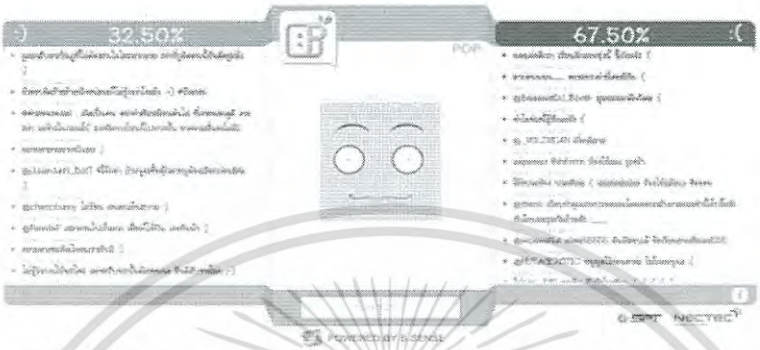
การประยุกต์ใช้การทำเหมืองข้อมูลด้านการประมวลผลทางภาษา มีงานวิจัย (Haruechaiyasak, et al., 2013) นำเสนอเครื่องมือสำหรับการวิเคราะห์ตลาดคือ ระบบวิเคราะห์ความคิดเห็นของบุคคลทั่วไปหรือกลุ่มลูกค้าบนโซเชียลมีเดีย (S-sense) สร้างขึ้นเพื่อวัดอารมณ์หรือความรู้สึกของบุคคลภายในโซเชียลมีเดีย ตัวอย่างเช่น ทวิตเตอร์ (Twitter) และเว็บบอร์ดพันทิพย์ (Pantip webboard) ด้วยเทคนิคการประมวลผลทางภาษา ประกอบกับเทคนิคการทำเหมืองข้อความ เพื่อดูความรู้สึกหรืออารมณ์ของผู้ใช้หรือลูกค้าที่มีต่อผลิตภัณฑ์ที่ผู้ใช้หรือลูกค้าให้ความสนใจ

Intention		Example
Announcement		อัตราค่าบริการ Happy Bonus ปรับปรุงใหม่จ้ะ เริ่มใช้วันที่ 1 คะ The new service fee for Happy Bonus will start on the 1st of this month.
		โปรใหม่! ทรูมูฟ... ซิมสุดคุ้ม โปรวินาทีละ 1 ส.ต. ตลอด 24 ชั่วโมง New promotion!! True Move... Best-deal SIM, 1 satang / second all day and night.
Request		สมัครใช้บริการ Call Screening เองไม่ได้ CC ช่วยด้วยครับ I can't apply for Call Screening myself. CC (Call Center), please help me.
		ขอทวนCC AISหน่อยคะ..เงินในโทรศัพท์หายไปไหนไม่รู้ (- -)?? AIS Call Center, please.. My pre-paid balance has gone missing without a clue ??
Question		โทรศัพท์หาย จะทำซิมใหม่เบอร์เดิมของ ais ต้องใช้เอกสารอะไรบ้างครับ I lost my phone. To get a new SIM card, what documents are required?
		โปรไหนของ one-2-call ที่รอรับสายได้นานสุดครับ Which promotion package of one-2-call allows the longest call waiting time?
Sentiment	Negative	น่ารำคาญมาก DTAC เมื่อไหร่จะปรับปรุงสัญญาณสักที โดยเฉพาะบนBTS Very annoyed. DTAC, when will you improve the signal? Especially on the BTS.
	Positive	ขอบคุณและชื่นชม เจ้าหน้าที่ AIS serenade call center ประทับใจครับ Thank you to the operator at AIS serenade call center. Very impressive.

ภาพที่ 2.6 รูปแบบการวิเคราะห์ของ S-sense (Haruechaiyasak, et al., 2013)

ระบบของ S-sense มุ่งเน้นไปที่การวิเคราะห์ 2 ประเภทคือ 1) วัตถุประสงค์ของการโพสต์และ 2) อารมณ์ ตัวอย่างเช่น การประกาศ (Announcement), การร้องขอ (Request), การถามคำถาม (Question) และอารมณ์ (Sentiment) โดนระบบสามารถวิเคราะห์ในหัวข้อต่าง ๆ ประกอบด้วย การติดตามแบรนด์ (Brand monitoring), การติดตามกิจกรรมที่จัด (Campaign monitoring), การวิเคราะห์การแข่งขัน (Competitive analysis) และความผูกพันของลูกจ้าง (Employee engagement) และแสดงผลในรูปแบบอีโมติคอนที่สามารถปรับเปลี่ยนหน้าตาได้ (Adaptive emoticon) ตามอารมณ์หรือความรู้สึกของข้อความ

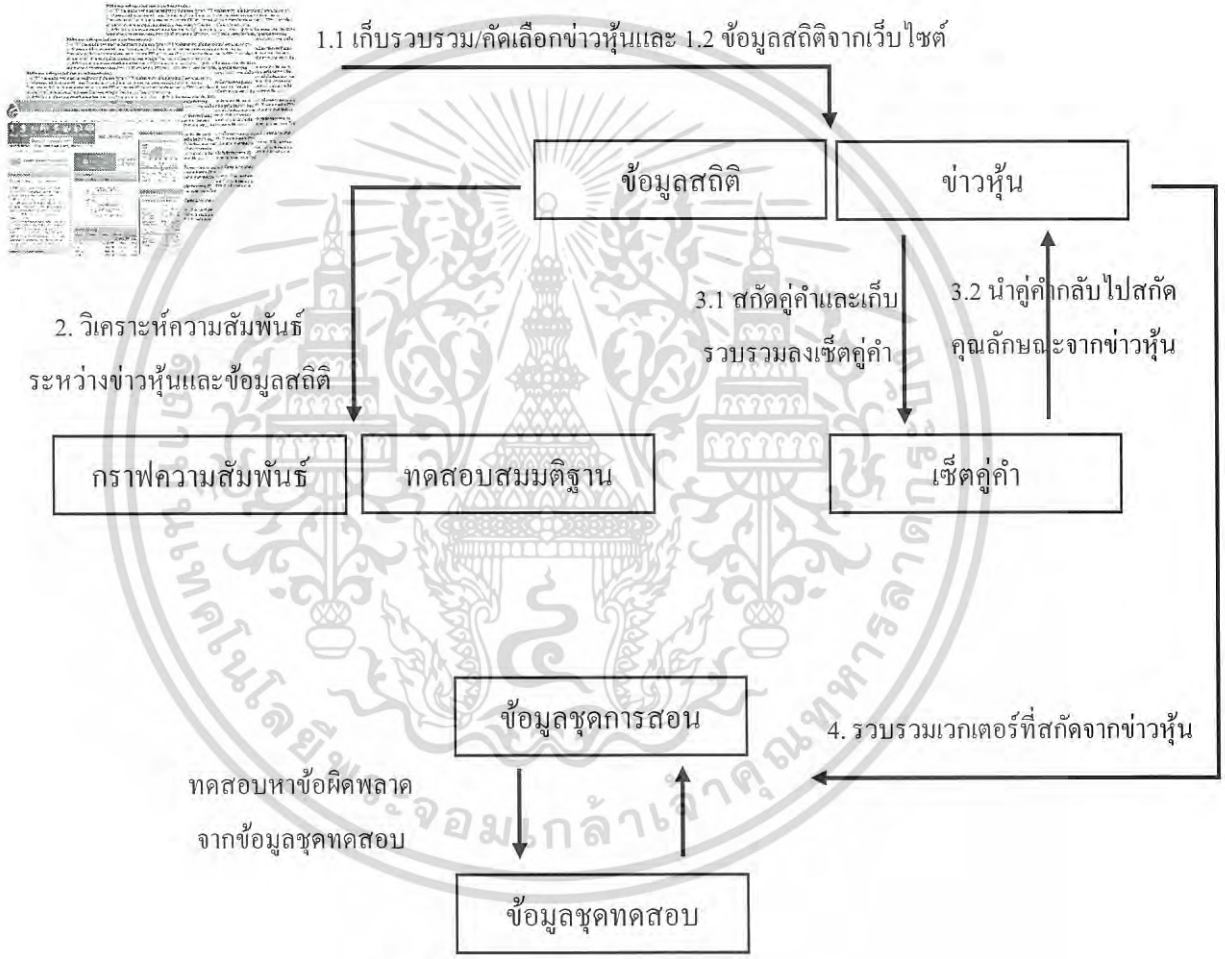
ตัวอย่างของแอพพลิเคชันที่มีการนำ S-sense ไปประยุกต์ใช้คือ แอพพลิเคชันวัดความรู้สึกของคนไทยบนโซเชียลมีเดีย (POP) โดยมีการเก็บรวบรวมข้อความเพื่อใช้สำหรับวิเคราะห์อารมณ์ในเชิงบวก และอารมณ์ในเชิงลบ คำนวณค่าน้ำหนักออกเป็นเปอร์เซนต์ในขณะนั้น แสดงตัวอย่างแอพพลิเคชันวัดความรู้สึก POP ดังภาพที่ 2.7



ภาพที่ 2.7 ตัวอย่างแอพพลิเคชันวัดความรู้สึก POP (NECTEC, 2013)

บทที่ 3
วิธีการวิจัยและการเก็บรวบรวมข้อมูล

ในบทนี้จะกล่าวถึงวิธีการวิจัย โดยแสดงขั้นตอนต่าง ๆ ตั้งแต่การคัดเลือกและการเก็บรวบรวมข้อมูล วิธีการประเมินความสอดคล้องกันของข้อมูลที่เก็บรวบรวมและได้นำเสนอการสกัดคู่คำเพื่อใช้เป็นคุณลักษณะในการวิเคราะห์ข่าวหุ้น โดยแต่ละขั้นตอนแสดงในภาพที่ 3.1 ดังต่อไปนี้



ภาพที่ 3.1 แผนภาพวิธีการวิจัย

3.1 การคัดเลือกและการเก็บรวบรวมข้อมูล

การคัดเลือกและการเก็บรวบรวมข้อมูล ในงานวิจัยนี้มีการแบ่งกลุ่มของข้อมูลออกเป็นสองประเภทประกอบด้วย ข้อมูลข่าวหุ้น โดยของเขตของข่าวหุ้นที่ใช้ในงานวิจัยนี้มุ่งสนใจเฉพาะข่าวหุ้นที่เป็นภาษาไทยเท่านั้น และข้อมูลทางสถิติ ประกอบด้วยการเปลี่ยนแปลงราคาหุ้นและปริมาณการซื้อขายหุ้น

สำหรับข้อมูลข่าวหุ้นมีการคัดเลือกข้อมูลจากโบรคเกอร์และเว็บไซต์ให้บริการข่าวหุ้นต่าง ๆ ตัวอย่างของโบรคเกอร์ได้แก่ บล. บัวหลวง, บล. ธนชาต และ บล. กรุงศรี เป็นต้น สำหรับตัวอย่างของเว็บไซต์ให้บริการข่าวหุ้นได้แก่ ข่าวหุ้นธุรกิจออนไลน์ เป็นต้น ข่าวหุ้นที่คัดเลือกมาจากโบรคเกอร์และเว็บไซต์ จะถูกแบ่งออกเป็นสองกลุ่ม ประกอบด้วย ข่าวหุ้นสำหรับชุดการสอน (Training set) และข่าวหุ้นสำหรับชุดทดสอบ (Testing set) ข่าวหุ้นสำหรับชุดการสอนจะคัดเลือกและเก็บรวบรวมจาก บล. บัวหลวง เท่านั้น เนื่องด้วยเหตุผลว่า ข่าวหุ้นที่เก็บรวบรวมจากโบรคเกอร์นี้จะมีการแสดงเครื่องหมายกำกับอารมณ์มาพร้อมกับข่าวเรียบร้อยแล้ว การกำกับเครื่องหมายโดยโบรคเกอร์ทำให้มีความน่าเชื่อถือในระดับหนึ่ง สำหรับตัวอย่างของเครื่องหมายกำกับอารมณ์ ได้แก่ +, -, 0 หรือ * ส่วนข่าวหุ้นสำหรับชุดทดสอบจะคัดเลือกและเก็บรวบรวมจากโบรคเกอร์ต่าง ๆ และเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์ เนื่องด้วยเหตุผลว่า ข่าวหุ้นสำหรับชุดทดสอบควรจะมีความหลากหลายของข้อมูลและเป็นข้อมูลที่ชุดการสอนไม่เคยพบมาก่อน

ข่าวหุ้นที่เก็บรวบรวมจาก บล. บัวหลวง ถูกเก็บรวบรวมระหว่างวันที่ 4 เมษายน 2014 ถึงวันที่ 27 พฤษภาคม 2015 มีจำนวนข่าวทั้งสิ้น 1,381 ข่าว สำหรับเงื่อนไขหลักในการคัดเลือกข่าวหุ้นก่อนการเก็บรวบรวมมี 2 ประการ

1) ภายในข่าวหุ้นจะต้องมีการกล่าวถึงสัญลักษณ์หุ้น (Stock symbol) อย่างน้อยหนึ่งตัวขึ้นไป เพื่อใช้ในการอ้างถึงความสัมพันธ์ระหว่างข้อความภายในข่าวและตรวจสอบความสัมพันธ์กับข้อมูลทางสถิติ สำหรับตัวอย่างของสัญลักษณ์หุ้น ได้แก่ CK, BTS และ BMCL เป็นต้น
2) ภายในข่าวหุ้นจะต้องสามารถสกัดคู่คำได้อย่างน้อยหนึ่งคู่ เนื่องจากเนื้อหาของข่าวหุ้นบางส่วนจะเป็นการแนะนำราคาหรือเพียงแค่อธิบายภาพรวมตลาด ไม่ได้กล่าวถึงสัญลักษณ์หุ้นตัวใดตัวหนึ่งอย่างเจาะจง สำหรับคู่คำภายในข่าวจะใช้ในการวิเคราะห์หรือเป็นตัวแทนของข่าว จำนวนข่าวที่เก็บรวบรวมจาก บล. บัวหลวง เมื่อสกัดสัญลักษณ์หุ้นออกจากข่าวเรียบร้อยแล้ว โดยในหนึ่งข่าวจะกล่าวถึงเพียงหนึ่งสัญลักษณ์หุ้น จะมีจำนวนทั้งสิ้น 6,596 ข่าว โดยความยาวเฉลี่ยของหนึ่งข่าวยาวประมาณ 200 ถึง 500 ตัวอักษร

ตารางที่ 3.1 แสดงตัวอย่างของข่าวหุ้นที่เก็บรวบรวมจาก บล. บัวหลวง สำหรับข่าวที่หนึ่งแสดงสัญลักษณ์หุ้นภายในข่าว และข่าวที่สองแสดงข่าวที่อธิบายภาพรวมตลาด ไม่มีการอ้างอิงถึงสัญลักษณ์หุ้นและไม่สามารถสกัดคู่คำอย่างน้อยหนึ่งคู่ออกจากข่าวได้ โดยทั้งสองตัวอย่างจะมีเครื่องหมายกำกับอารมณ์ของข่าวกำกับไว้ด้วย และตารางที่ 3.2 แสดงตัวอย่างการสกัดสัญลักษณ์หุ้นออกจากข่าว

ตารางที่ 3.1 ข่าวหุ้นที่เก็บรวบรวมจาก บล. บัวหลวง

วันที่	ข่าว	เครื่องหมายกำกับอารมณ์	สัญลักษณ์หุ้น
06/03/2015	CK เราคงคำแนะนำ ซื้อ TP 32.75 บาท คงมุมมองเชิงบวกหลังประชุมนักวิเคราะห์เมื่อวาน และเราคงประมาณการณ์กำไรปีนี้ 1.86 พันลบ. (-22% y-y) ประเด็นการลงทุน 1) คาดได้งานระบบ สายสีน้ำเงินส่วนต่อขยายจาก BMCL 2) มีโอกาสได้งานอุโมงค์ระบายน้ำ กทม. 6 พันลบ. และ คาดได้แบ่งงาน ถนนมอเตอร์เวย์พัทยา-มาบตาพุด 3) แนวโน้มกำไรสุทธิ 1Q15 เติบโตสูงจากกำไรขาย ไซยะบุรี	+	CK, BMCL
17/02/2015	วันนี้ยังมีการประชุม EU group meeting ต่อรอลุ้นแผนอุ้มหนี้กรีซ	+	-

ตารางที่ 3.2 การสกัดสัญลักษณ์หุ้นออกจากข่าว

วันที่	ข่าว	เครื่องหมายกำกับอารมณ์	สัญลักษณ์หุ้น
06/03/2015	CK เราคงคำแนะนำ ซื้อ TP 32.75 บาท คงมุมมองเชิงบวกหลังประชุมนักวิเคราะห์เมื่อวาน และเราคงประมาณการณ์กำไรปีนี้ 1.86 พันลบ. (-22% y-y) ประเด็นการลงทุน…	+	CK
06/03/2015	… 1) คาดได้งานระบบ สายสีน้ำเงินส่วนต่อขยายจาก BMCL…	+	BMCL

ข่าวหุ้นที่เก็บรวบรวมจากโบรคเกอร์อื่น ๆ และเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์ ยกเว้น บล. บัวหลวง เพื่อใช้เป็นข้อมูลสำหรับชุดทดสอบ ถูกเก็บรวบรวมระหว่างวันที่ 10 มีนาคม 2015 ถึงวันที่ 3 กรกฎาคม 2015 มีจำนวนข่าวทั้งสิ้น 327 ข่าว

สำหรับเงื่อนไขหลักในการคัดเลือกข่าวหุ้นก่อนการเก็บรวบรวมใช้เงื่อนไขเดียวกันกับชุดข้อมูลการสอนที่กล่าวไปก่อนหน้า จำนวนข่าวที่เก็บรวบรวมเมื่อสกัดสัญลักษณ์หุ้นออกจากข่าวเรียบร้อยแล้ว โดยในหนึ่งข่าวจะกล่าวถึงเพียงหนึ่งสัญลักษณ์หุ้น จะมีจำนวนทั้งสิ้น 3,489 ข่าว โดยความยาวเฉลี่ยของหนึ่งข่าวยาวประมาณ 500 ถึง 1,000 ตัวอักษร

จะเห็นได้ว่าปริมาณความยาวเฉลี่ยของข่าวในชุดทดสอบจะมีความยาวมากกว่าชุดการสอน ส่งผลให้ปริมาณของข่าวหลังจากการสกัดสัญลักษณ์หุ้นออกจากข่าวเรียบร้อยแล้วมีปริมาณมากขึ้นกว่าปริมาณข่าวเดิมเป็นอย่างมาก ตารางที่ 3.3 แสดงตัวอย่างของข่าวหุ้นที่เก็บรวบรวมจากโบรคเกอร์อื่น ๆ และเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์ใน

สำหรับข่าวที่เก็บรวบรวมจากโบรคเกอร์อื่น ๆ และเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์จะไม่มีเครื่องหมายกำกับอารมณ์จากโบรคเกอร์มาให้ ในการเก็บรวบรวมจึงใช้เครื่องหมาย '?' เพื่อใช้ในการทำนายเครื่องหมายกำกับอารมณ์จากโมเดลที่สร้างขึ้นจากข้อมูลชุดการสอนในหัวข้อถัดไป

ตารางที่ 3.3 ข่าวหุ้นที่เก็บรวบรวมจากโบรคเกอร์อื่น ๆ และเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์

วันที่	ข่าว	เครื่องหมายกำกับอารมณ์	สัญลักษณ์หุ้น
02/04/2015	บล.ธนชาต ระบุในบทวิเคราะห์ (1 เม.ย.) ว่า แนวโน้ม SET ระยะสั้นมีจังหวะ ฟื้นตัว... 1) ยกเลิกกฎอัยการศึก หนุนกลุ่มท่องเที่ยว ... 5) SET เกิดสัญญาณ Morning Star เป็นสัญญาณกลับตัว ทางเทคนิค ตั้งแต่ต้นสัปดาห์ หุ้นกลุ่มรับเหมาฯ และเกี่ยวข้องกับนโยบายภาครัฐฯ อาจได้รับผลดีจากการใช้ ม.44 เนื่องจากทำให้กระบวนการในลงทุนกระชับ และใช้กระตุ้นเศรษฐกิจได้ดีขึ้น แนะนำ ซื้อ กลุ่มรับเหมาฯ วัสดุก่อสร้าง SCC CK STEC และ SEAFCO กลุ่มที่ลงทุนเกี่ยวกับพลังงานทางเลือก EA SAMART และ TPIPL กลุ่มท่องเที่ยว AOT รวมถึงกลุ่มหุ้นใหญ่ KBANK	?	SCC, CK, STEC, SEAFCO, EA, SAMART, TPIPL, AOT, KBANK

นอกจากสัญลักษณ์หุ้นที่สกัดออกจากข่าวแล้ว ยังมีการเพิ่มสัญลักษณ์ของกลุ่มอุตสาหกรรมที่หุ้นตัวนั้น ๆ สังกัดอยู่ ด้วยเหตุผลสำหรับการเปรียบเทียบกับข้อมูลทางสถิติในระดับหุ้นและระดับกลุ่มอุตสาหกรรม โดยกลุ่มอุตสาหกรรมที่ใช้มีทั้งหมด 8 กลุ่ม แสดงตัวอย่างสัญลักษณ์ของกลุ่มอุตสาหกรรมพร้อมกับสัญลักษณ์หุ้นบางส่วนในกลุ่มอุตสาหกรรมนั้น ๆ ในตารางที่ 3.4

ตารางที่ 3.4 รายชื่อกลุ่มอุตสาหกรรมทั้ง 8 กลุ่ม สัญลักษณ์ย่อและตัวอย่างของสัญลักษณ์หุ้นที่สังกัดในกลุ่มอุตสาหกรรมนั้น ๆ

กลุ่มอุตสาหกรรม	สัญลักษณ์ย่อ	สัญลักษณ์หุ้น
เกษตรและอุตสาหกรรมอาหาร	Agro	ICHI, MINT, OISHI
สินค้าอุปโภคบริโภค	Consump	SABINA, SIAM, TR
ธุรกิจการเงิน	Fincial	SCB, KTC, BLA
สินค้าอุตสาหกรรม	Indus	CTW, IVL, ALUCON
อสังหาริมทรัพย์และก่อสร้าง	Propcon	DCON, SCC, AP
ทรัพยากร	Resourc	BANPU, ESSO, PTT
บริการ	Service	ROBINS, PLANB, BTS
เทคโนโลยี	Tech	DELTA, DTAC, JAS

จำนวนของสัญลักษณ์หุ้นที่เก็บรวบรวมจากข้อมูลชุดการสอนและข้อมูลชุดทดสอบ มีการวิเคราะห์เบื้องต้นด้วยการนับความถี่ แสดงในตารางที่ 3.5 โดยกลุ่มอุตสาหกรรมที่มีหุ้นถูกกล่าวถึงมากที่สุดคือกลุ่มอสังหาริมทรัพย์และก่อสร้าง (Propcon) รองลงมาคือกลุ่มบริการ (Service) โดยทั้งในข้อมูลชุดการสอนจาก บล. บัวหลวง และในข้อมูลชุดทดสอบจากโบรคเกอร์อื่น ๆ พบว่ามีความถี่เป็นไปในทิศทางเดียวกัน ในทางตรงกันข้ามกลุ่มอุตสาหกรรมที่มีหุ้นถูกกล่าวถึงน้อยที่สุดคือกลุ่มสินค้าอุปโภคบริโภค (Consump) รองลงมาคือกลุ่มสินค้าอุตสาหกรรม (Indus) ตามลำดับ โดยข้อมูลทั้งชุดการสอน และข้อมูลชุดทดสอบ มีความสอดคล้องไปในทิศทางเดียวกัน

สำหรับตารางที่ 3.6 แสดงความถี่เฉพาะของสัญลักษณ์หุ้น เทียบกับปริมาณเฉพาะของสัญลักษณ์หุ้นในดัชนี SET พบว่าในข้อมูลชุดการสอนมีสัญลักษณ์หุ้นเฉพาะ 296 สัญลักษณ์ และ 219 สัญลักษณ์ในชุดทดสอบ จากปริมาณสัญลักษณ์หุ้นในดัชนี SET ทั้งหมด 572 สัญลักษณ์ คิดเป็น 51.74% สำหรับปริมาณเฉพาะที่ถูกกล่าวถึงในข้อมูลชุดการสอน และ 38.28% สำหรับปริมาณเฉพาะที่ถูกกล่าวถึงในข้อมูลชุดทดสอบ

ตารางที่ 3.5 ปริมาณความถี่ของสัญลักษณ์หุ้นในข้อมูลชุดการสอน และข้อมูลชุดทดสอบ จัดกลุ่มตามกลุ่มอุตสาหกรรม

กลุ่มอุตสาหกรรม	บล. บัวหลวง		โบรคเกอร์อื่น ๆ	
	ความถี่	ลำดับ	ความถี่	ลำดับ
เกษตรและอุตสาหกรรมอาหาร	780	3	303	5
สินค้าอุปโภคบริโภค	310	8	139	8
ธุรกิจการเงิน	514	5	345	3
สินค้าอุตสาหกรรม	359	7	189	7
อสังหาริมทรัพย์และก่อสร้าง	2512	1	1356	1
ทรัพยากร	629	4	265	6
บริการ	1094	2	562	2
เทคโนโลยี	398	6	330	4

ตารางที่ 3.6 ปริมาณสัญลักษณ์หุ้นเฉพาะในดัชนี SET, ข้อมูลชุดทดสอบ และข้อมูลชุดการสอน จัดกลุ่มตามกลุ่มอุตสาหกรรม

กลุ่มอุตสาหกรรม	สัญลักษณ์หุ้นเฉพาะในดัชนี SET	สัญลักษณ์หุ้นเฉพาะในข้อมูลชุดการสอน	สัญลักษณ์หุ้นเฉพาะในข้อมูลชุดทดสอบ
เกษตรและอุตสาหกรรมอาหาร	54	32	29
สินค้าอุปโภคบริโภค	42	12	8
ธุรกิจการเงิน	61	32	27
สินค้าอุตสาหกรรม	87	32	17
อสังหาริมทรัพย์และก่อสร้าง	152	76	53
ทรัพยากร	36	28	20
บริการ	99	57	41
เทคโนโลยี	41	27	23
ผลรวมกลุ่มอุตสาหกรรม	572	296	219

3.2 การวิเคราะห์ความสอดคล้องของข่าวหุ้น กับข้อมูลราคาและปริมาณการซื้อขายด้วยการจำลองกราฟ

การวิเคราะห์ความสอดคล้องของข่าวหุ้นกับข้อมูลทางสถิติด้วยการจำลองกราฟ ผู้วิจัยได้หาความสัมพันธ์ระหว่างการเปลี่ยนแปลงราคาหุ้นตัวนั้น ๆ และปริมาณการซื้อขายของหุ้นตัวนั้น ๆ โดยเมื่อทำการเปรียบเทียบกับข้อมูลดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรมของหุ้นตัวนั้น ๆ และดันชีกลุ่มธุรกิจของหุ้นตัวนั้น ๆ จะได้สมการในการแสดงผลกราฟ ดังนี้

$$nP_{(Stock,\ SET)} = (P_{Stock} + e) \div (P_{SET} + e) \tag{7}$$

$$nP_{(Stock,\ Industry)} = (P_{Stock} + e) \div (P_{Industry} + e) \tag{8}$$

$$nP_{(Stock,\ Sector)} = (P_{Stock} + e) \div (P_{Sector} + e) \tag{9}$$

จากสมการที่ (7) ถึง (9) จะเป็นสมการที่ใช้ในการแสดงผลกราฟของการเปลี่ยนแปลงราคาหุ้นตัวนั้น ๆ โดยที่ nP หมายถึง ค่าการเปลี่ยนแปลงราคาหุ้นของหุ้นและดัชนีตัวนั้น ๆ ที่ผ่านการคำนวณและการปรับค่า (Normalized) เรียบร้อยแล้ว

P_{Stock} หมายถึง การเปลี่ยนแปลงราคาของหุ้นตัวนั้น ๆ สำหรับ P_{SET}, $P_{Industry}$ และ P_{Sector} หมายถึง การเปลี่ยนแปลงราคาของดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม และดัชนีกลุ่มธุรกิจ ตามลำดับ และค่า e ที่อยู่ในสมการทั้งเศษและส่วน เป็นการเพิ่มความคลาดเคลื่อนเพื่อป้องกันค่าของตัวหารเป็นศูนย์ ในกรณีที่ไม่มีการเปลี่ยนแปลงราคาของวันก่อนหน้า และวันปัจจุบัน โดยค่า e ที่ใช้ในการวิเคราะห์นี้มีค่า 0.01

$$nV_{(Stock,\ SET)} = (V_{Stock} \div 10^{3}) \div (V_{SET} \div 10^{9}) \tag{10}$$

$$nV_{(Stock,\ Industry)} = (V_{Stock} \div 10^{3}) \div (V_{Industry} \div 10^{9}) \tag{11}$$

$$nV_{(Stock,\ Sector)} = (V_{Stock} \div 10^{3}) \div (V_{Sector} \div 10^{9}) \tag{12}$$

จากสมการที่ (10) ถึง (12) จะเป็นสมการที่ใช้ในการแสดงผลกราฟของปริมาณการซื้อขายของหุ้นตัวนั้น ๆ โดยที่ nV หมายถึง ค่าปริมาณการซื้อขายของหุ้นและดัชนีตัวนั้น ๆ ที่ผ่านการคำนวณและการปรับค่า (Normalized) เรียบร้อยแล้ว

V_{Stock} หมายถึง ปริมาณการซื้อขายของหุ้นตัวนั้น ๆ สำหรับ V_{SET}, $V_{Industry}$ และ V_{Sector} หมายถึง ปริมาณการซื้อขายของดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม และดัชนีกลุ่มธุรกิจ ตามลำดับ และค่า 10^3 และ 10^9 ที่อยู่ในสมการทั้งเศษและส่วน เป็นการปรับค่าเพื่อให้อยู่ในช่วงที่สามารถคำนวณค่าให้วิเคราะห์ได้อย่างสะดวก

สำหรับเส้นกราฟมาตรฐานที่ใช้ในการเปรียบเทียบ จะใช้ข้อมูลสถิติของดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม (Industry) และดัชนีระดับกลุ่มธุรกิจ (Sector) มาทำการปรับให้ได้ค่าคงที่ ที่ระดับ 1 ด้วยการหารด้วยค่าของตัวดัชนีเอง ดังสมการ (13) ถึง (15) สำหรับดัชนีการเปลี่ยนแปลงราคา

$$P_{(Standard,\ SET)} = P_{SET} \div P_{SET} \tag{13}$$

$$P_{(Standard,\ Industry)} = P_{Industry} \div P_{Industry} \tag{14}$$

$$P_{(Standard,\ Sector)} = P_{Sector} \div P_{Sector} \tag{15}$$

สำหรับปริมาณการซื้อขายของดัชนี จะใช้ข้อมูลสถิติของดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม (Industry) และดัชนีระดับกลุ่มธุรกิจ (Sector) มาทำการปรับให้ได้ค่าคงที่ ที่ระดับ 1 ด้วยการหารด้วยค่าของตัวดัชนีเอง ดังสมการ (16) ถึง (18)

$$V_{(Standard,\ SET)} = V_{SET} \div V_{SET} \tag{16}$$

$$V_{(Standard,\ Industry)} = V_{Industry} \div V_{Industry} \tag{17}$$

$$V_{(Standard,\ Sector)} = V_{Sector} \div V_{Sector} \tag{18}$$

3.3 การวิเคราะห์ความสอดคล้องแนวโน้มของข่าวหุ้น ด้วยการวิเคราะห์ทางสถิติ

การทดสอบสมมติฐาน (Hypothesis testing) เป็นการวัดความสอดคล้องของวันที่มีการประกาศข่าวจากโบรคเกอร์ว่ามีความแตกต่างจากค่าเฉลี่ยหรือค่าปกติหรือไม่ โดยจะใช้สถิติเครื่องหมาย-อันดับของวิลคอกซัน (Wilcoxon signed-rank test) ทำการทดสอบสมมติฐาน โดยมีการเลือกสมมติฐานสำหรับการเปลี่ยนแปลงราคาที่ได้กล่าวไปก่อนหน้านี้ในบทที่ 2 คือ

1)

$$H_0: x_i = \mu_0 \quad H_1: x_i \neq \mu_0$$

โดยที่สมมติฐานการเปลี่ยนแปลงราคาแย้ง $H_1: x_i \neq \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าการเปลี่ยนแปลงราคา 5 วันที่มีข่าว (x_i) ไม่เท่ากับค่าเฉลี่ยการเปลี่ยนแปลงราคาของวันที่ไม่มีข่าว (μ_0)

สมมติฐานสำหรับปริมาณการซื้อขาย คือ

2)

$$H_0: x_i \leq \mu_0 \quad H_1: x_i > \mu_0$$

โดยที่สมมติฐานปริมาณการซื้อขายแย้ง $H_1: x_i > \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าปริมาณการซื้อขายช่วง 5 วันที่มีข่าว (x_i) มากกว่าค่าเฉลี่ยปริมาณการซื้อขายของวันที่ไม่มีข่าว (μ_0)

แสดงตัวอย่างข้อมูลการเปลี่ยนแปลงราคาที่ใช้ในการคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซันในตารางที่ 3.7

ตารางที่ 3.7 แสดงตัวอย่างข้อมูลการเปลี่ยนแปลงราคาที่ใช้ในการคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน

วันที่	การเปลี่ยนแปลงราคา					วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)
	วันที่มีข่าว-2	วันที่มีข่าว-1	วันที่มีข่าว	วันที่มีข่าว+1	วันที่มีข่าว+2		
12/11/2014	0	0.95	-1.89	1.92	0	0.19	1.44
08/12/2014	0	0	-4.55	-0.95	-3.85	-1.87	1.44
23/01/2015	3.88	6.54	0	0	3.51	2.78	1.44
18/01/2015	0	-3.45	0.89	0	-0.88	-0.68	1.44

จากตารางที่ 3.7 ขั้นตอนการคำนวณ คือ

1) คำนวณค่าความแตกต่างของค่าสังเกตแต่ละตัว x_i กับค่าเฉลี่ย μ_0 คือ $d_i = x_i - \mu_0$
2) เรียงลำดับ d_i จากน้อยสุดไปมากที่สุดโดยไม่สนใจเครื่องหมาย
3) ให้แต่ละอันดับมีเครื่องหมายของ d_i
4) หา T^+ คือผลบวกของอันดับที่มีเครื่องหมายบวก
   T^- คือผลบวกของอันดับที่มีเครื่องหมายลบ

วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)	ค่าความแตกต่าง (d_i)	เรียงลำดับ d_i	เพิ่มเครื่องหมายของ d_i
0.19	1.44	-1.25	1	-1
-1.87	1.44	-3.31	4	-4
2.78	1.44	1.34	2	2
-0.68	1.44	-2.12	3	-3

$T^+ = 2$

$T^- = -8$

ภาพที่ 3.2 แสดงตัวอย่างการคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน

หลังจากคำนวณค่า T^+ และ T^- เรียบร้อยจะนำไปเปรียบเทียบกับค่าวิกฤต μ ดังสมการที่ (19)

$$\mu = n(n+1)/2 \tag{19}$$

จากสมการที่ (19) μ คำนวณได้จากจำนวนข้อมูล (n) มาเข้าสมการ เพื่อให้ได้ค่าวิกฤตที่ใช้ในการเปรียบเทียบ หลังจากนั้นจะนำไปเปรียบเทียบระหว่าง T และ μ ตามสมมติฐานที่ได้กำหนดไว้ก่อนหน้า

3.4 การวิเคราะห์รูปแบบและการสกัดคู่คำ

ความหมายของคู่คำ (Wordpairs) หมายถึง รูปแบบหรือตัวแทนของข้อความสำหรับใช้ในการวิเคราะห์ สำหรับโครงสร้างหรือรูปแบบของคู่คำ ประกอบด้วยข้อมูลสามกลุ่มเรียงต่อกันคั่นด้วยเครื่องหมายจุลภาค ข้อมูลทั้ง 3 กลุ่มประกอบด้วย

1) คีย์เวิร์ด (Keyword)
2) คำที่มีขั้ว (Polarity word)
3) เครื่องหมายกำกับอารมณ์ (Sentiment)

คีย์เวิร์ด หมายถึง คำหรือกลุ่มคำที่มีความหมายเฉพาะ และไม่บ่งบอกค่าอารมณ์ หรือขั้วภายในตัวเอง โดยส่วนใหญ่จะมีชนิดของคำเป็นคำนาม หรือคำกริยาในบางส่วน ตัวอย่างของคีย์เวิร์ดได้แก่ แนะนำ, ราคา และ รายได้ เป็นต้น

คำที่มีขั้ว หมายถึง คำหรือกลุ่มคำที่ทำหน้าที่ขยายความหมายของคีย์เวิร์ดให้มีความชัดเจนขึ้นในด้านของอารมณ์และความสมบูรณ์ของความหมาย คำที่มีขั้วจะมีค่าอารมณ์ หรือขั้วภายในตัวเอง โดยส่วนใหญ่จะมีชนิดของคำเป็นคำกริยา คำคุณศัพท์ และคำวิเศษณ์ ตัวอย่างของคำที่มีขั้ว ได้แก่ กำไร, ซื้อ และ คงที่ เป็นต้น

เครื่องหมายกำกับอารมณ์ หมายถึง เครื่องหมายที่บ่งบอกด้านอารมณ์ของคำที่มีขั้ว โดยในงานวิจัยนี้มีการจำแนกเครื่องหมายกำกับอารมณ์ออกเป็น 3 รูปแบบ ได้แก่

1) '1' หมายถึง คำที่มีขั้วเป็นบวก
2) '-1' หมายถึง คำที่มีขั้วเป็นลบ
3) '0' หมายถึง คำที่มีขั้วเป็นกลาง (หมายเหตุ คำที่มีขั้วเป็นกลาง และคำที่ไม่มีขั้ว มีค่า 'ไม่เท่ากัน')

การเก็บรวบรวมและการสกัดคู่คำ สำหรับงานวิจัยนี้มีการแบ่งคู่คำออกเป็น 3 รูปแบบอ้างอิงตามรูปแบบการสร้างคู่คำ ประกอบด้วย

1) การสกัดและการเก็บรวบรวมคู่คำด้วยมือ (Manual wordpairs extraction – ME)
2) การสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ (Manual wordpairs addition – MA)
3) การสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน (Automate wordpairs combination – AC)

การสกัดและการเก็บรวบรวมคู่คำด้วยมือ (Manual wordpairs extraction – ME) หมายถึง การใช้มนุษย์ในการสกัด และเก็บรวบรวมคู่คำออกจากข่าวหุ้น โดยใช้การอ้างอิงจากรูปแบบที่กล่าวไปก่อนหน้านี้ คือ คีย์เวิร์ด, คำที่มีขั้ว และเครื่องหมายกำกับอารมณ์ สำหรับการสกัดคู่คำด้วยวิธีการนี้

จะได้คู่คำส่วนใหญ่ที่ปรากฏอยู่ในข่าวจริง ตารางที่ 3.8 แสดงคู่คำที่ได้จากวิธีการสกัดและเก็บรวบรวมคู่คำด้วยมือ

ปริมาณของคู่คำที่ใช้วิธีการสกัดและเก็บรวบรวมคู่คำด้วยมือนี้มีจำนวนทั้งสิ้น 133 คู่คำ แบ่งออกเป็น

1) คู่คำที่มีขั้วอารมณ์เป็นบวกจำนวน 68 คู่คำ
2) คู่คำที่มีขั้วอารมณ์เป็นลบจำนวน 50 คู่คำ
3) คู่คำที่มีขั้วอารมณ์เป็นกลางจำนวน 15 คู่คำ

ตารางที่ 3.8 คู่คำจากวิธีการสกัดและเก็บรวบรวมคู่คำด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
1	เงินบาท	อ่อนค่า	1
2	เงินปันผล	จ่าย	1
3	เงินปันผล	สูง	1
4	เป้า	เติบโต	1
5	แนะ	เก็งกำไร	1
6	แนะ	ซื้อ	1
7	แนะ	รอซื้อ	1
8	แนะนำ	เก็งกำไร	1
9	แนะนำ	ซื้อ	1
10	โมเมนตั้ม	บวก	1
11	โอกาส	กำไร	1
12	โอกาส	ขาย	1
13	การเติบโต	ดี	1
14	การเติบโต	สูง	1
15	การเติบโต	หนุน	1
16	การเปลี่ยนแปลง	เชิงบวก	1
17	การลงทุน	ฟื้นตัว	1
18	กำไร	เด่น	1
19	กำไร	เติบโต	1
20	กำไร	เพิ่มขึ้น	1
21	กำไร	โต	1
22	กำไร	ดี	1
23	กำไร	ปรับเพิ่ม	1
24	กำไร	ปรับขึ้น	1
25	กำไร	พิเศษ	1
26	กำไร	ฟื้นตัว	1
27	กำไร	หนุน	1
28	กิจการ	ซื้อ	1
29	ขาดทุน	น้อยลง	1
30	คาด	กำไร	1
31	คาดการ	กำไร	1
32	จังหวะ	ดี	1
33	ตลาด	เติบโต	1
34	ตลาด	ฟื้นตัว	1
35	ทุน	เพิ่ม	1
36	ธุรกิจ	เติบโต	1
37	ธุรกิจ	ดีขึ้น	1
38	ธุรกิจ	ฟื้นตัว	1
39	ปัจจัย	หนุน	1
40	ปันผล	จ่าย	1
41	ผลการดำเนินงาน	เด่น	1

ตารางที่ 3.8 (ต่อ) คู่คำจากวิธีการสกัดและเก็บรวบรวมคู่คำด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
42	ผลการดำเนินงาน	โต	1
43	ผลการดำเนินงาน	กำไร	1
44	ผลการดำเนินงาน	ดี	1
45	ผลตอบแทน	ดี	1
46	ผลตอบแทน	สูง	1
47	ภาษี	ลด	1
48	มุมมอง	บวก	1
49	มูลค่าพื้นฐาน	บวก	1
50	ยอดขาย	เติบโต	1
51	ยอดขาย	เพิ่ม	1
52	ยอดขาย	ดี	1
53	ยอดจอง	ปรับสูง	1
54	ราคา	ปรับเพิ่ม	1
55	ราคาหุ้น	ขึ้น	1
56	ราคาหุ้น	บวก	1
57	รายได้	เติบโต	1
58	รายได้	เพิ่ม	1
59	รายได้	ดี	1
60	รายได้	สูง	1
61	รายได้	หนุน	1
62	รายงาน	กำไร	1
63	ลงทุน	เพิ่ม	1
64	สินทรัพย์	เพิ่ม	1
65	หุ้น	บวก	1
66	อุตสาหกรรม	เด่น	1
67	อุตสาหกรรม	ดีขึ้น	1
68	น้ำหนักลงทุน	เพิ่ม	1
69	เศรษฐกิจ	ชะลอ	-1
70	แนะ	ขาย	-1
71	แนะ	หลีกเลี่ยง	-1
72	แนะนำ	ขาย	-1
73	การเมือง	ผลกระทบ	-1
74	การฟื้นตัว	ชะลอ	-1
75	การฟื้นตัว	ช้า	-1
76	การลงทุน	ไม่อนุมัติ	-1
77	กำไร	แย่	-1
78	กำไร	ชะลอลง	-1
79	กำไร	ต่ำ	-1
80	กำไร	ถ่วง	-1
81	กำไร	น้อย	-1
82	กำไร	ปรับลด	-1
83	กำไร	ลงต่อเนื่อง	-1
84	กำไร	ลดลง	-1
85	กิจการ	ขาย	-1
86	ข่าว	ลบ	-1
87	ค่าใช้จ่าย	เพิ่มขึ้น	-1
88	คาด	แย่	-1
89	งบ	ขาดทุน	-1
90	ดอกเบี้ย	แพง	-1
91	ดอกเบี้ย	ลด	-1
92	ดอกเบี้ย	สูง	-1
93	ทรัพย์สิน	ขาย	-1
94	ธุรกิจ	ไม่พัฒนา	-1
95	น้ำหนักลงทุน	ลง	-1

ตารางที่ 3.8 (ต่อ) คู่คำจากวิธีการสกัดและเก็บรวบรวมคู่คำด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
96	ปรับกำไร	ลง	-1
97	ผลกระทบ	เชิงลบ	-1
98	ผลการดำเนินงาน	แย่	-1
99	ผลการดำเนินงาน	ขาดทุน	-1
100	มุมมอง	ลบ	-1
101	ราคา	ปรับลง	-1
102	ราคา	ปรับลด	-1
103	ราคา	ลง	-1
104	ราคาน้ำมัน	ลง	-1
105	ราคาน้ำมัน	ลบ	-1
106	ราคาหุ้น	เสี่ยง	-1
107	ราคาหุ้น	ไม่ตอบรับ	-1
108	ราคาหุ้น	ขาด	-1
109	ราคาหุ้น	ปรับลง	-1
110	ราคาหุ้น	ลง	-1
111	ราคาหุ้น	อ่อนลง	-1
112	รายได้	เสีย	-1
113	รายได้	แย่	-1
114	รายได้	ลดลง	-1
115	รายงาน	ขาดทุน	-1
116	สต็อก	ขาดทุน	-1
117	หุ้น	ลง	-1
118	อุตสาหกรรม	กังวล	-1
119	แนะ	ถือ	0
120	แนะนำ	ถือ	0
121	การลงทุน	ชะลอ	0
122	กำไร	ทรงตัว	0
123	กำไร	ทรง ๆ	0
124	คาด	Neutral	0
125	มุมมอง	Neutral	0
126	รัฐบาล	อนุมัติ	0
127	ราคา	เป้าหมาย	0
128	ราคา	ต่ำ	0
129	ราคา	ถูก	0
130	ราคา	พักตัว	0
131	รายได้	ทรงตัว	0
132	หุ้น	ขาย	0
133	หุ้น	สะสม	0

การสกัดและการเก็บรวบรวมคู่คำด้วยมือ มีข้อสังเกตในเรื่องของปริมาณคู่คำในแต่ละขั้วอารมณ์ เนื่องจากคู่คำที่เก็บรวบรวมส่วนใหญ่จะมีขั้วอารมณ์ด้านบวกและขั้วอารมณ์ด้านลบมากกว่าขั้วอารมณ์เป็นกลาง จึงเกิดความไม่สมดุลกันระหว่างคู่คำแต่ละด้าน ดังนั้นการสร้างคู่คำใหม่ขึ้นจึงมีความสำคัญเพื่อแก้ไขปัญหาดังกล่าว

เทคนิคในการสร้างคู่คำเทคนิคแรก คือ การสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ (Manual wordpairs addition – MA) หมายถึง การสร้างคู่คำใหม่ด้วยการนำคำที่มีขั้วของคู่คำ มาหาคำที่มีขั้วตรงข้ามและคำที่มีขั้วเป็นกลางโดยมนุษย์ เพื่อเพิ่มคำที่มีขั้วเหล่านั้นให้กับคีย์เวิร์ดเดิม ตัวอย่างของคู่คำ ได้แก่ 'ราคาหุ้น, ขึ้น, 1' คำที่มีขั้วในคู่คำตัวอย่าง คือ 'ขึ้น' มีขั้วอารมณ์เป็นบวก ขั้วอารมณ์ตรง

ข้ามจึงเป็นลบ ตัวอย่าง คือ ‘ลง’ และขั้วอารมณ์เป็นกลาง ตัวอย่าง คือ ‘คงที่’ ดังนั้น คู่คำใหม่ที่เกิดจากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ ได้แก่ ‘ราคาหุ้น, ลง, -1’ และ ‘ราคาหุ้น, คงที่, 0’ คู่

ตารางที่ 3.9 แสดงคำที่ได้จากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ ปริมาณของคู่คำที่ใช้วิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือนี้มีจำนวนทั้งสิ้น 144 คู่คำ แบ่งออกเป็น

1) คู่คำที่มีขั้วอารมณ์เป็นบวกจำนวน 37 คู่คำ
2) คู่คำที่มีขั้วอารมณ์เป็นลบจำนวน 53 คู่คำ
3) คู่คำที่มีขั้วอารมณ์เป็นกลางจำนวน 54 คู่คำ

ตารางที่ 3.9 คู่คำจากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
134	เศรษฐกิจ	ฟื้นตัว	1
135	โอกาส	ซื้อ	1
136	การเมือง	ไม่มีผลกระทบ	1
137	การฟื้นตัว	เพิ่มขึ้น	1
138	การฟื้นตัว	เร็ว	1
139	การลงทุน	เพิ่มขึ้น	1
140	การลงทุน	อนุมัติ	1
141	กำไร	เพิ่ม	1
142	กำไร	เพิ่มต่อเนื่อง	1
143	กำไร	มาก	1
144	กำไร	สูง	1
145	ข่าว	บวก	1
146	ค่าใช้จ่าย	ลดลง	1
147	คาด	Positive	1
148	คาด	ดี	1
149	งบ	กำไร	1
150	ดอกเบี้ย	เพิ่ม	1
151	ดอกเบี้ย	ต่ำ	1
152	ดอกเบี้ย	ถูก	1
153	ทรัพย์สิน	ซื้อ	1
154	ธุรกิจ	พัฒนา	1
155	น้ำหนักลงทุน	ขึ้น	1
156	ปรับกำไร	ขึ้น	1
157	ผลกระทบ	เชิงบวก	1
158	มุมมอง	Positive	1
159	ราคา	ขึ้น	1
160	ราคา	ปรับขึ้น	1
161	ราคาน้ำมัน	ขึ้น	1
162	ราคาน้ำมัน	บวก	1
163	ราคาหุ้น	แข็งขึ้น	1
164	ราคาหุ้น	ตอบรับ	1
165	ราคาหุ้น	ปรับขึ้น	1
166	ราคาหุ้น	ปลอดภัย	1
167	รายได้	เพิ่มขึ้น	1
168	สต็อก	กำไร	1
169	หุ้น	ขึ้น	1
170	อุตสาหกรรม	สดใส	1
171	เงินบาท	แข็งค่า	-1
172	เงินปันผล	ไม่จ่าย	-1
173	เงินปันผล	ต่ำ	-1
174	เป้า	ลดลง	-1
175	แนะนำ	ขาย	-1

ตารางที่ 3.9 (ต่อ) คู่คำจากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
176	โมเมนตั้ม	ลบ	-1
177	โอกาส	ขาดทุน	-1
178	การเติบโต	ไม่ดี	-1
179	การเติบโต	กดดัน	-1
180	การเติบโต	ต่ำ	-1
181	การเปลี่ยนแปลง	เชิงลบ	-1
176	โมเมนตั้ม	ลบ	-1
177	โอกาส	ขาดทุน	-1
178	การเติบโต	ไม่ดี	-1
179	การเติบโต	กดดัน	-1
180	การเติบโต	ต่ำ	-1
181	การเปลี่ยนแปลง	เชิงลบ	-1
182	การลงทุน	ซบเซา	-1
183	กำไร	กดดัน	-1
184	กำไร	ถดถอย	-1
185	กำไร	ลด	-1
186	ขาดทุน	เพิ่มขึ้น	-1
187	คาด	Negative	-1
188	คาด	ขาดทุน	-1
189	คาดการ	ขาดทุน	-1
190	จังหว่ะ	ไม่ดี	-1
191	ตลาด	ซบเซา	-1
192	ตลาด	ซึม	-1
193	ตลาด	ถดถอย	-1
194	ทุน	ลด	-1
195	ธุรกิจ	แย่ลง	-1
196	ธุรกิจ	ถดถอย	-1

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
197	น้ำหนักลงทุน	ลด	-1
198	ปัจจัย	กดดัน	-1
199	ปันผล	ไม่จ่าย	-1
200	ผลการดำเนินงาน	ไม่เด่น	-1
201	ผลการดำเนินงาน	ไม่ดี	-1
202	ผลการดำเนินงาน	ลด	-1
203	ผลตอบแทน	ไม่ดี	-1
204	ผลตอบแทน	ต่ำ	-1
205	ภาษี	เพิ่ม	-1
206	ภาษี	คงที่	-1
207	มุมมอง	Negative	-1
208	มูลค่าพื้นฐาน	ลบ	-1
209	ยอดขาย	ไม่ดี	-1
210	ยอดขาย	ลด	-1
211	ยอดขาย	ลดลง	-1
212	ยอดจอง	ปรับต่ำ	-1
213	รัฐบาล	ไม่อนุมัติ	-1
214	ราคาหุ้น	ลบ	-1
215	รายได้	ไม่ดี	-1
216	รายได้	กดดัน	-1
217	รายได้	ต่ำ	-1
218	รายได้	ลด	-1
219	ลงทุน	ลด	-1
220	สินทรัพย์	ลด	-1
221	หุ้น	ลบ	-1
222	อุตสาหกรรม	แย่ลง	-1
223	อุตสาหกรรม	ไม่เด่น	-1

ตารางที่ 3.9 (ต่อ) คู่คำจากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
224	เงินบาท	คงตัว	0
225	เป้า	คงที่	0
226	เศรษฐกิจ	คงตัว	0
227	แนะ	รอถือ	0
228	โมเมนตั้ม	กลาง	0
229	โอกาส	ถือ	0
230	การเติบโต	คงที่	0
231	การเปลี่ยนแปลง	เป็นกลาง	0
232	การฟื้นตัว	คงที่	0
233	การลงทุน	คงที่	0
234	กำไร	คงตัว	0
235	กำไร	คงที่	0
236	กำไร	คงที่ต่อเนื่อง	0
237	ข่าว	กลาง	0
238	ค่าใช้จ่าย	คงที่	0
239	คาด	เท่าทุน	0
240	คาด	คงที่	0
241	คาดการ	เท่าทุน	0
242	งบ	คงที่	0
243	ดอกเบี้ย	คงที่	0
244	ตลาด	คงตัว	0
245	ทรัพย์สิน	ถือ	0
246	ทุน	คงที่	0
247	ธุรกิจ	คงตัว	0
248	น้ำหนักลงทุน	เท่าเดิม	0
249	น้ำหนักลงทุน	คงที่	0
250	ปรับกำไร	คงที่	0
251	ปัจจัย	ไม่ส่งผล	0
252	ผลกระทบ	เป็นกลาง	0
253	ผลการดำเนินงาน	เท่าทุน	0
254	ผลการดำเนินงาน	คงตัว	0
255	ผลการดำเนินงาน	คงที่	0
256	ผลตอบแทน	คงที่	0
257	มุมมอง	เป็นกลาง	0
258	มุมมอง	กลาง	0
259	มูลค่าพื้นฐาน	เป็นกลาง	0
260	ยอดขาย	คงที่	0
261	ยอดจอง	คงที่	0
262	ราคา	แพง	0
263	ราคา	คงตัว	0
264	ราคา	คงที่	0
265	ราคา	สูง	0
266	ราคาน้ำมัน	คงที่	0
267	ราคาหุ้น	คงตัว	0
268	ราคาหุ้น	คงที่	0
269	รายได้	คงที่	0
270	ลงทุน	คงที่	0
271	สต็อก	คงที่	0
272	สินทรัพย์	คงที่	0
273	หุ้น	เป็นกลาง	0

ตารางที่ 3.9 (ต่อ) คู่คำจากวิธีการสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
274	หุ้น	คงที่	0
275	หุ้น	ซื้อ	0
276	อุตสาหกรรม	กลาง	0
277	อุตสาหกรรม	คงตัว	0

เทคนิคที่ใช้ในการสร้างคู่คำเทคนิคที่สอง คือ การสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน (Automate wordpairs combination – AC) หมายถึง การสร้างคู่คำใหม่ด้วยการใช้คอมพิวเตอร์ตรวจสอบคีย์เวิร์ดที่มีความเหมือนกันในบางส่วนของคำ แล้วนำกลุ่มของคำที่มีขั้วจากคีย์เวิร์ดที่มีความเหมือนกันในบางส่วนมาสลับคู่กัน เพื่อเพิ่มคู่คำที่มีความหมายใกล้เคียงกัน ตัวอย่างของคีย์เวิร์ด ได้แก่ 'ราคาหุ้น' และ 'ราคา' คีย์เวิร์ดทั้งสองมีบางส่วนของคำเหมือนกัน คือ 'ราคา' แสดงตัวอย่างของการสลับคู่ของกลุ่มคำที่มีขั้วจากคู่คำ 'ราคาหุ้น, ขึ้น, 1', 'ราคาหุ้น, บวก, 1' และ 'ราคา, ปรับลด, -1' หลังจากการสลับคู่ของกลุ่มคำที่มีขั้วแล้วจะได้คู่คำใหม่ ได้แก่ 'ราคาหุ้น, ปรับลด, -1', 'ราคา, ขึ้น, 1' และ 'ราคา, บวก, 1'

ตารางที่ 3.10 แสดงคู่คำที่ได้จากการสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน ปริมาณของคู่คำที่ใช้การสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกันนี้มีจำนวนทั้งสิ้น 54 คู่คำ แบ่งออกเป็น

1) คู่คำที่มีขั้วอารมณ์เป็นบวกจำนวน 17 คู่คำ
2) คู่คำที่มีขั้วอารมณ์เป็นลบจำนวน 20 คู่คำ
3) คู่คำที่มีขั้วอารมณ์เป็นกลางจำนวน 17 คู่คำ

ตารางที่ 3.10 คู่คำจากวิธีการสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
278	เงินบาท	จ่าย	1
279	เงินบาท	สูง	1
280	เงินปันผล	อ่อนค่า	1
281	แนะนำ	รอซื้อ	1
282	คาดการ	Positive	1
283	คาดการ	ดี	1
284	ราคา	บวก	1
285	ราคา	แข็งขึ้น	1
286	ราคา	ตอบรับ	1
287	ราคา	ปลอดภัย	1
288	ราคาน้ำมัน	ปรับเพิ่ม	1
289	ราคาน้ำมัน	ปรับขึ้น	1
290	ราคาน้ำมัน	แข็งขึ้น	1
291	ราคาน้ำมัน	ตอบรับ	1
312	ราคาน้ำมัน	ขาด	-1
313	ราคาน้ำมัน	อ่อนลง	-1
292	ราคาน้ำมัน	ปลอดภัย	1
293	ราคาหุ้น	ปรับเพิ่ม	1

ตารางที่ 3.10 (ต่อ) คู่คำจากวิธีการสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน

ลำดับ	คีย์เวิร์ด	คำที่มีขั้ว	เครื่องหมายอารมณ์
294	ราคาหุ้น	แข็งขึ้น	1
295	เงินบาท	ไม่จ่าย	-1
296	เงินบาท	ต่ำ	-1
297	เงินปันผล	แข็งค่า	-1
298	แนะ	รอขาย	-1
299	แนะนำ	รอขาย	-1
300	แนะนำ	หลีกเลี่ยง	-1
301	คาดการ	Negative	-1
302	คาดการ	แย่	-1
303	ราคา	ลบ	-1
304	ราคา	เสี่ยง	-1
305	ราคา	ไม่ตอบรับ	-1
306	ราคา	ขาด	-1
307	ราคา	อ่อนลง	-1
308	ราคาน้ำมัน	ปรับลง	-1
309	ราคาน้ำมัน	ปรับลด	-1
310	ราคาน้ำมัน	เสี่ยง	-1
311	ราคาน้ำมัน	ไม่ตอบรับ	-1
312	ราคาน้ำมัน	ขาด	-1
313	ราคาน้ำมัน	อ่อนลง	-1
314	ราคาหุ้น	ปรับลด	-1
315	เงินปันผล	คงตัว	0
316	แนะนำ	รอถือ	0
317	คาดการ	Neutral	0
318	คาดการ	คงที่	0
319	ราคาน้ำมัน	เป้าหมาย	0
320	ราคาน้ำมัน	แพง	0
321	ราคาน้ำมัน	คงตัว	0
322	ราคาน้ำมัน	ต่ำ	0
323	ราคาน้ำมัน	ถูก	0
324	ราคาน้ำมัน	พักตัว	0
325	ราคาน้ำมัน	สูง	0
326	ราคาหุ้น	เป้าหมาย	0
327	ราคาหุ้น	แพง	0
328	ราคาหุ้น	ต่ำ	0
329	ราคาหุ้น	ถูก	0
330	ราคาหุ้น	พักตัว	0
331	ราคาหุ้น	สูง	0

ตารางที่ 3.11 สรุปปริมาณจำนวนคู่คำในเซ็ต ME, MA และ AC

เซ็ต	คู่คำที่มีอารมณ์บวก	คู่คำที่มีอารมณ์ลบ	คู่คำที่มีอารมณ์เป็นกลาง	รวม	ปริมาณสะสม
ME	68	50	15	133	133
MA	37	53	54	144	277
AC	17	20	17	54	331

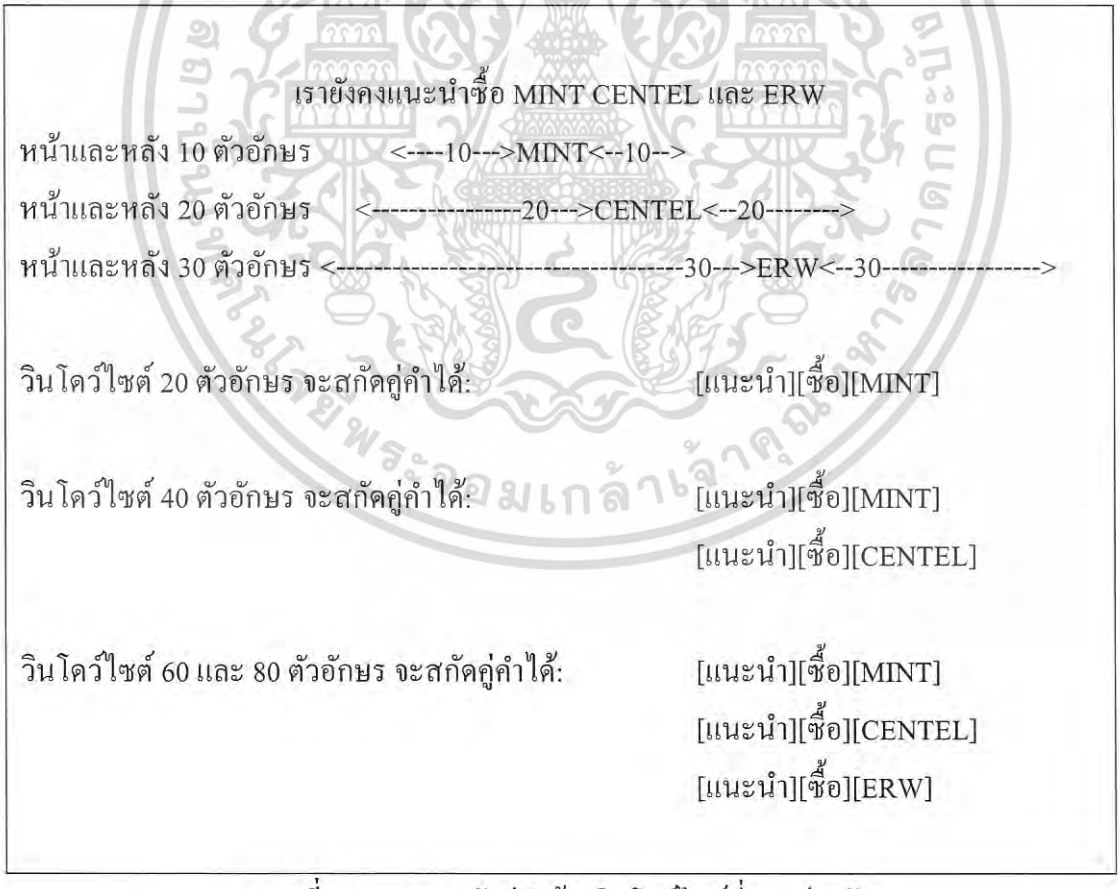
ความถี่สะสมในการเก็บรวบรวมคู่คำจากเซ็ต ME มีจำนวนทั้งสิ้น 133 คู่คำ และในเซ็ต AC เมื่อมีการสร้างคู่คำใหม่เพิ่มเติมขึ้น เมื่อรวบกับคู่คำจากเซ็ต ME เดิม จะมีจำนวนทั้งสิ้น 277 คู่คำ และ 331 คู่คำเมื่อเพิ่มคู่คำจากเซ็ต AC เข้าไป

หลังจากขั้นตอนการเก็บรวบรวมคู่คำในแต่ละรูปแบบเรียบร้อยแล้ว คู่คำในแต่ละรูปแบบจะถูกนำไปสกัดข่าวหุ้น เพื่อใช้วิเคราะห์และเปรียบเทียบประสิทธิภาพของคู่คำแต่ละรูปแบบคือ ME, MA และ AC รวมไปถึงการทดสอบประสิทธิภาพของการใช้คำ (Individual words) ในหัวข้อถัดไป นอกจากนี้ยังมีการทดลองเพิ่มเติมเกี่ยวกับคำที่แสดงเฉพาะอารมณ์ (Polarity words) ซึ่งเป็นเซ็ตย่อย (Subset) ภายในคำในภาคผนวก ก อีกด้วย

3.5 วินโดว์ไซต์ของการสกัดคู่คำ

หลังจากขั้นตอนการเก็บรวบรวมเซ็ตของคู่คำเรียบร้อยแล้ว เซ็ตของคู่คำเหล่านั้นจะถูกนำไปสกัดตัวแทนของข้อมูลออกจากข่าวหุ้น ทั้งในข้อมูลชุดการสอน และข้อมูลชุดทดสอบ โดยมีสมมติฐานดังนี้

คู่คำที่ส่งผลต่อหุ้นตัวนั้น ๆ จะอยู่บริเวณรอบ ๆ สัญลักษณ์หุ้น ดังนั้นจึงได้ทดลองสกัดคู่คำ โดยกำหนดวินโดว์ไซต์โดยใช้สัญลักษณ์หุ้นเป็นกึ่งกลาง โดยทดลองวินโดว์ไซต์ตั้งแต่ 20, 40, 60 และ 80 ตัวอักษร ดังภาพที่ 3.3



ภาพที่ 3.3 ผลการสกัดคู่คำด้วยวินโดว์ไซต์ที่แตกต่างกัน

จากการทดลองเบื้องต้นพบว่าการใช้วินโดว์ไซต์ที่ขนาด 60 ตัวอักษรให้ประสิทธิภาพที่ครอบคลุมบริเวณของข่าวหุ้นอย่างเหมาะสม โดยสามารถสกัดคู่คำได้ตั้งแต่ 1 ถึง 3 คู่คำ

นอกจากนี้เหตุผลที่วินโดว์ไซต์ขนาด 60 ตัวอักษรมีประสิทธิภาพเหมาะสมมากกว่า 80 ตัวอักษร เนื่องมาจากขนาด 80 ตัวอักษรสามารถสกัดคู่คำที่ไม่สอดคล้องกับสัญลักษณ์หุ้นที่สนใจออกมาได้

หลังจากที่สกัดคู่คำออกจากข้อมูลชุดการสอนและข้อมูลชุดทดสอบ เพื่อใช้เป็นคุณลักษณะเรียบร้อยแล้ว จึงนำไปสู่ขั้นตอนการจำแนกข้อมูลในส่วนถัดไป

3.6 การเรียงลำดับตำแหน่งคู่คำและสัญลักษณ์หุ้น

หลังจากขั้นตอนการสกัดคู่คำออกจากข่าวหุ้น เกิดข้อสังเกตขึ้นว่า การเรียงลำดับตำแหน่งคู่คำและสัญลักษณ์หุ้น ส่งผลต่อการจำแนกข้อมูลหรือไม่ จึงได้ทำการกำหนดรูปแบบของสัญลักษณ์หุ้นและคู่คำ ออกได้เป็น 6 รูปแบบ โดยใช้ตำแหน่งของ สัญลักษณ์หุ้น (S), คีย์เวิร์ด (K) และคำที่มีขั้ว (P) เรียงสับเปลี่ยนกัน แสดงในตารางที่ 3.12

ตารางที่ 3.12 รูปแบบตำแหน่งของสัญลักษณ์หุ้น อ้างอิงตามข่าวหุ้นที่ประกาศจากโบรคเกอร์

รูปแบบที่	**สัญลักษณ์หุ้น, คีย์เวิร์ด, คำที่มีขั้ว**
1: S-K-P	**RATCH** **แนะนำ** **ถือ** ราคาเป้าหมาย 64 บาท... สัญลักษณ์หุ้น คีย์เวิร์ด คำที่มีขั้ว
2: S-P-K	**SAMTEL** SAT **กำไร**ตาม**คาด**... สัญลักษณ์หุ้น คำที่มีขั้ว คีย์เวิร์ด
3: K-S-P	รายงานกลุ่ม Small Cap **คาด** **UNIQ** และ TRC มี**โอกาส**ทำ**กำไร**... คีย์เวิร์ด สัญลักษณ์หุ้น คำที่มีขั้ว
4: K-P-S	คง**คำแนะนำ** **ซื้อ** **TP** 24 บาท... คีย์เวิร์ด คำที่มีขั้ว สัญลักษณ์หุ้น
5: P-S-K	เรา**ลด** **ราคา**เป้าหมาย และคำแนะนำ **SIM** และ SAMTEL ลงเหลือถือจาก**ซื้อ**... คำที่มีขั้ว คีย์เวิร์ด สัญลักษณ์หุ้น
6: P-K-S	เดือน สค. พบว่า**ปรับสูงขึ้น** 0.5% และ +4.3% โดย **BBL** BAY TMB KBANK SCB คำที่มีขั้ว สัญลักษณ์หุ้น รายงาน**สินเชื่อ**เติบโต... คีย์เวิร์ด

ในชุดข้อมูลการสอนพบว่า รูปแบบที่พบมากที่สุดของข่าวหุ้นคือ รูปแบบที่สาม (KPS) โดยมีจำนวน 335, 387 และ 387 ข่าว เมื่อทำการสกัดโดยใช้คู่คำเซ็ต ME, MA และ AC ตามลำดับ และรูปแบบที่พบรองลงมาคือ รูปแบบที่ห้า (PSK) โดยมีจำนวน 215, 254 และ 257 ข่าว เมื่อทำการสกัดโดยใช้คู่คำเซ็ต ME, MA และ AC ตามลำดับ ในทางกลับกันรูปแบบที่พบได้น้อยที่สุดคือรูปแบบที่สี่ (KPS) ซึ่งมีจำนวน 47,55 และ 57 ข่าว ตามลำดับ

สำหรับชุดข้อมูลทดสอบพบว่ารูปแบบที่พบได้มากที่สุดและรองลงมาสอดคล้องตามชุดข้อมูลการสอน โดยรูปแบบที่สามมีจำนวน 599, 674 และ 675 ข่าว และรูปแบบที่ห้ามีจำนวน 577, 669 และ 669 ข่าว ตามลำดับ หลังจากที่สกัดรูปแบบครบทั้งหมดแล้วจะนำไปวัดประสิทธิภาพโดยการจำแนกข้อมูลในส่วนถัดไป

บทที่ 4
ผลการทดลองความสอดคล้องและประสิทธิภาพของคู่คำ

ในหัวข้อผลการทดลองความสอดคล้องและประสิทธิภาพจะกล่าวถึง ผลลัพธ์ที่ได้จากการทดลองในแต่ละขั้นตอน เพื่อเป็นการประเมินความสอดคล้อง และความสมเหตุสมผลในขั้นตอนต่างๆ ของงานวิจัย อ้างอิงตามคำถามการวิจัย สำหรับหัวข้อการประเมินผลของการทดลองประกอบไปด้วย

1. ผลการประเมินความสอดคล้องของข่าวหุ้นกับข้อมูลทางสถิติด้วยการจำลองกราฟ (Graph visualization)
2. ผลการประเมินความสอดคล้องแนวโน้มของข่าวหุ้น โดยใช้การวิเคราะห์ทางสถิติด้วยการทดสอบสมมติฐาน (Hypothesis testing)
3. ผลการทดลองประสิทธิภาพโมเดลจำแนกข้อมูลด้วยคู่คำ (Wordpairs classification)
4. ผลการทดลองประสิทธิภาพโมเดลจำแนกข้อมูลด้วยคู่คำ แยกตามตำแหน่งของคู่คำ (Wordpair patterns)
5. ผลการเปรียบเทียบประสิทธิภาพระหว่างการใช้คำ (Individual words) และคู่คำ (Wordpairs) เป็นคุณลักษณะ

4.1 ผลการประเมินความสอดคล้องของข่าวหุ้นกับข้อมูลทางสถิติด้วยการจำลองกราฟ (Graph visualization)

การประเมินความสอดคล้องของข่าวหุ้นกับข้อมูลทางสถิติด้วยการจำลองกราฟ เป็นวิธีการตรวจสอบความสัมพันธ์ของการเปลี่ยนแปลงราคาหุ้น (Price changes) และปริมาณการซื้อขายหุ้น (Trading volumes)

โดยหุ้นตัวอย่างที่นำมาแสดงการวิเคราะห์คือหุ้น CK ในกลุ่มอุตสาหกรรมอสังหาริมทรัพย์และก่อสร้าง (Propcon) เนื่องจากเป็นหุ้นในกลุ่มที่มีการพูดถึงมากที่สุดจากความถี่ในบทที่ 3 ตารางที่ 3.5 และจากงานวิจัยก่อนหน้านี้ (Chattupan and Netisopakul, 2014) ได้ทำการทดสอบค่าสหสัมพันธ์ (Pearson correlation) ระหว่างอารมณ์ข่าวและการเปลี่ยนแปลงราคาหุ้น ดังนั้นก่อนการสร้างกราฟจำลองจึงมีการทดสอบค่าสหสัมพันธ์ก่อนเป็นอันดับแรก

การทดลองพบว่าอารมณ์ข่าวภายในการทดลองนี้และการเปลี่ยนแปลงราคาหุ้น CK มีค่าสหสัมพันธ์เป็นบวกที่ 0.60 ดังนั้นการสร้างกราฟจำลองจึงแสดงตัวอย่างของหุ้น CK ในการวิเคราะห์

การตรวจสอบการเปลี่ยนแปลงของราคาหุ้นและปริมาณการซื้อขายหุ้น ในช่วงระยะเวลา 5 วัน ก่อนหน้าและหลังจากการประกาศข่าว แสดงข้อมูลอ้างอิงประกอบในตารางที่ 4.1 สำหรับแนวโน้มช่วงระยะเวลาที่มีการประกาศข่าวด้านลบจากโบรคเกอร์ ตารางที่ 4.2 สำหรับแนวโน้มช่วงระยะเวลาที่ไม่มีการประกาศข่าวจากโบรคเกอร์ และตารางที่ 4.3 สำหรับแนวโน้มช่วงระยะเวลาที่มีการประกาศข่าวด้านบวกจากโบรคเกอร์

จากตารางที่ 4.1 ข่าวหุ้นที่มีการประกาศจากโบรคเกอร์ ส่งผลอารมณ์ในด้านลบออกมา ดังนั้น การเปลี่ยนแปลงของราคาหุ้นจึงมีค่าเป็นลบ ในขณะเดียวกันสาเหตุที่ปริมาณการซื้อขายมีปริมาณมากขึ้นเนื่องมาจากเป็นการรวมระหว่างปริมาณการซื้อ และปริมาณการขายของหุ้น CK ซึ่งข่าวหุ้นอาจส่งผลให้นักลงทุนเทขายหุ้นออกมาได้

ตารางที่ 4.1 ข่าวหุ้นที่มีการประกาศข่าวด้านลบจากโบรคเกอร์และแนวโน้มช่วงระยะเวลาหน้าและหลัง 5 วันของหุ้น CK

วันที่	**ข่าวหุ้น**			
08/12/14	เราปรับลดน้ำหนักกลุ่มรับเหมา (ITD STEC CK) ลงเหลือ NEUTRAL			
วันที่	**ราคาปิดวันก่อนหน้า**	**ราคาปิดวันนี้**	**การเปลี่ยนแปลงราคา (%)**	**ปริมาณการซื้อขาย ('000 หุ้น)**
03/12/14	27.50	27.50	0	12,116
04/12/14	27.50	27.50	0	17,477
08/12/14	27.50	26.25	-4.55	23,651
09/12/14	26.25	26.00	-0.95	55,561
11/12/14	26.00	25.00	-3.85	34,039

จากตารางที่ 4.2 เป็นวันที่ไม่มีการประกาศข่าวจากโบรคเกอร์ ส่งผลอารมณ์ในด้านเป็นกลางออกมา ดังนั้นการเปลี่ยนแปลงของราคาหุ้นจึงมีการเปลี่ยนแปลงเพียงเล็กน้อย ในขณะเดียวกันปริมาณการซื้อขายมีปริมาณลดลง อาจสรุปได้ว่ามีปริมาณการซื้อขายเป็นกลาง ไม่มีปัจจัยหนุนให้เกิดความตื่นตัวของหุ้น

ตารางที่ 4.2 วันที่ไม่มีการประกาศข่าวหุ้นจากโบรคเกอร์และแนวโน้มช่วงระยะเวลาหน้าและหลัง 5 วันของหุ้น CK

วันที่	**ข่าวหุ้น**			
14/01/15	ไม่มีข่าว			
วันที่	**ราคาปิดวันก่อนหน้า**	**ราคาปิดวันนี้**	**การเปลี่ยนแปลงราคา (%)**	**ปริมาณการซื้อขาย ('000 หุ้น)**
12/01/15	26.00	25.75	-0.96	10,031
13/01/15	25.75	26.00	0.97	9,616
14/01/15	26.00	25.75	-0.96	7,450
15/01/15	25.75	25.75	0	5,655
16/01/15	25.75	25.50	-0.97	5,544

จากตารางที่ 4.3 ข่าวที่มีการประกาศจากโบรคเกอร์ ส่งผลอารมณ์ในด้านบวกออกมา ดังนั้นการเปลี่ยนแปลงของราคาหุ้นจึงมีค่าเป็นบวก แต่ใน 1 วันถัดไปอาจจะยังไม่เกิดการเปลี่ยนแปลงที่เด่นชัดในทันที โดยจะมีการเปลี่ยนแปลงราคาหุ้น และปริมาณการซื้อขายหุ้นที่สามารถสังเกตได้ชัดเจน ใน 2 วันถัดไปหลังจากที่มีการประกาศข่าว

ตารางที่ 4.3 ข่าวหุ้นที่มีการประกาศข่าวด้านบวกจากโบรคเกอร์และแนวโน้มช่วงระยะเวลาหน้าและหลัง 5 วันของหุ้น CK

วันที่	**ข่าวหุ้น**			
23/01/15	CK คาดได้กำไรพิเศษจากการขาย ไซยะบุรี ราว 1.6 พันล้านบาท... แนะนำซื้อ			
วันที่	**ราคาปิดวันก่อนหน้า**	**ราคาปิดวันนี้**	**การเปลี่ยนแปลงราคา (%)**	**ปริมาณการซื้อขาย ('000 หุ้น)**
21/01/15	25.75	26.75	3.88	20,974
22/01/15	26.75	28.50	6.45	142,372
23/01/15	28.50	28.50	0	46,529
24/01/15	28.50	28.50	0	20,760
27/01/15	28.50	29.50	3.51	52,154

การประเมินความสอดคล้องเบื้องต้นของข่าวหุ้น และแนวโน้มช่วงระยะเวลาที่มีการประกาศข่าวของข้อมูลทางสถิติ เป็นเพียงการประเมินเบื้องต้นเท่านั้น โดยในส่วนถัดไปจะเป็นการสร้างกราฟจำลองเพื่อใช้ตรวจสอบการขึ้นลงของราคาหุ้น เมื่อเทียบกับการเปลี่ยนแปลงของตลาด

จากสมการ (7) ถึง (18) ในบทที่ 3 สามารถจำลองกราฟการเปลี่ยนแปลงของราคาหุ้น (Price changes) และปริมาณการซื้อขายหุ้น (Trading volumes) ที่ทำการปรับค่าเรียบร้อยแล้ว (Normalized) โดยใช้ตัวอย่างข้อมูลประมาณ 6 เดือน ระหว่างวันที่ 14 ตุลาคม 2014 ถึง 24 กุมภาพันธ์ 2015 ของหุ้น CK เพื่อจำลองกราฟ แสดงดัง

ภาพที่ 4.1 และ 4.2 สำหรับการเปรียบเทียบการเปลี่ยนแปลงของราคาหุ้น และปริมาณการซื้อขายหุ้นกับดัชนี SET (SET index)

ภาพที่ 4.3 และ 4.4 สำหรับการเปรียบเทียบการเปลี่ยนแปลงของราคาหุ้น และปริมาณการซื้อขายหุ้นกับดัชนีระดับกลุ่มอุตสาหกรรม (Industry index)

ภาพที่ 4.5 และ 4.6 สำหรับการเปรียบเทียบการเปลี่ยนแปลงของราคาหุ้น และปริมาณการซื้อขายหุ้นกับดัชนีระดับกลุ่มหมวดธุรกิจ (Sector index)

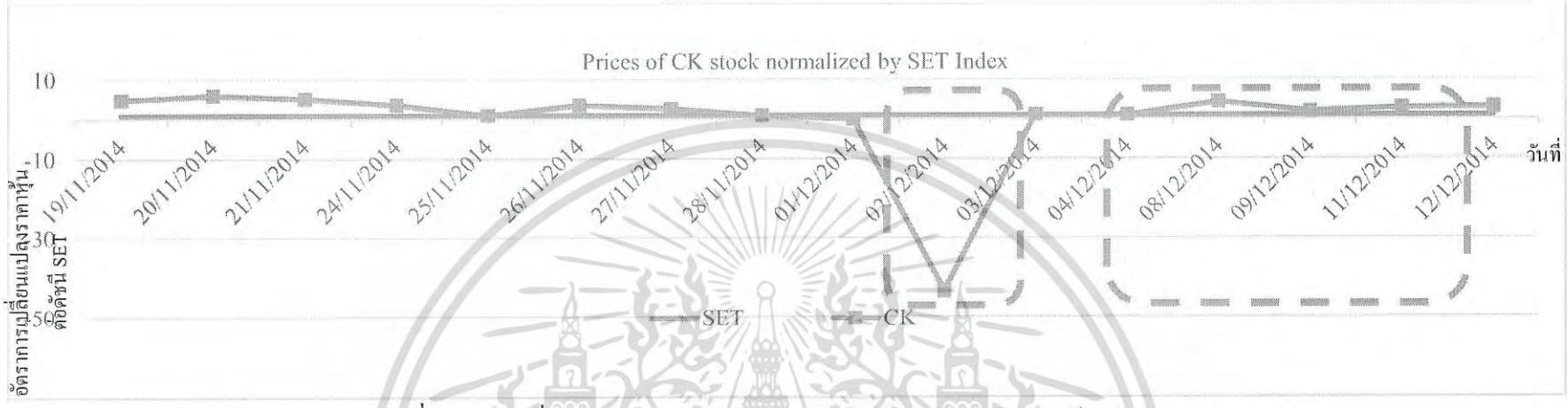
โดยข้อมูลการเปลี่ยนแปลงราคา ปริมาณการซื้อขาย และดัชนี SET, ระดับอุตสาหกรรม และกลุ่มหมวดธุรกิจที่ใช้คำนวณประกอบการสร้างกราฟแสดงดังตารางที่ 4.4 และ 4.5

ตารางที่ 4.4 แสดงข้อมูลการเปลี่ยนแปลงราคา และดัชนี SET, ระดับอุตสาหกรรม และกลุ่มหมวดธุรกิจ

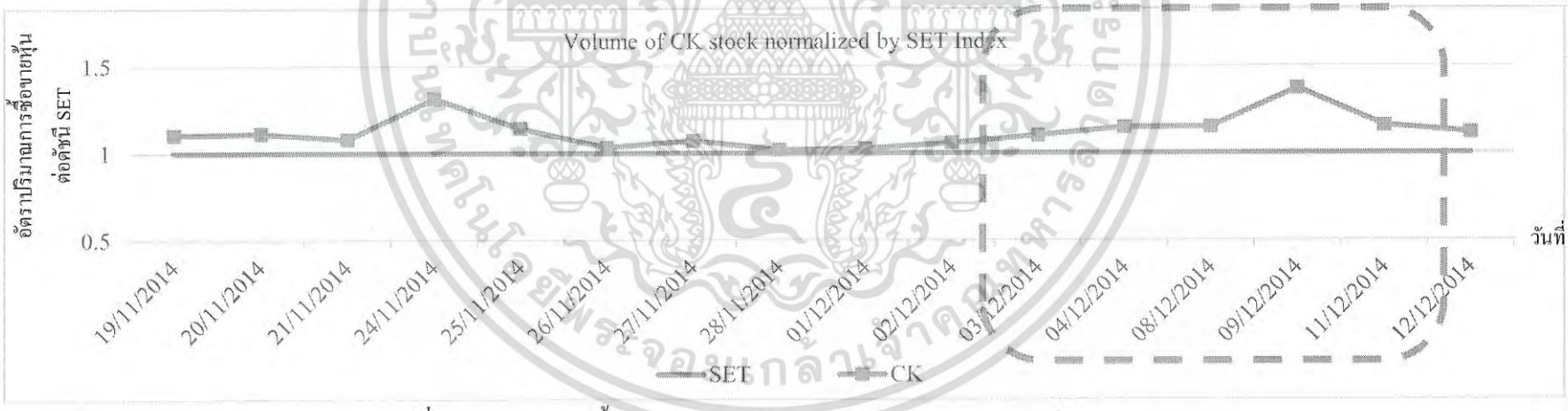
Date	CK+ 0.01	SET+ 0.01	Propcon+ 0.01	Cons+ 0.01	CK/ SET	CK/ Propcon	CK/ Cons
19/11/2014	-0.9	-0.23	-0.78	-0.72	3.91	1.15	1.25
20/11/2014	-2.74	-0.55	-0.57	-0.59	4.98	4.81	4.64
21/11/2014	2.84	0.68	0.16	2.48	4.18	17.75	1.15
24/11/2014	1.84	0.7	1.22	2.25	2.63	1.51	0.82
25/11/2014	0.01	0.43	-0.27	0.08	0.02	-0.04	0.13
26/11/2014	-0.89	-0.35	-0.65	-1.92	2.54	1.37	0.46
27/11/2014	0.92	0.56	0.81	1.54	1.64	1.14	0.60
28/11/2014	0.01	-0.36	0.32	0.82	-0.03	0.03	0.01
01/12/2014	0.01	-0.009	-0.03	0.59	-1.11	-0.33	0.02
02/12/2014	-0.89	0.02	0.02	-0.86	-44.50	-44.50	1.03
03/12/2014	0.01	0.05	0.38	0.29	0.20	0.03	0.03
04/12/2014	0.01	0.21	-0.03	0.13	0.05	-0.33	0.08
08/12/2014	-4.54	-1.38	-1.79	-4.38	3.29	2.54	1.04
09/12/2014	-0.94	-1	-0.5	0.65	0.94	1.88	-1.45
11/12/2014	-3.84	-2.09	-1.65	-3.55	1.84	2.33	1.08
12/12/2014	-1.59	-0.77	-0.38	-2.07	2.06	4.18	0.77

ตารางที่ 4.5 แสดงข้อมูลปริมาณการซื้อขาย และดัชนี SET, ระดับอุตสาหกรรม และกลุ่มหมวดธุรกิจ

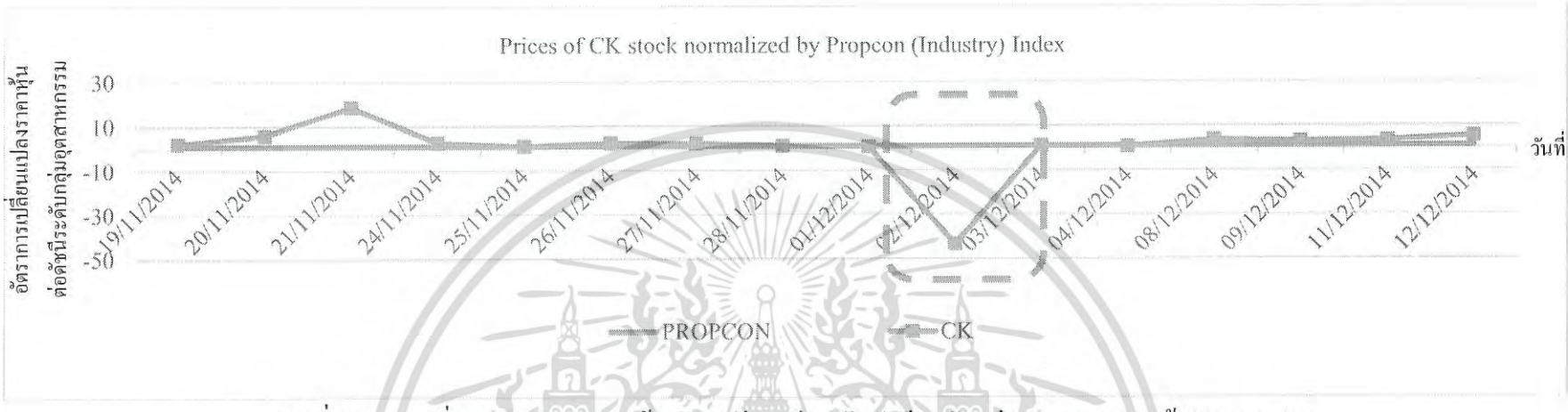
Date	CK/ 1000	SET/ 100M	Propcon/ 100M	Cons/ 100M	CK/ SET	CK/ Propcon	CK/ Cons
19/11/2014	12.81	123.12	48.28	4.64	0.10	0.27	2.76
20/11/2014	14.69	130.80	64.44	5.49	0.11	0.23	2.68
21/11/2014	11.59	144.23	74.46	7.47	0.08	0.16	1.55
24/11/2014	49.72	158.13	80.84	6.32	0.31	0.61	7.87
25/11/2014	19.94	140.73	47.23	4.92	0.14	0.42	4.05
26/11/2014	8.35	255.91	85.17	6.52	0.03	0.10	1.28
27/11/2014	13.30	186.15	91.20	5.00	0.07	0.15	2.66
28/11/2014	4.46	243.57	40.07	3.42	0.02	0.11	1.31
01/12/2014	8.71	358.06	30.98	3.11	0.02	0.28	2.80
02/12/2014	7.22	127.16	32.06	2.51	0.06	0.23	2.87
03/12/2014	12.12	120.97	29.55	1.95	0.10	0.41	6.21
04/12/2014	17.48	118.98	24.79	2.35	0.15	0.70	7.45
08/12/2014	23.65	157.18	30.15	3.17	0.15	0.78	7.46
09/12/2014	55.56	149.93	30.94	3.92	0.37	1.80	14.18
11/12/2014	34.04	216.80	45.09	4.83	0.16	0.75	7.05
12/12/2014	18.66	160.32	53.42	3.35	0.12	0.35	5.56



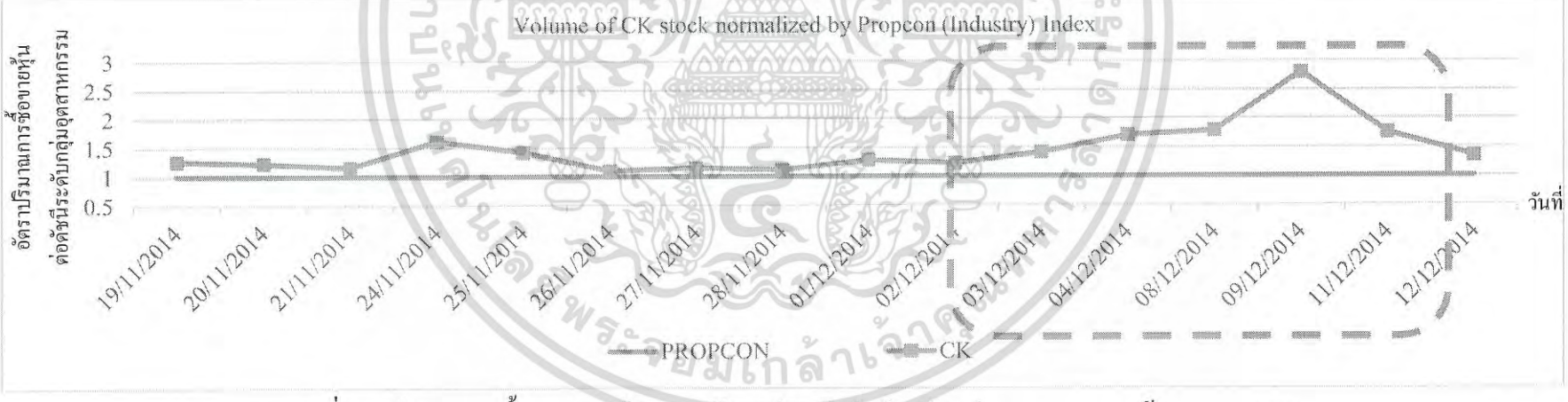
ภาพที่ 4.1 การเปลี่ยนแปลงของราคาหุ้น CK เปรียบเทียบกับดัชนี SET ด้วยสมการ (7)



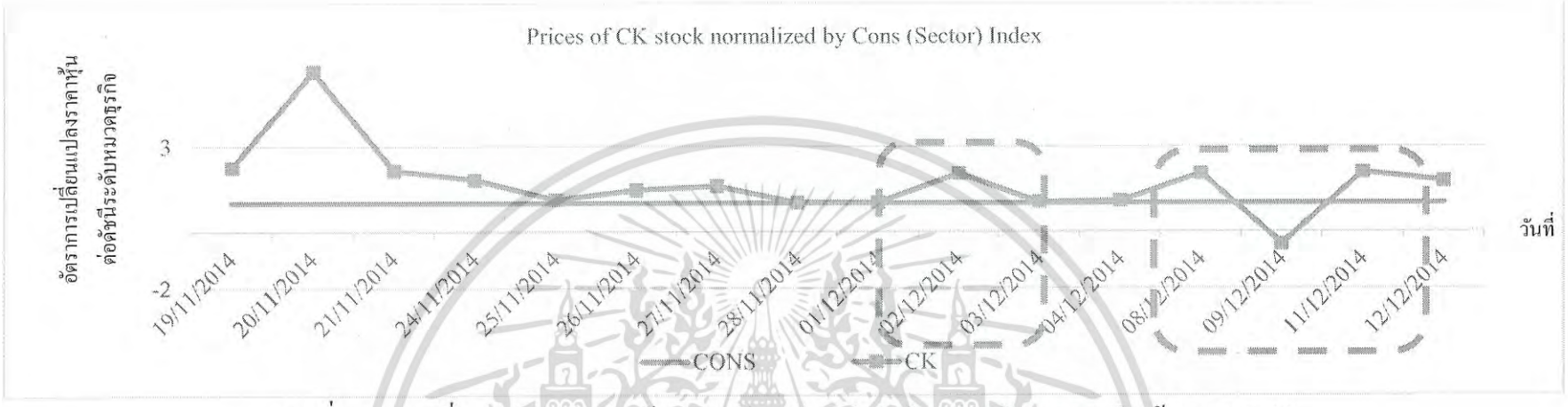
ภาพที่ 4.2 ปริมาณการซื้อขายของหุ้น CK เปรียบเทียบกับดัชนี SET ด้วยสมการ (10)



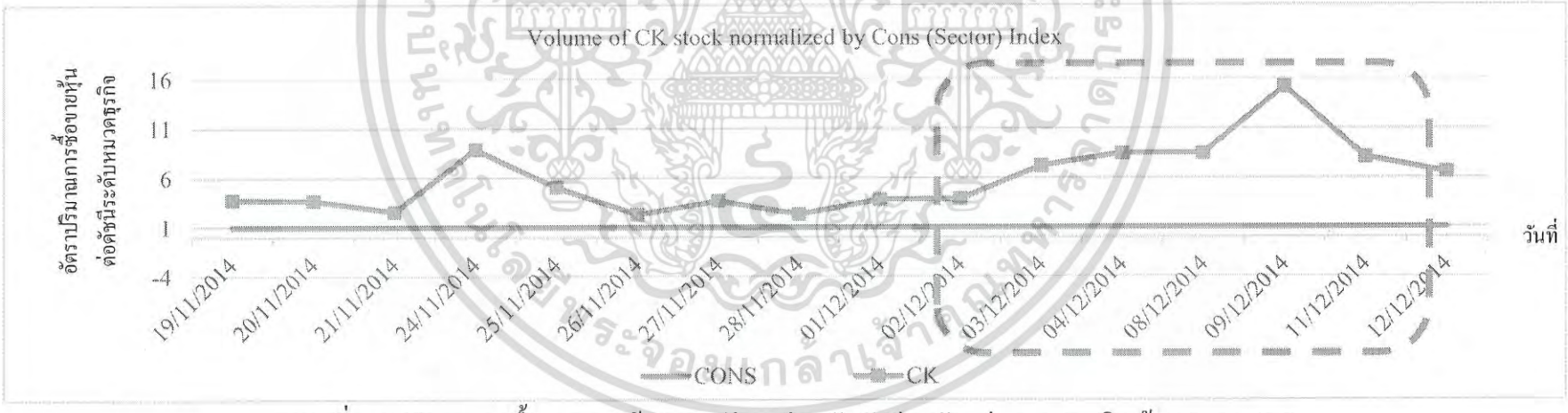
ภาพที่ 4.3 การเปลี่ยนแปลงของราคาหุ้น CK เปรียบเทียบกับดัชนีระดับกลุ่มอุตสาหกรรม ด้วยสมการ (8)



ภาพที่ 4.4 ปริมาณการซื้อขายของหุ้น CK เปรียบเทียบกับดัชนีระดับกลุ่มอุตสาหกรรม ด้วยสมการ (11)



ภาพที่ 4.5 การเปลี่ยนแปลงของราคาหุ้น CK เปรียบเทียบกับดัชนีระดับกลุ่มหมวดธุรกิจ ด้วยสมการ (9)



ภาพที่ 4.6 ปริมาณการซื้อขายของหุ้น CK เปรียบเทียบกับดัชนีระดับกลุ่มหมวดธุรกิจ ด้วยสมการ (12)

ภาพที่ 4.1 ถึง 4.6 จะเห็นได้ว่าเมื่อทำการปรับค่าดัชนี SET, ดัชนีระดับกลุ่มอุตสาหกรรม และดัชนีระดับกลุ่มหมวดธุรกิจให้เป็นค่ามาตรฐาน (เท่ากับ 1 เสมอทั้งเส้น) จะมีการเพิ่มขึ้นหรือลดลงของทั้งกราฟการเปลี่ยนแปลงของราคา และปริมาณการซื้อขายของหุ้น อย่างชัดเจน

ดังนั้นผลการประเมินความสอดคล้องของข่าวหุ้นกับข้อมูลทางสถิติด้วยการใช้สมการสร้างกราฟจำลอง (Graph visualization) จึงเป็นอีกวิธีการหนึ่งที่ใช้ประกอบกับผลจากการประเมินความสอดคล้องเบื้องต้นของข่าวหุ้น

นอกจากนี้ในงานวิจัยได้ทำการประเมินความสอดคล้องแนวโน้มของข่าวหุ้น โดยใช้การวิเคราะห์ทางสถิติด้วยการทดสอบสมมติฐาน (Hypothesis testing) เพิ่มเติมในหัวข้อที่ 4.2

4.2 ผลการประเมินความสอดคล้องแนวโน้มของข่าวหุ้น โดยใช้การวิเคราะห์ทางสถิติด้วยการทดสอบสมมติฐาน (Hypothesis testing)

การทดสอบสมมติฐาน (Hypothesis testing) ด้วยสถิติเครื่องหมาย-อันดับของวิลคอกซัน (Wilcoxon signed-rank test) ระหว่าง แนวโน้มการเปลี่ยนแปลงราคาช่วง 5 วันของวันที่มีข่าวกับการเปลี่ยนแปลงราคาเฉลี่ย 6 เดือนของหุ้น CK แสดงวิธีการคำนวณในตารางที่ 4.6 และสรุปผลลัพธ์ในตารางที่ 4.7

สำหรับแนวโน้มปริมาณการซื้อขายช่วง 5 วันของวันที่มีข่าวกับปริมาณการซื้อขายเฉลี่ย 6 เดือนของหุ้น CK แสดงในตารางที่ 4.8 และสรุปผลลัพธ์ในตารางที่ 4.9

ตารางที่ 4.6 การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มการเปลี่ยนแปลงราคาช่วง 5 วันของวันที่มีข่าวกับการเปลี่ยนแปลงราคาเฉลี่ย 6 เดือนของหุ้น CK

วันที่	ช่วงราคา 5 วันของวันที่มีข่าว					ช่วงการเปลี่ยนแปลงราคา 5 วันของวันที่มีข่าว					การเปลี่ยนแปลงราคาวันที่มีข่าวเฉลี่ย	การเปลี่ยนแปลงราคาเฉลี่ย 6 เดือนของวันที่ไม่มีข่าว
	วันที่มีข่าว-2	วันที่มีข่าว-1	วันที่มีข่าว	วันที่มีข่าว+1	วันที่มีข่าว+2	วันที่มีข่าว-2	วันที่มีข่าว-1	วันที่มีข่าว	วันที่มีข่าว+1	วันที่มีข่าว+2		
12/11/2014	26.25	26.50	26.00	26.50	26.50	0	0.95	-1.89	1.92	0	0.19	1.44
08/12/2014	27.50	27.50	26.25	26.00	25.00	0	0	-4.55	-0.95	-3.85	-1.87	1.44
23/01/2015	26.75	28.50	28.50	28.50	29.50	3.88	6.54	0	0	3.51	2.78	1.44
18/02/2015	29.00	28.00	28.25	28.25	28.00	0	-3.45	0.89	0	-0.88	-0.68	1.44

ตารางที่ 4.6 (ต่อ) การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มการเปลี่ยนแปลงราคาช่วง 5 วันของวันที่มีข่าวกับการเปลี่ยนแปลงราคาเฉลี่ย 6 เดือนของหุ้น CK

วันที่	วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)	ค่าความแตกต่าง (d_i)	เรียงลำดับ d_i	เพิ่มเครื่องหมายของ d_i
12/11/2014	0.19	1.44	-1.25	1	-1
08/12/2014	-1.87	1.44	-3.31	4	-4
23/01/2015	2.78	1.44	1.34	2	2
18/02/2015	-0.68	1.44	-2.12	3	-3

จากตารางที่ 4.6 สามารถสรุปค่า $T^+ = 2$

$T^- = -1-4-3 = -8$

$T = 2$ (ค่าที่น้อยที่สุดระหว่าง T^+ และ T^- โดยไม่สนใจเครื่องหมาย)

คำนวณค่าวิกฤต μ จากสมการที่ (19) $\mu = n(n+1)/2$ โดย $n = 4$

$= 10$

ตารางที่ 4.7 ผลการทดสอบสมมติฐานความแตกต่างของการเปลี่ยนแปลงราคาหุ้น CK ในวันที่มีข่าว

	ค่าจากการคำนวณของวิลคอกซัน
T^+	2
T^-	-8
T	2
μ	10

จากสมมติฐาน 1 ที่กล่าวถึงในบทที่ 3

$$H_0: x_i = \mu_0 \quad H_1: x_i \neq \mu_0$$

โดยที่สมมติฐานการเปลี่ยนแปลงราคาแย้ง $H_1: x_i \neq \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤต

2 น้อยกว่า 10 ดังนั้น $x_i \neq \mu_0$ การเปลี่ยนแปลงของราคาช่วงเวลา 5 วัน (x_i) จึงแตกต่างจากค่าเฉลี่ยการเปลี่ยนแปลงของราคาหุ้น เป็นเวลา 6 เดือน (μ_0)

ตารางที่ 4.8 การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มปริมาณการซื้อขายช่วง 5 วันของวันที่มีข่าวกับปริมาณการซื้อขายเฉลี่ย 6 เดือนของหุ้น CK

วันที่	ช่วงราคา 5 วันของวันที่มีข่าว					ปริมาณซื้อขายวันที่มีข่าวเฉลี่ย	ปริมาณซื้อขายเฉลี่ย 6 เดือนของวันที่ไม่มีข่าว
	วันที่มีข่าว-2	วันที่มีข่าว-1	วันที่มีข่าว	วันที่มีข่าว+1	วันที่มีข่าว+2		
12/11/2014	12,104,145	6,948,882	10,419,182	9,992,154	7,362,298	9,365,332	16,654,994
08/12/2014	12,116,896	17,477,093	23,651,232	55,561,056	34,039,695	28,569,194	16,654,994
23/01/2015	20,974,109	142,372,681	46,529,697	20,760,530	52,154,529	56,558,309	16,654,994
18/02/2015	11,505,810	12,299,395	11,789,915	4,297,686	7,485,436	9,475,648	16,654,994

ตารางที่ 4.8 (ต่อ) การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มปริมาณการซื้อขายช่วง 5 วันของวันที่มีข่าวกับปริมาณการซื้อขายเฉลี่ย 6 เดือนของหุ้น CK

วันที่	วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)	ค่าความแตกต่าง (d_i)	เรียงลำดับ d_i	เพิ่มเครื่องหมายของ d_i
12/11/2014	9,365,332	16,654,994	-7,289,662	2	-2
08/12/2014	28,569,194	16,654,994	11,914,200	3	3
23/01/2015	56,558,309	16,654,994	39,903,315	4	4
18/02/2015	9,475,648	16,654,994	-7,179,346	1	-1

จากตารางที่ 4.8 สามารถสรุปค่า $T^+ = 3+4 = 7$

$T^- = -2-1 = -3$

$T = 3$ (ค่าที่ T^- โดยไม่สนใจเครื่องหมาย)

คำนวณค่าวิกฤต μ จากสมการที่ (19) $\mu = n(n+1)/2$ โดย $n = 4$

$= 10$

ตารางที่ 4.9 ผลการทดสอบสมมติฐานความแตกต่างของปริมาณการซื้อขายหุ้น CK ในวันที่มีข่าว

	ค่าจากการคำนวณของวิลคอกซัน
T^+	7
T^-	-3
T	3
μ	10

จากสมมติฐาน 2 ที่กล่าวถึงในบทที่ 3

$$H_0: x_i \leq \mu_0 \quad H_1: x_i > \mu_0$$

โดยที่สมมติฐานการเปลี่ยนแปลงราคาแย้ง $H_1: x_i > \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤต

3 น้อยกว่า 10 ดังนั้น $H_1: x_i > \mu_0$ ปริมาณการซื้อขายช่วงเวลา 5 วัน (x_i) จึงแตกต่างจากค่าเฉลี่ยปริมาณการซื้อขาย เป็นเวลา 6 เดือน (μ_0)

จากตารางที่ 4.7 และ 4.9 สามารถสรุปผลการทดสอบสมมติฐานได้ว่าแนวโน้มช่วงเวลาที่มีการประกาศข่าวจากโบรคเกอร์ การเปลี่ยนแปลงราคาและปริมาณการซื้อขายหุ้น มีความแตกต่างอย่างมีนัยสำคัญจากค่าเฉลี่ยหรือค่าปกติของการเปลี่ยนแปลงราคาและปริมาณการซื้อขาย

จากการทดสอบในหัวข้อที่ 4.1 และ 4.2 สามารถสรุปคำถามวิจัยข้อแรกได้ว่า ข่าวหุ้นที่ประกาศจากโบรคเกอร์ หรือมีการกล่าวถึงหุ้นตัวนั้นๆ ภายในข่าวจะส่งผลต่อการเปลี่ยนแปลงราคา และปริมาณการซื้อขายของหุ้นตัวนั้นๆ อย่างมีนัยสำคัญ

นอกจากนี้ยังได้มีการทดสอบสมมติฐานการเปลี่ยนแปลงราคากับกลุ่มตัวอย่างที่มีขนาดใหญ่ขึ้นจากข่าวหุ้นทั้งหมดด้วยการสุ่ม แสดงในตารางที่ 4.10

ตารางที่ 4.10 การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มการเปลี่ยนแปลงราคาช่วง 5 วันของวันที่มีข่าวกับการเปลี่ยนแปลงราคาเฉลี่ย 6 เดือนของหุ้นกลุ่มตัวอย่างโดยการสุ่มจากข่าวทั้งหมด

n	สัญลักษณ์หุ้น	วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)	ค่าความแตกต่าง (d_i)	เรียงลำดับ d_i	เพิ่มเครื่องหมายของ d_i
n=1	BECL	2.40	1.20	1.20	384	384
n=2	CPALL	1.29	0.96	0.33	145	145
n=3	DEMCO	1.23	2.14	-0.91	323	-323
n=668	…	…	…	…	…	…

จากตารางที่ 4.10 สามารถสรุปค่า $T^+ = 60,559$

$T^- = -162,481$

$T = 60,559$ (ค่าที่น้อยที่สุดระหว่าง T^+ และ T^- โดยไม่สนใจเครื่องหมาย)

คำนวณค่าวิกฤต μ จากสมการที่ (19) $\mu = n(n+1)/4$ โดย $n = 668$ (n มีขนาดมากกว่า 50)

$= 111,723$

ตารางที่ 4.11 ผลการทดสอบสมมติฐานความแตกต่างของการเปลี่ยนแปลงราคาหุ้นจากกลุ่มตัวอย่างในวันที่มีข่าว

	ค่าจากการคำนวณของวิลคอกซัน
T^+	60,559
T^-	-162,481
T	60,559
μ	111,723

จากสมมติฐาน 1 ที่กล่าวถึงในบทที่ 3

$$H_0: x_i = \mu_0 \quad H_1: x_i \neq \mu_0$$

โดยที่สมมติฐานการเปลี่ยนแปลงราคาแย้ง $H_1: x_i \neq \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤต

60,559 น้อยกว่า 111,723 ดังนั้น $x_i \neq \mu_0$ การเปลี่ยนแปลงของราคาช่วงเวลา 5 วัน (x_i) จึงแตกต่างจากค่าเฉลี่ยการเปลี่ยนแปลงของราคาหุ้น เป็นเวลา 6 เดือน (μ_0)

นอกจากนี้ยังได้มีการทดสอบสมมติฐานปริมาณการซื้อขายเพิ่มเติมกับกลุ่มตัวอย่างที่มีขนาดใหญ่ขึ้น จากข่าวหุ้นทั้งหมดด้วยการสุ่ม แสดงในตารางที่ 4.12

ตารางที่ 4.12 การคำนวณสถิติเครื่องหมาย-อันดับของวิลคอกซัน ระหว่าง แนวโน้มปริมาณการซื้อขายช่วง 5 วันของวันที่มีข่าวกับปริมาณการซื้อขายเฉลี่ย 6 เดือนของหุ้นกลุ่มตัวอย่างโดยการสุ่มจากข่าวทั้งหมด

n	สัญลักษณ์หุ้น	วันที่มีข่าวเฉลี่ย (x_i)	วันที่ไม่มีข่าว (μ_0)	ค่าความแตกต่าง (d_i)	เรียงลำดับ d_i	เพิ่มเครื่องหมายของ d_i
n=1	QH	48,280,029.6	34,659,245.78	13,620,783.82	460	460
n=2	KTB	39,249,653.2	37,057,475.34	2,192,177.86	313	313
n=3	CPALL	10,314,981.8	11,477,883.41	-1,162,901.61	244	244
n=559	...	...	...	...	...	...

จากตารางที่ 4.12 สามารถสรุปค่า $T^+ = 94,881$

$T^- = -84,616$

$T = 84,616$ (ค่าที่ T^- โดยไม่สนใจเครื่องหมาย)

คำนวณค่าวิกฤต μ จากสมการที่ (19) $\mu = n(n+1)/4$ โดย $n = 559$

$= 89,850$

ตารางที่ 4.13 ผลการทดสอบสมมติฐานความแตกต่างของปริมาณการซื้อขายหุ้นจากกลุ่มตัวอย่างในวันที่มีข่าว

	ค่าจากการคำนวณของวิลคอกซัน
T^+	94,881
T^-	-84,616
T	84,616
μ	89,850

จากสมมติฐาน 2 ที่กล่าวถึงในบทที่ 3

$$H_0: x_i \leq \mu_0 \quad H_1: x_i > \mu_0$$

โดยที่สมมติฐานการเปลี่ยนแปลงราคาแย้ง $H_1: x_i > \mu_0$ จะตัดสินใจปฏิเสธ H_0 ที่ระดับนัยสำคัญ ถ้าค่า T ที่คำนวณได้น้อยกว่าหรือเท่ากับค่าวิกฤต

84,616 น้อยกว่า 89,850 ดังนั้น $H_1: x_i > \mu_0$ ปริมาณการซื้อขายช่วงเวลา 5 วัน (x_i) จึงแตกต่างจากค่าเฉลี่ยปริมาณการซื้อขาย เป็นเวลา 6 เดือน (μ_0)

4.3 ผลการทดลองประสิทธิภาพโมเดลจำแนกข้อมูลด้วยคู่คำ (Wordpairs classification)

โมเดลการจำแนกข้อมูลด้วยคู่คำ สร้างขึ้นจากคู่คำทั้ง 3 รูปแบบประกอบด้วย การสกัดและเก็บรวบรวมคู่คำด้วยมือ (ME), การสร้างคู่คำใหม่จากคู่คำดั้งเดิมด้วยมือ (MA) และ การสร้างคู่คำใหม่อัตโนมัติด้วยบางส่วนของคีย์เวิร์ดที่เหมือนกัน (AC)

โดยใช้วิธีการสร้างต้นไม้ตัดสินใจ (Decision tree) และใช้ฟังก์ชัน SVM (Support vector machine) เพื่อสร้างโมเดลการจำแนกข้อมูลขึ้นมา

ผลของความถูกต้องจากโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM ของคู่คำเซ็ต ME แสดงในตารางที่ 4.14

ผลของความถูกต้องจากโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM ของคู่คำเซ็ต MA แสดงในตารางที่ 4.15

ผลของความถูกต้องจากโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM ของคู่คำเซ็ต AC แสดงในตารางที่ 4.16

ตารางที่ 4.14 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM โดยใช้คู่คำเซ็ต ME

	Decision tree			SVM		
Class	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
-1	0.675	0.042	0.079	0.632	0.036	0.068
0	0.732	0.094	0.167	0.568	0.066	0.118
1	0.760	0.996	0.862	0.758	0.995	0.860
Average	0.741	0.758	0.669	0.723	0.755	0.663

ตารางที่ 4.15 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM โดยใช้คู่คำเซ็ต MA

	Decision tree			SVM		
Class	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
-1	0.600	0.042	0.079	0.574	0.040	0.075
0	0.640	0.100	0.173	0.579	0.069	0.123
1	0.761	0.994	0.862	0.758	0.993	0.860
Average	0.722	0.757	0.669	0.712	0.755	0.665

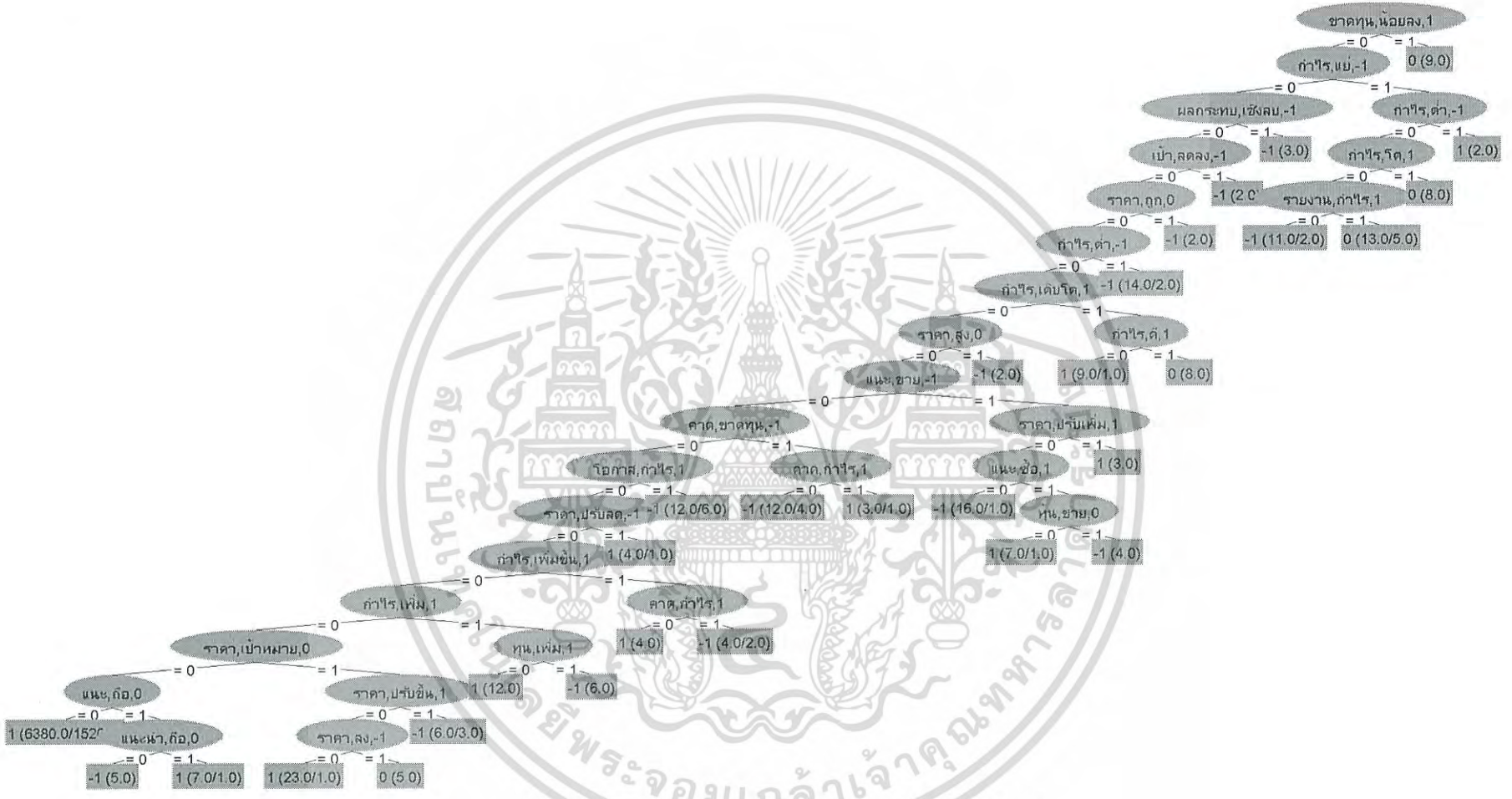
ตารางที่ 4.16 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM โดยใช้คู่คำเซ็ต AC

	Decision tree			SVM		
Class	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
-1	0.591	0.041	0.077	0.576	0.039	0.074
0	0.653	0.100	0.174	0.564	0.069	0.123
1	0.760	0.994	0.862	0.758	0.993	0.860
Average	0.721	0.757	0.669	0.712	0.754	0.664

จากผลการทดลองในตารางที่ 4.14 ถึง 4.16 พบว่าประสิทธิภาพของการจำแนกข้อมูล แต่ละเซ็ตของคู่คำ ให้ประสิทธิภาพที่ค่อนข้างใกล้เคียงกัน ดังนั้นผู้วิจัยจึงได้ทำการทดลองต่อเพื่อหาความแตกต่างที่ชัดเจนมากยิ่งขึ้น ภายในหัวข้อที่ 4.4

ภาพที่ 4.7, 4.8 และ 4.9 แสดงภาพจำลองต้นไม้ตัดสินใจ ด้วยการใช้คู่คำเซ็ต ME, MA และ AC ตามลำดับ

ภาพที่ 4.7 ภาพจำลองต้นไม้ตัดสินใจ ด้วยการใช้คู่คำเซ็ต ME



ภาพที่ 4.8 ภาพจำลองต้นไม้ตัดสินใจ ด้วยการใช้คู่คำเซ็ต MA

ภาพที่ 4.9 ภาพจำลองต้นไม้ตัดสินใจ ด้วยการใช้คู่คำเซ็ต AC

4.4 ผลการทดลองประสิทธิภาพโมเดลจำแนกข้อมูลด้วยคู่คำ แยกตามตำแหน่งของคู่คำ (Wordpair patterns)

จากผลการทดลองในตารางที่ 4.14 ถึง 4.16 ก่อนหน้านี้ ผู้วิจัยได้ทำการทดลองที่ 2 โดยทำการทดลองแยกตามตำแหน่งของสัญลักษณ์หุ้น, คีย์เวิร์ด และ คำที่มีขั้ว ออกเป็น 6 รูปแบบ ได้แก่ #1:S-K-P, #2:S-P-K, #3:K-S-P, #4:K-P-S, #5:P-S-K และ #6:P-K-S โดยที่ S หมายถึง สัญลักษณ์หุ้น, K หมายถึง คีย์เวิร์ด และ P หมายถึง คำที่มีขั้ว แสดงผลการทดลองประสิทธิภาพด้วยโมเดลต้นไม้ตัดสินใจในตารางที่ 4.17 และ ผลการทดลองประสิทธิภาพด้วยฟังก์ชัน SVM ในตารางที่ 4.18

ตารางที่ 4.17 ผลการทดลองประสิทธิภาพด้วยโมเดลต้นไม้ตัดสินใจ แยกตามตำแหน่งของคู่คำ

Wordpair features		ME			MA			AC		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Decision tree	1:SKP	0.68	0.69	0.66	0.77	0.80	0.77	0.77	0.80	0.77
	2:SPK	0.93	0.92	0.91	0.90	0.91	0.90	0.90	0.91	0.90
	3:KSP	0.63	0.64	0.62	0.58	0.61	0.58	0.58	0.61	0.58
	4:KPS	0.75	0.76	0.74	0.74	0.75	0.72	0.74	0.75	0.72
	5:PSK	0.65	0.74	0.69	0.63	0.69	0.66	0.68	0.71	0.69
	6:PKS	0.75	0.76	0.75	0.79	0.80	0.79	0.76	0.77	0.76
	Avg.	0.73	0.75	0.73	0.74	0.76	0.74	0.74	0.76	0.74

ตารางที่ 4.18 ผลการทดลองประสิทธิภาพด้วยฟังก์ชัน SVM แยกตามตำแหน่งของคู่คำ

Wordpair features		ME			MA			AC		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM	1:SKP	0.70	0.71	0.70	0.81	0.83	0.81	0.79	0.81	0.80
	2:SPK	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
	3:KSP	0.70	0.71	0.70	0.69	0.70	0.70	0.69	0.70	0.69
	4:KPS	0.70	0.71	0.70	0.71	0.73	0.71	0.71	0.73	0.72
	5:PSK	0.77	0.78	0.77	0.72	0.72	0.72	0.75	0.77	0.76
	6:PKS	0.69	0.71	0.69	0.74	0.76	0.75	0.73	0.75	0.74
	Avg.	0.75	0.76	0.75	0.77	0.78	0.77	0.77	0.79	0.78

จากผลการทดลองในตารางที่ 4.17 และ 4.18 พบว่าประสิทธิภาพของฟังก์ชัน SVM ให้ผลดีกว่าโมเดลต้นไม้ตัดสินใจทั้ง 3 เซ็ตของคู่คำ เมื่อเปรียบเทียบประสิทธิภาพเฉพาะโมเดลฟังก์ชัน SVM พบว่า เซ็ตของคู่คำ AC ให้ผลที่ดีที่สุด และรูปแบบตำแหน่งของคู่คำที่ให้ผลดีที่สุด จะอยู่ในรูปแบบ S, P, K หรือ สัญลักษณ์หุ้น ตามด้วย คำที่มีขั้ว และ คีย์เวิร์ด เป็นลำดับสุดท้าย ส่วนรูปแบบที่ให้ผลดีรองลงมาคือ S, K, P หรือ สัญลักษณ์หุ้น ตามด้วย คีย์เวิร์ด และ คำที่มีขั้ว

4.5 ผลการเปรียบเทียบประสิทธิภาพระหว่างการใช้คำ (Individual words) และคู่คำ (Wordpairs) เป็นคุณลักษณะ

จากประสิทธิภาพของการใช้คู่คำในหัวข้อที่ 4.3 เมื่อนำมาเปรียบเทียบประสิทธิภาพกับการใช้คำเป็นคุณลักษณะ แสดงในตารางที่ 4.19 สามารถสรุปผลได้ว่า ประสิทธิภาพของคุณลักษณะทั้งสองมีความใกล้เคียงกัน แต่เมื่อแยกรูปแบบการเขียนพบว่า ประสิทธิภาพของคู่คำเซ็ต MA และ AC จะสูงกว่าเล็กน้อย ซึ่งในแง่ของขนาดคุณลักษณะ การใช้คู่คำจะมีปริมาณคุณลักษณะต่อข่าวน้อยกว่าการใช้คำ เนื่องจากคู่คำสามารถสกัดรอบสัญลักษณ์หุ้นได้ตั้งแต่ 1 ถึง 3 คู่คำ แต่การใช้คำเป็นคุณลักษณะจะใช้ทุกคำรอบๆ บริบทของสัญลักษณ์หุ้น ซึ่งจะมีมากกว่าคู่คำ

ตารางที่ 4.19 ผลการทดลองประสิทธิภาพระหว่างการใช้คำ และคู่คำเป็นคุณลักษณะ

	133 คำ			277 คู่คำ			331 คู่คำ		
Individual words	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	0.743	0.774	0.728	0.744	0.777	0.738	0.747	0.778	0.739
	ME (133 คู่คำ)			MA (277 คู่คำ)			AC (331 คู่คำ)		
Wordpairs	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	0.75	0.76	0.75	0.77	0.78	0.77	0.77	0.79	0.78

นอกจากนี้ยังได้มีการทดลองประสิทธิภาพของการใช้คู่คำโดยไม่มีวินโดว์ไซต์ เพื่อแสดงให้เห็นว่าประสิทธิภาพของคู่คำสามารถนำไปประยุกต์ใช้กับโดเมนอื่นๆ ที่ไม่มีความจำเป็นต้องจำกัดข้อความด้วยวินโดว์ไซต์ แสดงในตารางที่ 4.20

ตารางที่ 4.20 ผลการทดลองประสิทธิภาพระหว่างการใช้คู่คำเป็นคุณลักษณะโดยไม่จำกัดวินโดว์ไซต์ และใช้กับขนาดข้อมูลที่แตกต่างกัน

	Wordpairs (AC)		
Input per class	**Precision**	**Recall**	**F-measure**
100	0.604	0.603	0.600
500	0.853	0.843	0.770
1,000	0.880	0.876	0.877
1,500	0.868	0.852	0.853
2,000	0.852	0.817	0.818
2,500	0.846	0.801	0.799
3,000	0.833	0.787	0.782
3,500	0.836	0.787	0.782
4,000	0.835	0.780	0.774
4,500	0.842	0.793	0.789
5,000	0.839	0.786	0.781

จากตารางที่ 4.20 จะเห็นได้ว่าเมื่อไม่จำกัดวินโดว์ไซต์ ค่าความถูกต้องของการจำแนกข้อมูลจะเพิ่มสูงขึ้น จึงเหมาะสำหรับการนำไปประยุกต์ใช้กับโดเมนอื่นๆ ที่ไม่มีความจำเป็นในการกำหนดขนาดวินโดว์ไซต์

สำหรับผลการทดลองอื่นๆ เพิ่มเติม ตัวอย่างเช่น ความแตกต่างของปริมาณข้อมูลในการสร้างโมเดลจำแนกข้อมูลชุดการสอน แสดงประกอบเพิ่มเติมในส่วนของ ภาคผนวก ก

4.6 การวิเคราะห์ข้อผิดพลาดจากโมเดลการเรียนรู้ชุดการสอน ME, MA และ AC

หลังจากการสร้างโมเดลการจำแนกข้อมูลด้วยคู่คำเซ็ต ME, MA และ AC เรียบร้อยแล้ว จึงมีการตรวจสอบข้อผิดพลาดของโมเดลการจำแนกข้อมูล จากขั้นตอนการสร้างโดยการใช้ค่าความถูกต้องจากการตรวจสอบไขว้ (K-folds validation) ในแต่ละครั้งทั้งหมด 10 ครั้ง แสดงค่าความถูกต้องและความผิดพลาดในตารางที่ 4.21 สำหรับโมเดลการเรียนรู้ที่สร้างขึ้นด้วยการใช้ต้นไม้ตัดสินใจ และตารางที่ 4.22 สำหรับโมเดลการเรียนรู้ที่สร้างขึ้นด้วยฟังก์ชัน SVM

ตารางที่ 4.21 ค่าความถูกต้องจากการตรวจสอบไขว้ของโมเดลการเรียนรู้ต้นไม้ตัดสินใจ

เซ็ตคู่คำ	รอบที่ (Fold)	ปริมาณที่ทดสอบ (ข่าว)	ปริมาณที่ถูกต้อง (ข่าว)	ปริมาณที่ผิดพลาด (ข่าว)	ความถูกต้อง (%)	ความผิดพลาด (%)
ME	1	660	497	163	75.30	24.70
	2	660	504	156	76.36	23.64
	3	660	503	157	76.21	23.79
	4	660	499	161	75.61	24.39
	5	660	501	159	75.91	24.09
	6	660	503	157	76.21	23.79
	7	659	497	162	75.42	24.58
	8	659	497	162	75.42	24.58
	9	659	499	160	75.72	24.28
	10	659	503	156	76.33	23.67
MA	1	660	498	162	75.45	24.55
	2	660	504	156	76.36	23.64
	3	660	502	158	76.06	23.94
	4	660	494	166	74.85	25.15
	5	660	501	159	75.91	24.09
	6	660	501	159	75.91	24.09
	7	659	497	162	75.42	24.58
	8	659	497	162	75.42	24.58
	9	659	499	160	75.72	24.28
	10	659	502	157	76.18	23.82

ตารางที่ 4.21 (ต่อ) ค่าความถูกต้องจากการตรวจสอบไขว้ของโมเดลการเรียนรู้ต้นไม้ตัดสินใจ

เซ็ตคู่คำ	รอบที่ (Fold)	ปริมาณที่ทดสอบ (ข่าว)	ปริมาณที่ถูกต้อง (ข่าว)	ปริมาณที่ผิดพลาด (ข่าว)	ความถูกต้อง (%)	ความผิดพลาด (%)
AC	1	660	498	162	75.45	24.55
	2	660	503	157	76.21	23.79
	3	660	502	158	76.06	23.94
	4	660	494	166	74.85	25.15
	5	660	501	159	75.91	24.09
	6	660	501	159	75.91	24.09
	7	659	497	162	75.42	24.58
	8	659	497	162	75.42	24.58
	9	659	499	160	75.72	24.28
	10	659	502	157	76.18	23.82

จากตารางที่ 4.21 จะพบว่าเซ็ตคู่คำ MA และ AC มีปริมาณความผิดพลาดของการจำแนกข้อมูลมาที่สุดในรอบที่ 4 (ข่าวลำดับที่ 1,981 – 2,640) ที่จำนวน 166 ข่าว คิดเป็น 25.15%

ในการตรวจสอบความผิดพลาดจากข่าวโดยมนุษย์พบว่ามีข่าวหุ้นบางส่วน (ช่วงที่ 1,981 – 2,640) ที่ไม่สามารถสกัดคู่คำออกมาได้ เนื่องจากโครงสร้างของข่าวไม่อยู่ในรูปแบบของ คีย์เวิร์ด และคำที่มีขั้ว แสดงดังภาพที่ 4.10 ส่งผลให้ไม่มีคุณลักษณะสำหรับข่าวนั้น

> CPF TUF แจงไร้ผลกระทบ:หลัง EU แจกใบเหลืองเป็นทางการเตือนประมงไทยใช้แรงงานผิด กม.ให้เวลาแก้ไข 6 เดือน เล็งมาตรการค่ำบาตรไทย

ภาพที่ 4.10 ประโยคที่ไม่สามารถสกัดคู่คำได้

ซึ่งข่าวหุ้นดังกล่าวมีอารมณ์ของข่าวในด้านลบ แต่เมื่อทำการจำแนกข้อมูลด้วยโมเดลชุดการสอนที่สร้างขึ้น ผลของการจำแนกข้อมูลให้อารมณ์ของข่าวในด้านบวก ดังนั้นข่าวหุ้นที่ไม่พบคู่คำจึงถูกจำแนกเป็นบวกทั้งหมด

ตารางที่ 4.22 ค่าความถูกต้องจากการตรวจสอบไขว้ของโมเดลฟังก์ชัน SVM

เซ็ตคู่คำ	รอบที่ (Fold)	ปริมาณที่ทดสอบ (ข่าว)	ปริมาณที่ถูกต้อง (ข่าว)	ปริมาณที่ผิดพลาด (ข่าว)	ความถูกต้อง (%)	ความผิดพลาด (%)
ME	1	660	494	166	74.85	25.15
	2	660	494	166	74.85	25.15
	3	660	494	166	74.85	25.15
	4	660	494	166	74.85	25.15
	5	660	494	166	74.85	25.15
	6	660	493	167	74.70	25.30
	7	659	493	166	74.81	25.19
	8	659	493	166	74.81	25.19
	9	659	493	166	74.81	25.19
	10	659	493	166	74.81	25.19
MA	1	660	494	166	74.85	25.15
	2	660	494	166	74.85	25.15
	3	660	494	166	74.85	25.15
	4	660	494	166	74.85	25.15
	5	660	494	166	74.85	25.15
	6	660	493	167	74.70	25.30
	7	659	493	166	74.81	25.19
	8	659	493	166	74.81	25.19
	9	659	493	166	74.81	25.19
	10	659	493	166	74.81	25.19

ตารางที่ 4.22 (ต่อ) ค่าความถูกต้องจากการตรวจสอบไขว้ของโมเดลฟังก์ชัน SVM

เซ็ตคู่คำ	รอบที่ (Fold)	ปริมาณที่ทดสอบ (ข่าว)	ปริมาณที่ถูกต้อง (ข่าว)	ปริมาณที่ผิดพลาด (ข่าว)	ความถูกต้อง (%)	ความผิดพลาด (%)
AC	1	660	494	166	74.85	25.15
	2	660	494	166	74.85	25.15
	3	660	494	166	74.85	25.15
	4	660	494	166	74.85	25.15
	5	660	494	166	74.85	25.15
	6	660	493	167	74.70	25.30
	7	659	493	166	74.81	25.19
	8	659	493	166	74.81	25.19
	9	659	493	166	74.81	25.19
	10	659	493	166	74.81	25.19

จากตารางที่ 4.22 จะพบว่าเซ็ตคู่คำ ME, MA และ AC มีปริมาณความผิดพลาดของการจำแนกข้อมูลมาที่สุดในรอบที่ 6 (ข่าวลำดับที่ 3,301 – 3,960) ที่จำนวน 167 ข่าว คิดเป็น 25.30%

ในการตรวจสอบความผิดพลาดจากข่าวโดยมนุษย์พบว่ามีข่าวหุ้นบางส่วน (ช่วงที่ 3,301 – 3,960) ที่ไม่สามารถสกัดคู่คำออกมาได้ เนื่องจากโครงสร้างของข่าวไม่อยู่ในรูปแบบของ คีย์เวิร์ด และคำที่มีขั้ว แสดงดังภาพที่ 4.11 ส่งผลให้ไม่มีคุณลักษณะสำหรับข่าวนั้น

IFEC (ที่มาข่าวหุ้น) เซ็น MOU โครงการโรงไฟฟ้าพลังงานลม 200 MW ที่ประเทศลาววันนี้ / จากการสอบถามทางบริษัทฯ แจ้งว่าเช้านี้ผู้บริหารเดินทางไปประเทศลาว เพื่อเซ็นโครงการโรงไฟฟ้าลม คาดว่าจะใช้เวลาก่อสร้างราว 1 ปี

ภาพที่ 4.11 ประโยคที่ไม่สามารถสกัดคู่คำได้

นอกจากปัญหาของข่าวหุ้นที่ไม่สามารถสกัดคู่คำได้แล้ว ยังพบข้อผิดพลาดจากเก็บรวบรวมคู่คำที่ยังมีปริมาณไม่มากพอตัวอย่างเช่น คู่คำ 'การรับซื้อ, ชะลอ' ในภาพที่ 4.12 ซึ่งคำนี้มีการใช้คำนำหน้า (Prefix) คือคำว่า 'การ' ซึ่งมีผลต่อการเก็บรวบรวมคู่คำในการทดลองกับข้อมูลชุดทดสอบในหัวข้อถัดไป

+DEMCO IFEC SUPER / -TPCH TPOLY BWGกระทรวงพลังงานระบุต้องการรับซื้อไฟฟ้าจากพลังงานหมุนเวียน ประเภท แดด และลม เพิ่ม ขณะที่พลังงานไฟฟ้าจาก ชีวมวล และขยะ อาจต้องชะลอการรับซื้อ เพราะสายส่งไฟฟ้าไม่พอ

ภาพที่ 4.12 ประโยคที่มีคู่คำแต่ไม่ถูกสกัด

4.7 การวิเคราะห์ข้อผิดพลาดจากผลการจำแนกข้อมูลชุดทดสอบ ด้วยโมเดลการเรียนรู้ชุดการสอน ME, MA และ AC

เป้าหมายของการวิเคราะห์ข้อผิดพลาดของการวิจัยนี้ เพื่อตรวจสอบความถูกต้องของโมเดลการเรียนรู้ของชุดการสอน และเป็นการตรวจสอบย้อนกลับไปจนถึงคู่คำที่เก็บรวบรวมเพื่อวิเคราะห์โครงสร้างและหาข้อผิดพลาด โดยมีเงื่อนไขการหาข้อผิดพลาดดังต่อไปนี้

1. การจำแนกข้อมูลชุดทดสอบ ด้วยโมเดลการเรียนรู้ชุดการสอนของคู่คำ ME, MA และ AC ให้ผลการจำแนกข้อมูลเหมือนกันทั้ง 3 รูปแบบ จะถือว่าผลที่ได้มีความน่าเชื่อถือ
2. การจำแนกข้อมูลชุดทดสอบ ด้วยโมเดลการเรียนรู้ชุดการสอนของคู่คำที่ไม่ตรงกัน จะทำการเปรียบเทียบด้วยการให้มนุษย์เป็นผู้ตรวจสอบ

หลังจากตรวจสอบผลจากขั้นตอนการจำแนกข้อมูลชุดทดสอบ ด้วยโมเดลการเรียนรู้ชุดการสอน ME, MA และ AC พบว่าข่าวหุ้นโดย บล. ธนชาต ประจำวันที่ 22 มิถุนายน 2015 ที่แสดงดังภาพที่ 4.13 ให้ผลลัพธ์การจำแนกข้อมูล ด้วยโมเดลการเรียนรู้ชุดการสอน ME, MA และ AC เป็น 0, -1 และ -1 ตามลำดับ โดยสัญลักษณ์หุ้นที่สนใจในข้อความ คือ PTTGC

บล.ธนชาต ระบุในบทวิเคราะห์ (22 มิ.ย.) ว่า SET เตรียมสร้างฐานใหม่อีกครั้ง หลังจากปรับลดลงแรงก่อนหน้า โดยกลุ่มหุ้นหลักที่แนะนำซื้อ ได้แก่ กลุ่มรับเหมา CK-STEC-SEAFCOคาดกำไรในไตรมาส 2/58 เติบโตดีอย่าง PTTGC ขณะที่หุ้นกำไร Turnaround จากไตรมาส 1/58 อย่าง SAMART-BLAโดยยังมอง Downside Risk จากการปรับลดดอกเบี้ยจำกัด...

ภาพที่ 4.13 ข่าวหุ้นโดย บล. ธนชาต ประจำวันที่ 22 มิถุนายน 2015

จากเงื่อนไขการหาข้อผิดพลาดข้อที่ 2 เมื่อผลลัพธ์การจำแนกข้อมูลชุดทดสอบไม่ตรงกัน จะมีการตรวจสอบผลของการจำแนกข้อมูลด้วยเครื่องหรือคอมพิวเตอร์เปรียบเทียบกับผลของการจำแนกข้อมูลด้วยมนุษย์ โดยภาพที่ 4.14 แสดงประโยคที่พบข้อผิดพลาดในข่าวหุ้น

...คาดกำไรในไตรมาส 2/58 เติบโตดีอย่าง PTTGC...

ภาพที่ 4.14 ประโยคที่พบข้อผิดพลาดภายในข่าวหุ้นโดย

จากการตรวจสอบด้วยมนุษย์พบว่า ความเห็นของมนุษย์ให้ผลลัพธ์เป็น 1 ต่อข่าวหุ้นนี้ ซึ่งให้ผลลัพธ์ที่ไม่ตรงกับโมเดลที่สร้างด้วยคู่คำ ME, MA และ AC ซึ่งให้ผลลัพธ์เป็น 0, -1 และ -1 ตามลำดับ

ในลำดับถัดมาเมื่อลองวิเคราะห์รูปแบบประโยคจะพบว่ากลุ่มคำที่น่าจะส่งผลต่อขั้วด้านบวกของข่าวหุ้นนี้จะเป็นกลุ่มคำว่า 'เติบโตดี' และเมื่อแยกกลุ่มคำให้อยู่ในรูปแบบของคู่คำจะได้ 'เติบโต, ดี, 1' มีเครื่องหมายกำกับอารมณ์เป็น 1 ซึ่งมีขั้วด้านบวก

รูปแบบหรือชนิดของคู่คำที่เก็บรวบรวมในงานวิจัยนี้ประกอบด้วย คีย์เวิร์ด (Keyword), คำที่มีขั้ว (Polarity word) และเครื่องหมายกำกับอารมณ์ (Sentiment) ตามที่ได้กล่าวถึงก่อนหน้าแล้ว เมื่อวิเคราะห์ชนิดของคำพบว่าเหตุผลที่คู่คำว่า 'เติบโต, ดี, 1' ไม่ถูกสกัดหรือเก็บรวบรวมลงในชุดของคู่คำ เนื่องจากคำว่า 'เติบโต' มีชนิดของคำคือคำกริยาซึ่งชนิดของคำที่เป็นคีย์เวิร์ดส่วนใหญ่จะเป็นคำนาม

ตารางที่ 4.23 เปรียบเทียบชนิดของคำที่ไม่ถูกเก็บรวบรวม และชนิดของคำภายในชุดคู่คำ ME, MA และ AC

ชุดคู่คำ	คู่คำ	ชนิดของคู่คำ
-	เติบโต, ดี	กริยา, กริยา
ME	การเติบโต, ดี	นาม, กริยา
MA	การเติบโต, ดี	นาม, กริยา
AC	การเติบโต, ดี	นาม, กริยา

จากตารางที่ 4.23 เมื่อตรวจสอบกลับไปที่ชุดของคู่คำทั้งสามรูปแบบ คือ ME, MA, AC พบว่าชุดของคู่คำทั้งสามรูปแบบมีคำที่ใกล้เคียงคำว่า 'เติบโต, ดี' คือ 'การเติบโต, ดี' ถูกเก็บรวบรวมอยู่ก่อนแล้วหากเปรียบเทียบคำเหล่านี้กับภาษาอังกฤษ ดังตารางที่ 4.24

ตารางที่ 4.24 เปรียบเทียบคู่คำระหว่างภาษาไทย และภาษาอังกฤษ

คู่คำภาษาไทย	คู่คำภาษาอังกฤษ
เติบโต (กริยา), ดี	growth (verb), good
การเติบโต (นาม), ดี	growth (noun), good

ตารางที่ 4.24 จะเห็นได้ว่าในภาษาอังกฤษคำว่า 'เติบโต' และ 'การเติบโต' ใช้คำที่เขียนเหมือนกัน แต่ชนิดของคำเป็นคนละชนิดกัน

จากรูปแบบปัญหาที่เกิดขึ้น พบว่าลักษณะการเขียนของภาษาไทยมีการเพิ่มคำนำหน้าเพื่อให้เกิดความแตกต่างระหว่างชนิดของคำต่างจากภาษาอังกฤษที่ใช้คำเดียวกัน สำหรับการแก้ไขปัญหาเรื่องชนิดของคำจะถูกแบ่งออกเป็นสองแนวทาง ได้แก่

1) ทำการเก็บรวบรวมคู่คำเพิ่มเติมด้วยรูปแบบของ 'คำกริยา, คำกริยา, ขั้วอารมณ์' เพื่อให้คู่คำครอบคลุมกรณีที่เกิดปัญหาจากการเขียนในภาษาไทย แสดงตัวอย่างในตารางที่ 4.25

ตารางที่ 4.25 การตัดคำนำหน้า (Prefix) ออกจากคู่คำเดิมเพื่อให้อยู่ในรูปคำกริยา

คำนำหน้า	คู่คำ
การ	~~การ~~ฟื้นตัว, ช้า ~~การ~~เติบโต, ดี ~~การ~~เปลี่ยนแปลง, ลง

2) ทำการเพิ่มคำนำหน้า (Prefix) ให้กับคู่คำเดิมที่ทำการเก็บรวบรวมไปแล้วเพื่อเพิ่มปริมาณลักษณะการเขียนที่เป็นไปได้แต่ยังคงรูปแบบของ 'คำนาม, คำกริยา, ขั้วอารมณ์' เดิมเอาไว้ ตัวอย่างของคำนำหน้า (Prefix) ในภาษาไทยที่นิยมใช้ได้แก่ 'การ' และ 'ความ' เป็นต้น แสดงตัวอย่างในตารางที่ 4.26

ตารางที่ 4.26 การเพิ่มคำนำหน้า (Prefix) ลงในคู่คำเดิม

คำนำหน้า	คู่คำ
การ	(การ)ปันผล, จ่าย (การ)ตลาด, ฟื้นตัว (การ)ลงทุน, เพิ่ม
ความ	(ความ)คาดหวัง, ลดลง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์เล่มนี้นำเสนอการวิเคราะห์ข่าวหุ้นภาษาไทยด้วยการใช้คู่คำ (Wordpair) เป็นคุณลักษณะ โดยวัตถุประสงค์ที่สำคัญเพื่อเป็นการศึกษาตัวแทนของข้อมูลในรูปแบบอื่นเพื่อทำการวิเคราะห์ข้อความ

ภายในการทดลองมีการเก็บรวบรวม และคัดเลือกข่าวหุ้น, การวิเคราะห์ความสัมพันธ์ก่อนการนำข้อมูลมาใช้, การสร้างและรวบรวมเซ็ตของคู่คำเพื่อใช้ในการวิเคราะห์ และการสร้างโมเดลการจำแนกข้อมูล โดยขั้นตอนที่สำคัญในการวิเคราะห์ข่าวหุ้นภาษาไทยด้วยการใช้คู่คำ ประกอบด้วย การสร้างกราฟความสัมพันธ์, การทดสอบสมมติฐาน และการจำแนกข้อมูล

ผลลัพธ์ที่ได้จากการสร้างกราฟความสัมพันธ์ (Graph visualization) เพื่อตรวจสอบสมมติฐานที่ว่า การกล่างถึงหุ้น (Stock symbol) ตัวใดตัวหนึ่ง ในข่าวหุ้นแต่ละวันจะส่งผลต่อการเปลี่ยนแปลงของราคา หรือมีการเปลี่ยนแปลงของปริมาณการซื้อขายหรือไม่ สำหรับผลลัพธ์พบว่า เมื่อทำการปรับค่าด้วยค่าดัชนีของ SET, ค่าดัชนีระดับกลุ่มอุตสาหกรรมเดียวกัน (Industry) หรือ แม้กระทั่งค่าดัชนีของกลุ่มธุรกิจเดียวกัน (Sector) เปรียบเทียบกับราคาและปริมาณการซื้อขายของหุ้นที่กล่าวถึงในข่าว พบการดีดตัวขึ้น หรือการลดต่ำลงของกราฟอย่างชัดเจนในวันที่มีข่าวหุ้นที่ค่อนข้างมีขั้วอารมณ์รุนแรง ทำให้สามารถสรุปข้อสมมติฐานได้ในระดับหนึ่งที่ว่า การกล่าวถึงหุ้นตัวใดตัวหนึ่งในข่าวหุ้นแต่ละวันจะส่งผลต่อการขึ้นลงของราคาและมีการเปลี่ยนแปลงของปริมาณการซื้อขาย

เพื่อเป็นการยืนยันผลของการวิเคราะห์ด้วยกราฟความสัมพันธ์ ผู้วิจัยจึงได้ใช้การวิเคราะห์สมมติฐานด้วยสถิติ เพื่อให้ข้อสมมติฐานมีความน่าเชื่อถือมากยิ่งขึ้น ด้วยการวิเคราะห์สมมติฐาน (Hypothesis testing) สำหรับแนวโน้มค่าเฉลี่ยภายในช่วงระยะเวลา 5 วันที่มีข่าวเพื่อดูความต่อเนื่อง สำหรับสถิติที่ใช้ในการวิเคราะห์จะใช้สถิติเครื่องหมาย-อันดับของวิลคอกซัน (Wilcoxon signed-rank test) โดยผลลัพธ์ที่ได้จากการสรุปก่อนหน้านี้พบว่าแนวโน้มค่าเฉลี่ยทั้ง 5 วันของการเปลี่ยนแปลงราคาและปริมาณการซื้อขายอยู่ในช่วงของ H_1 อย่างมีนัยสำคัญ ซึ่งหมายความว่าแนวโน้มค่าเฉลี่ยการเปลี่ยนแปลงราคาทั้ง 5 วันของวันที่มีข่าวมีความแตกต่างจากการเปลี่ยนแปลงราคาเฉลี่ยและปริมาณการซื้อขายของวันที่ไม่มีข่าว ดังนั้นการที่มีข่าวจากโบรคเกอร์ประกาศออกมาในวันนั้นจึงส่งผลต่อการขึ้นหรือลงของราคาและปริมาณการซื้อขาย

ในการวิจัยพบว่าการนำคู่คำซึ่งเป็นตัวแทนของข้อความมาเป็นคุณลักษณะ ด้วยเซ็ตคู่คำ AC ซึ่งมีจำนวนคู่คำทั้งสิ้น 331 คู่คำ เมื่อเปรียบเทียบกับการนำคำภายในประโยคมาเป็นคุณลักษณะ ที่ต้องใช้คำที่อยู่ภายในประโยคทุกคำในการวิเคราะห์ พบว่าทั้งสองรูปแบบให้ประสิทธิภาพในการ

วิเคราะห์ที่ใกล้เคียงกัน แต่ในแง่ของปริมาณคุณลักษะหรือขนาดของเวกเตอร์ที่ใช้ในการจำแนกข้อมูล การใช้คู่คำเป็นคุณลักษณะจะมีจำนวนน้อยกว่าการใช้คำเป็นคุณลักษณะเป็นอย่างมาก

นอกจากนี้การทดลองเพิ่มเติมเฉพาะคำที่แสดงเฉพาะอารมณ์ (Polarity words) อย่างเดียวซึ่งเป็นเซ็ตย่อยของคำ (Individual words) เพื่อทดสอบประสิทธิภาพเปรียบเทียบกับคู่คำ ในภาคผนวก ก ตารางที่ ก.6 พบว่าการใช้คู่คำให้ประสิทธิภาพที่ดีกว่า จึงเป็นเหตุผลในการเลือกคู่คำมาเป็นคุณลักษณะแทนการใช้คำเพียงอย่างเดียว

5.2 ข้อเสนอแนะ

การใช้คู่คำ (Wordpair) เพื่อเป็นคุณลักษณะในการจำแนกข้อมูล เป็นเพียงการนำเสนอตัวแทนของการวิเคราะห์ข้อความในรูปแบบหนึ่งเท่านั้น ทั้งนี้สาระสำคัญของการวิเคราะห์ข้อความอยู่ที่การคัดเลือกรูปแบบของข้อความที่จะใช้ในการวิเคราะห์เป็นหลัก ซึ่งเป็นปัจจัยที่ควบคุมได้ยากและมีความผันผวนสูง สำหรับการใช้คู่คำในการวิเคราะห์ข่าวหุ้นอาจจะเหมาะสม แต่สำหรับการนำคู่คำไปประยุกต์ใช้กับข้อความประเภทอื่นอาจจะต้องมีการปรับปรุงรูปแบบของการวิเคราะห์ในอนาคตดังต่อไปนี้

1) สำหรับเซ็ตของคู่คำ นอกจากคู่คำเฉพาะสำหรับข่าวหุ้นของ ME, MA และ AC แล้ว การนำไปประยุกต์ใช้กับเนื้อหาอื่นจำเป็นต้องมีการเพิ่มคู่คำใหม่เปลี่ยนแปลงไปตามแต่ละบริบทของข้อความ
2) การกำหนดระยะของข้อความที่ต้องการวิเคราะห์ และวินโดวไซต์ที่ใช้ในการสกัดคู่คำ เนื่องจากภาษาไทยมีรูปแบบของประโยคหลายประเภท ได้แก่ ประโยคความเดียว ประโยคความรวม และประโยคความซ้อน ซึ่งในการวิเคราะห์ข่าวหุ้น ผู้วิจัยจึงได้ใช้สัญลักษณ์หุ้นเพื่อเป็นตัวอ้างอิงและแยกแยะประโยค ดังนั้นในการนำไปประยุกต์ใช้กับการวิเคราะห์ข้อความประเภทอื่นจำเป็นจะต้องมีการจัดการในส่วนนี้เพื่อแยกตัวแทนของประโยคที่ซ่อนอยู่ออกมาวิเคราะห์
3) การวิเคราะห์อารมณ์ภายในข่าวหุ้นของโบรคเกอร์ ต้องมีการตรวจสอบจุดมุ่งหมายของโบรคเกอร์ว่าต้องการกล่าวถึงหุ้นตัวใดเป็นหลัก เนื่องจากการเขียนข่าวของโบรคเกอร์อาจจะมีการรวมกันของสัญลักษณ์หุ้น ซึ่งอาจจะให้อารมณ์ของแต่สัญลักษณ์ในทิศทางเดียวกันภายในข่าว แต่เป็นการส่งผลต่อสัญลักษณ์หุ้นนั้นในทางอ้อม

เอกสารอ้างอิง

Apinan Chattupan, and Ponrudee Netisopakul. "Stock sentiment analysis model using data mining (In Thai)." *In Knowledge and Smart Technology (KST), 2014. Proceeding of 6th National Conference on.* Chonburi, Thailand.

Changhou Wang, and Fei Wang. "A Bootstrapping Method for Extracting Sentiment Words Using Degree Adverb Patterns." *Computer Science & Service System (CSSS), 2012 International Conference on*. IEEE, 2012.

Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon, and Kanokorn Trakultaweekoon. "S-Sense: A Sentiment Analysis Framework for Social Media Sensing." *Sixth International Joint Conference on Natural Language Processing*.

Desheng Dash Wu, Lijuan Zheng, and David L. Olson. "A decision support approach for online stock forum sentiment analysis." *Systems, Man, and Cybernetics: Systems, IEEE Transactions on* 44.8 (2014): 1077-1087.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Wai Lam. "News sensitive stock trend prediction." *Advances in knowledge discovery and data mining*. Springer Berlin Heidelberg, 2002. 481-493.

Hui-Hsin Wu, Angela Charng-Rurng Tsai, Richard Tzong-Han Tsai, and Jane Yung-jen Hsu. "Sentiment value propagation for an integral sentiment dictionary based on commonsense knowledge." *Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on*. IEEE, 2011.

Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. "Contextifier: automatic generation of annotated stock visualizations." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013.

Jiang YANG, Min HOU, and Ning WANG. "Recognizing sentiment polarity in Chinese reviews based on topic sentiment sentences." *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*. IEEE, 2010.

Keisuke Mizumoto, Hidekazu Yanagimoto, and Michifumi Yoshioka. "Sentiment analysis of stock market news with semi-supervised learning." *Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on*. IEEE, 2012.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. "Lexicon-based methods for sentiment analysis." *Computational linguistics* 37.2 (2011): 267-307.

เอกสารอ้างอิง (ต่อ)

Marc-André Mittermayer. 2004. "Forecasting intraday stock price trends with text mining techniques." *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, 2004.

Marina Boia, Boi Faltings, Claudiu-Cristian Musat, and Pearl Pu. "A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets." *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013.

Na Fan, Wandong Cai, and Yu Zhao. "Research on the Model of Multiple Levels for Determining Sentiment of Text." *Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on*. Vol. 1. IEEE, 2008.

Nattadaporn Lertcheva, and Wirote Aroonmanakun. "A linguistic study of product names in Thai economic news." *Natural Language Processing, 2009. SNLP'09. Eighth International Symposium on*. IEEE, 2009.

Nattapong Tongtep, and Thanaruk Theeramunkong. "Pattern-based extraction of named entities in Thai news documents." *Thammasat International Journal of Science and Technology* 15.1 (2010).

Phaisarn Sutheebanjard, and Wichian Premchaiswadi. "Disambiguation of Thai personal name from online news articles." *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*. Vol. 3. IEEE, 2010.

Rathawut Lertsuksakda, Kitsuchart Pasupa, and Ponrudee Netisopakul. "Sentiment analysis of Thai children stories on support vector machine." *In Artificial Life and Robotics (AROB), 2015. Proceeding of the Twentieth International Symposium on*. Beppu, Japan.

Rathawut Lertsuksakda, Ponrudee Netisopakul, and Kitsuchart Pasupa. "Thai sentiment terms construction using the Hourglass of Emotions." *In Knowledge and Smart Technology (KST), 2014 6th International Conference on*. 46-50. IEEE.

Robert P. Schumaker, and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." *ACM Transactions on Information Systems (TOIS)* 27.2 (2009): 12.

เอกสารอ้างอิง (ต่อ)

Tosaporn Vichayakitti, and Chuleerat Jaruskulchai. "Automatic tagging time-revealing vocabulary from Thai article news." *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*. Vol. 2. IEEE, 2005.

Yang Gao, Li Zhou, Yong Zhang, Chunxiao Xing, Yigang Sun, and Xianzhong Zhu. "Sentiment classification for stock news." *Pervasive Computing and Applications (ICPCA), 2010 5th International Conference on*. IEEE, 2010.

Bualuang Securities. Retrieved February 15, 2015. http://www.bualuang.co.th/th/index.php

SET Market Analysis and Reporting Tool Retrieved 2015, February 15, http://www.setsmart.com/

Stock News Online. Retrieved May 15, 2015. http://www.kaohoon.com/online/content/category/13/ภาวะเศรษฐกิจและตลาดหุ้นในประเทศ

Weka 3.7.1. Retrieved October 1, 2013. http://www.cs.waikato.ac.nz/ml/weka

กิรดา กลีบมาลัย. "Data Visualization." Retrieved 2015. http://v54-30037.blogspot.com/2015/01/graph-chart.html

จิรา แก้วสุวรรณ์. "การตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน." Retrieved 2015. http://www.gits.kmutnb.ac.th/ethesis/data/4620781031.pdf

อัจฉริยา ปราบอริพ่าย. "สถิติที่ไม่ใช้พารามิเตอร์." Retrieved 2015. http://pirun.ku.ac.th/~faasatp/734415/data/chapter7.pdf

ภาคผนวก ก.

รายละเอียดผลการทดลอง

ผลการทดลองการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ และฟังก์ชัน SVM โดยใช้คู่คำเซ็ต ME, MA และ AC เป็นคุณลักษณะ ปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100, 500, 1,000 - 5,000 และกำหนดวินโดว์ไซต์ 60 ตัวอักษร

ตารางที่ ก.1 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจและฟังก์ชัน SVM โดยใช้คู่คำเซ็ต AC และปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด

Input per class	Class	Decision tree (AC)			SVM (AC)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
100	-1	0.324	0.800	0.461	0.338	0.890	0.490
	0	0.667	0.120	0.203	0.684	0.130	0.218
	1	0.286	0.100	0.148	0.611	0.110	0.186
	Average	0.425	0.340	0.271	0.545	0.377	0.298
500	-1	0.353	0.912	0.509	0.351	0.400	0.374
	0	0.717	0.142	0.237	0.686	0.140	0.233
	1	0.606	0.132	0.217	0.349	0.578	0.435
	Average	0.559	0.395	0.321	0.462	0.373	0.347
1,000	-1	0.369	0.941	0.530	0.366	0.939	0.526
	0	0.796	0.176	0.288	0.764	0.155	0.258
	1	0.780	0.177	0.289	0.664	0.152	0.247
	Average	0.648	0.431	0.369	0.598	0.415	0.344
1,500	-1	0.368	0.951	0.531	0.367	0.953	0.530
	0	0.840	0.183	0.300	0.773	0.168	0.276
	1	0.773	0.154	0.257	0.707	0.129	0.218
	Average	0.660	0.429	0.363	0.616	0.417	0.341

ตารางที่ ก.1 (ต่อ) แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจและฟังก์ชัน SVM โดยใช้คู่คำเซ็ต AC และปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด

Input per class	Class	Decision tree (AC)			SVM (AC)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
2,000	-1	0.362	0.950	0.524	0.362	0.960	0.526
	0	0.825	0.184	0.300	0.818	0.176	0.289
	1	0.731	0.113	0.196	0.714	0.096	0.169
	Average	0.639	0.416	0.340	0.631	0.410	0.328
2,500	-1	0.363	0.963	0.528	0.362	0.967	0.526
	0	0.867	0.190	0.312	0.846	0.180	0.297
	1	0.758	0.099	0.175	0.708	0.080	0.144
	Average	0.663	0.417	0.338	0.639	0.409	0.323
3,000	-1	0.405	0.245	0.305	0.762	0.070	0.127
	0	0.867	0.189	0.310	0.829	0.181	0.297
	1	0.364	0.792	0.499	0.359	0.965	0.524
	Average	0.545	0.409	0.371	0.638	0.405	0.316
3,500	-1	0.393	0.337	0.363	0.729	0.075	0.136
	0	0.867	0.193	0.315	0.861	0.182	0.301
	1	0.369	0.707	0.485	0.361	0.968	0.526
	Average	0.543	0.412	0.388	0.650	0.409	0.321
4,000	-1	0.465	0.159	0.237	0.706	0.071	0.129
	0	0.883	0.194	0.317	0.857	0.184	0.303
	1	0.364	0.888	0.516	0.359	0.965	0.524
	Average	0.570	0.413	0.357	0.641	0.407	0.319
4,500	-1	0.778	0.079	0.144	0.736	0.076	0.137
	0	0.873	0.192	0.315	0.859	0.186	0.305
	1	0.364	0.976	0.531	0.362	0.970	0.527
	Average	0.672	0.416	0.330	0.652	0.410	0.323

ตารางที่ ก.1 (ต่อ) แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจและฟังก์ชัน SVM โดยใช้คู่คำเซ็ต AC และปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด

Input per class	Class	Decision tree (AC)			SVM (AC)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
5,000	-1	0.847	0.073	0.135	0.745	0.076	0.138
	0	0.880	0.194	0.318	0.863	0.187	0.307
	1	0.364	0.981	0.532	0.361	0.969	0.526
	Average	0.697	0.416	0.328	0.657	0.410	0.324

ผลการทดลองการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คำเป็นคุณลักษณะ จำนวน 133, 277 และ 331 คำ อ้างอิงขนาดใกล้เคียงกันกับเซ็ตคู่คำ ME, MA และ AC และกำหนดวินโดว์ไซต์ 60 ตัวอักษร

ตารางที่ ก.2 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คำเป็นคุณลักษณะจำนวน 133, 277 และ 331 คำ และมีวินโดว์ไซต์ 60 ตัวอักษร

Amount of individual words	Class	Decision tree		
		Precision	Recall	F-measure
133	-1	0.614	0.231	0.336
	0	0.558	0.135	0.217
	1	0.790	0.963	0.868
	Average	0.743	0.774	0.728
277	-1	0.604	0.277	0.380
	0	0.483	0.135	0.211
	1	0.799	0.954	0.870
	Average	0.744	0.777	0.738
331	-1	0.604	0.270	0.373
	0	0.543	0.157	0.243
	1	0.799	0.956	0.870
	Average	0.747	0.778	0.739

ผลการทดลองการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คำเป็นคุณลักษณะ จำนวน 331 คำ อ้างอิงขนาดใกล้เคียงกันกับเซ็ตคู่คำ ME, MA และ AC ปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100, 500, 1,000 - 5,000 และกำหนดวินโดว์ไซต์ 60 ตัวอักษร

ตารางที่ ก.3 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คำเป็นคุณลักษณะจำนวน 331 คำ มีวินโดว์ไซต์ 60 ตัวอักษร และปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด

Input per class	Class	Decision tree (331 Individual words)		
		Precision	Recall	F-measure
100	-1	0.319	0.530	0.398
	0	0.388	0.330	0.357
	1	0.531	0.260	0.349
	Average	0.413	0.373	0.368
500	-1	0.419	0.620	0.500
	0	0.502	0.498	0.500
	1	0.762	0.404	0.528
	Average	0.561	0.507	0.510
1,000	-1	0.806	0.353	0.491
	0	0.478	0.984	0.643
	1	0.835	0.420	0.559
	Average	0.706	0.586	0.564
1,500	-1	0.848	0.398	0.542
	0	0.490	1.000	0.658
	1	0.882	0.433	0.581
	Average	0.740	0.610	0.594
2,000	-1	0.825	0.408	0.546
	0	0.494	1.000	0.661
	1	0.920	0.442	0.597
	Average	0.746	0.617	0.601

ตารางที่ ก.3 (ต่อ) แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คำเป็นคุณลักษณะจำนวน 331 คำ มีวินโดว์ไซต์ 60 ตัวอักษร และปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด

Input per class	Class	Decision tree (331 Individual words)		
		Precision	Recall	F-measure
2,500	-1	0.856	0.434	0.576
	0	0.497	1.000	0.664
	1	0.925	0.447	0.603
	Average	0.759	0.627	0.614
3,000	-1	0.845	0.454	0.590
	0	0.492	1.000	0.659
	1	0.972	0.419	0.585
	Average	0.770	0.624	0.612
3,500	-1	0.864	0.459	0.599
	0	0.493	1.000	0.661
	1	0.981	0.433	0.600
	Average	0.779	0.630	0.620
4,000	-1	0.869	0.478	0.617
	0	0.497	1.000	0.664
	1	0.989	0.433	0.602
	Average	0.785	0.637	0.628
4,500	-1	0.858	0.468	0.606
	0	0.493	1.000	0.661
	1	0.995	0.424	0.595
	Average	0.782	0.631	0.620
5,000	-1	0.867	0.483	0.620
	0	0.496	1.000	0.663
	1	0.996	0.425	0.596
	Average	0.786	0.636	0.626

ผลการทดลองการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต ME, MA และ AC เป็นคุณลักษณะ และไม่กำหนดวินโดว์ไซต์

ตารางที่ ก.4 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต ME, MA และ ACโดยไม่กำหนดวินโดว์ไซต์

Wordpair sets	Class	Decision tree		
		Precision	Recall	F-measure
ME	-1	0.798	0.412	0.543
	0	0.810	0.574	0.672
	1	0.853	0.981	0.913
	Average	0.840	0.846	0.826
MA	-1	0.845	0.447	0.585
	0	0.769	0.574	0.657
	1	0.861	0.986	0.919
	Average	0.854	0.856	0.839
AC	-1	0.831	0.449	0.583
	0	0.742	0.549	0.631
	1	0.861	0.983	0.918
	Average	0.849	0.854	0.836

ผลการทดลองการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต AC เป็นคุณลักษณะ ปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100, 500, 1,000 - 5,000 และไม่กำหนดวินโดว์ไซต์

ตารางที่ ก.5 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต AC ปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด และไม่กำหนดวินโดว์ไซต์

Input per class	Class	Decision tree (AC)		
		Precision	Recall	F-measure
100	-1	0.593	0.480	0.530
	0	0.583	0.700	0.636
	1	0.636	0.630	0.633
	Average	0.604	0.603	0.600
500	-1	0.860	0.788	0.741
	0	0.772	0.950	0.776
	1	0.925	0.790	0.792
	Average	0.853	0.843	0.770
1,000	-1	0.800	0.875	0.836
	0	0.893	0.868	0.880
	1	0.948	0.885	0.915
	Average	0.880	0.876	0.877
1,500	-1	0.737	0.900	0.811
	0	0.906	0.889	0.898
	1	0.960	0.766	0.852
	Average	0.868	0.852	0.853
2,000	-1	0.675	0.924	0.780
	0	0.916	0.872	0.893
	1	0.965	0.655	0.780
	Average	0.852	0.817	0.818

ตารางที่ ก.5 (ต่อ) แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต AC ปรับความเท่ากันของชุดข้อมูล -1, 0 และ 1 อย่างละ 100 - 5,000 ชุด และไม่กำหนดวินโดว์ไซต์

Input per class	Class	Decision tree (AC)		
		Precision	Recall	F-measure
2,500	-1	0.648	0.932	0.765
	0	0.927	0.884	0.905
	1	0.964	0.586	0.729
	Average	0.846	0.801	0.799
3,000	-1	0.932	0.538	0.683
	0	0.926	0.875	0.900
	1	0.641	0.947	0.764
	Average	0.833	0.787	0.782
3,500	-1	0.939	0.536	0.682
	0	0.932	0.871	0.900
	1	0.638	0.954	0.764
	Average	0.836	0.787	0.782
4,000	-1	0.937	0.508	0.658
	0	0.940	0.873	0.905
	1	0.627	0.959	0.758
	Average	0.835	0.780	0.774
4,500	-1	0.945	0.547	0.693
	0	0.936	0.870	0.902
	1	0.266	0.961	0.771
	Average	0.842	0.793	0.789
5,000	-1	0.944	0.524	0.674
	0	0.938	0.875	0.905
	1	0.635	0.960	0.764
	Average	0.839	0.786	0.781

ผลการทดลองเพิ่มเติมการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต ME, MA และ AC และไม่กำหนดวินโดว์ไซต์ เปรียบเทียบกับการใช้เฉพาะคำที่แสดงเฉพาะอารมณ์ (Polarity words) โดยเก็บคำจำนวน 133, 277 และ 331 และไม่กำหนดวินโดว์ไซต์

ตารางที่ ก.6 แสดงผลความถูกต้องของการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ โดยใช้คู่คำเซ็ต ME, MA และ AC และไม่กำหนดวินโดว์ไซต์ เปรียบเทียบกับการใช้เฉพาะคำที่แสดงเฉพาะอารมณ์ (Polarity words) โดยเก็บคำจำนวน 133, 277 และ 331 และไม่กำหนดวินโดว์ไซต์

Polarity words	133 คำ			277 คู่คำ			331 คู่คำ		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	0.807	0.823	0.800	0.836	0.846	0.828	0.835	0.846	0.828
Wordpairs	ME (133 คู่คำ)			MA (277 คู่คำ)			AC (331 คู่คำ)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	0.840	0.846	0.826	0.854	0.856	0.839	0.849	0.854	0.836

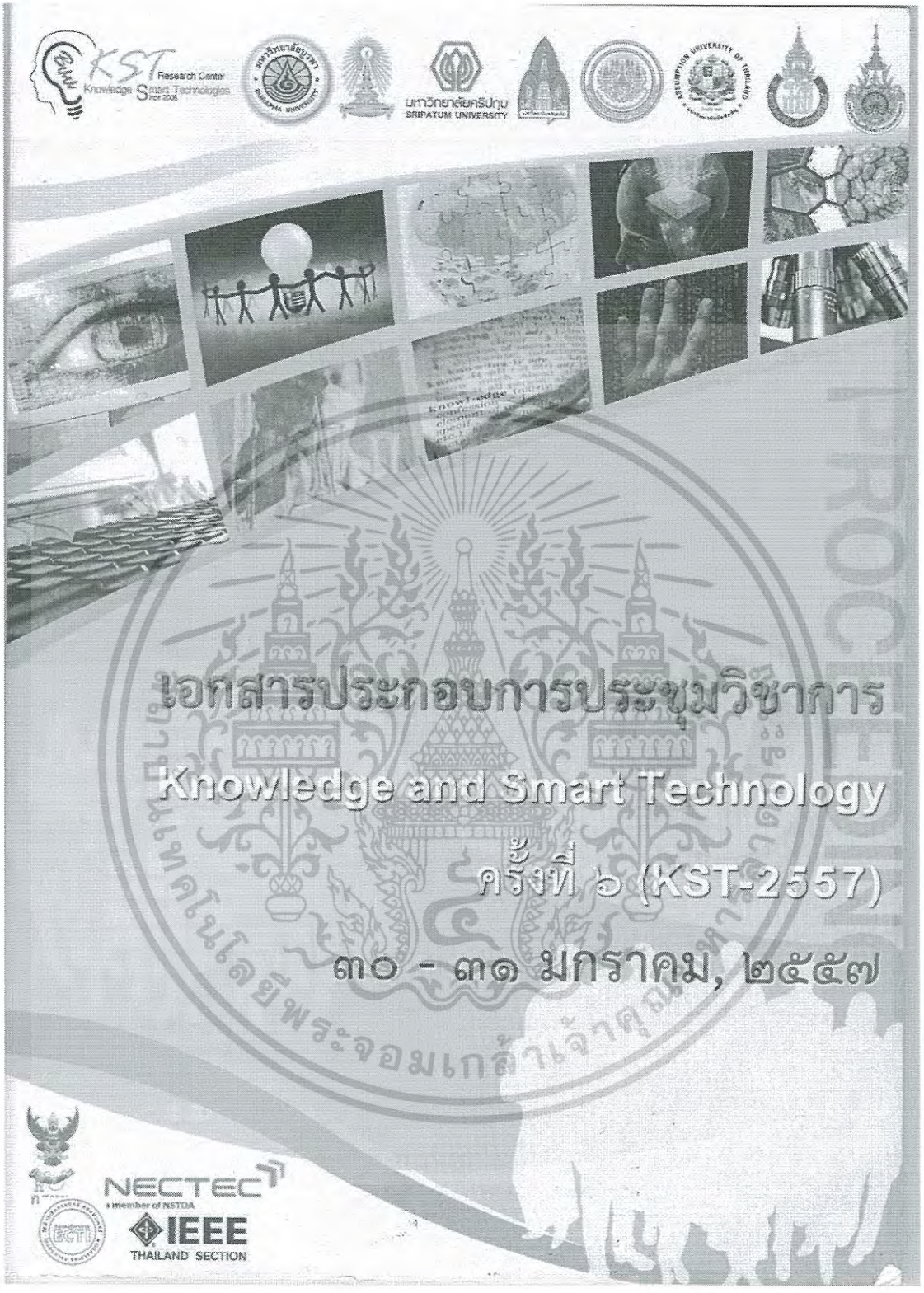
ภาคผนวก ข.

ผลงานวิจัยที่ได้รับการตีพิมพ์

ภาคผนวกนี้นำเสนอบทความฉบับสมบูรณ์ของผลงานวิจัย ซึ่งได้รับการตอบรับและตีพิมพ์ลงในวารสารการประชุมระดับชาติและวารสารการประชุมระดับนานาชาติ ซึ่งเป็นส่วนหนึ่งของการทำวิทยานิพนธ์เล่มนี้ โดยมีรายการบทความดังต่อไปนี้

- Apinan Chattupan, and Ponrudee Netisopakul. "Stock sentiment analysis model using data mining (In Thai)." *In Knowledge and Smart Technology (KST), 2014. Proceeding of 6th National Conference on.* Chonburi, Thailand.

- Apinan Chattupan, and Ponrudee Netisopakul. "Thai Stock News Sentiment Classification using Wordpair Features." *In Proceeding of 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29.* pp. 188-195. Shanghai, China. Oct. 30 – Nov. 1, 2015.



การประชุมวิชาการ Knowledge and Smart Technology ครั้งที่ ๖ (KST-2557)

แบบจำลองการวิเคราะห์อารมณ์หุ้นโดยการทำเหมืองข้อมูล
Stock Sentiment Analysis Model Using Data Mining

อภินันท์ จัตตุพันธ์[1], พรฤดี เนติโสภากุล[2]
คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
apinan.chn@gmail.com[1], ponrudee@it.kmitl.ac.th[2]

บทคัดย่อ

ปัจจุบันข่าวธุรกิจและรายงานตลาดหลักทรัพย์ มีบทบาทสำคัญในการใช้ประกอบการวิเคราะห์ และตัดสินใจในการซื้อ ขายหลักทรัพย์ งานวิจัยนี้จึงได้นำเสนอแบบจำลองการวิเคราะห์อารมณ์หุ้นสำหรับดัชนีตลาดหุ้นไทย โดยการทำเหมืองข้อความ จากข่าวธุรกิจและรายงานตลาดหลักทรัพย์ โดยรูปแบบของข่าวเป็นภาษาไทย และทำการจำแนกอารมณ์ของข่าว ว่ามีอารมณ์บวก เป็นกลาง หรือลบ จากนั้น ทำการพล็อตกราฟของอารมณ์ที่จำแนกกับเปอร์เซ็นต์เปลี่ยนแปลงราคาปิดแต่ละวันของหุ้นกลุ่มต่างๆ ผลการทดลองจากข้อมูลย้อนหลัง 6 เดือน พบว่า ค่าสหสัมพันธ์ระหว่างอารมณ์ข่าวหุ้นและเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของหุ้น ให้ค่าสัมประสิทธิสหสัมพันธ์ในหุ้นกลุ่มปิโตรเคมี และเคมีภัณฑ์ 0.74 และในหุ้นกลุ่มพลังงาน 0.61 ดังนั้น จึงมีความสัมพันธ์ระหว่างอารมณ์ข่าวหุ้นกับราคาเปลี่ยนแปลง และสามารถนำการวิเคราะห์อารมณ์ข่าวหุ้นไปประกอบการตัดสินใจซื้อ ขายหลักทรัพย์ต่อไปได้

คำสำคัญ: การวิเคราะห์อารมณ์, การจำแนกข้อความ, ข่าวธุรกิจ, ข่าวหุ้น, ค่าสัมประสิทธิสหสัมพันธ์

Abstract

The up-to-date stock exchange report and business news play important roles for stock trading decisions. This paper proposes a sentiment analysis model for Stock Exchange of Thailand (SET) using text classification. The stock report and business news texts are classified into positive, negative and neutral. Classification results are plotted against stock closing prices. The experiment, using 6 months data collection, gives a correlation coefficient of 0.74 in petrochemical and chemical product stocks and 0.61 in energy stocks. Therefore, there are correlations between stock sentiments and stock prices. Hence, stock sentiments may be used to predict stock price direction and used as one factor for stock trading decisions.

Key Words: Sentiment Analysis, Text Classification, Business news, Stock Exchange, Coefficient Correlation

1. บทนำ

การวิเคราะห์ปัจจัยที่เกี่ยวข้องกับหลักทรัพย์ มีบทบาทอย่างยิ่ง ต่อการตัดสินใจซื้อ ขายหลักทรัพย์ ซึ่งหนึ่งในนั้นคือการวิเคราะห์ข้อมูลที่เป็นข่าว เกี่ยวกับภาวะเศรษฐกิจและตลาดหุ้น ซึ่งการอ่านทำความเข้าใจ หรือวิเคราะห์ข่าวของแต่ละบุคคลมีวิจารณญาณที่แตกต่างกัน ทำให้เกิดปัญหาว่าแต่ละข่าวควรจะมีผลเป็นด้านใด

จากปัญหาที่พบ ผู้วิจัยได้เล็งเห็นว่าการทำเหมืองข้อมูลสามารถใช้วิเคราะห์เนื้อหาข่าว โดยการจำแนกเป็นอารมณ์บวก เป็นกลาง หรือลบ แล้วนำอารมณ์ที่ได้มาเปรียบเทียบกับราคาปิดของหุ้นในแต่ละวัน เพื่อหาทิศทางและความสัมพันธ์

งานวิจัยนี้ได้นำเสนอการวิเคราะห์ความสัมพันธ์ระหว่างอารมณ์ของข่าวภาวะเศรษฐกิจและตลาดหุ้น โดยใช้การจำแนกอารมณ์ข่าวจากการทำเหมืองข้อมูลว่ามีอารมณ์บวก เป็นกลาง หรือลบ และเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดกลุ่มหุ้น เพื่อเปรียบเทียบความสัมพันธ์โดยใช้ค่าสหสัมพันธ์เป็นตัวชี้วัด

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

คือ การค้นหารูปแบบ (Pattern) แนวทาง (Approach) และความสัมพันธ์ (Association) [1, 2] จากชุดข้อมูลที่มีจำนวนมาก โดยอ้างอิงหลักการทางสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักการทางคณิตศาสตร์ ทำให้สามารถวิเคราะห์รูปแบบของข้อมูลออกมาเป็นข้อสรุปที่ใช้ในการตัดสินใจได้ โดยการทำเหมืองข้อมูลประกอบไปด้วยขั้นตอนหลัก ดังนี้

1. การทำความเข้าใจปัญหา
2. การเตรียมข้อมูล
3. การสร้างแบบจำลอง
4. การประเมินผลลัพธ์
5. การนำไปใช้งาน

ซึ่งในการทำเหมืองข้อมูลเป็นที่นิยมสำหรับข้อมูลที่ต้องการจำแนกประเภท (Classification) โดยการระบุประเภทของข้อมูลจากคุณลักษณะของข้อมูล โดยในงานวิจัย [3, 4] มีการสกัดคุณลักษณะของข้อมูลออกจากความคิดเห็น และมีการจำแนกกลุ่มความคิดเห็นออกเป็นเชิงบวก และลบ

2.2 ต้นไม้ตัดสินใจ (Decision Tree)

คือ โครงสร้างของข้อมูลที่มีลักษณะเป็นลำดับชั้น (Hierarchy) ประกอบด้วยโหนดราก (Root node) ซึ่งอยู่ด้านบนสุด ในแต่ละโหนดจะมีคุณลักษณะ (Attribute) ใช้สำหรับทดสอบกิ่ง (Branch) ซึ่งแตกออกมาจากโหนดแสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกมาทดสอบ และโหนดด้านล่างสุดคือใบ (Leaf) ซึ่งแสดงผลลัพธ์ที่ได้จากการตัดสินใจ (Class)

งานวิจัยนี้ใช้อัลกอริทึม J48 หรือ C4.5 ในการสร้างการเรียนรู้ต้นไม้ตัดสินใจ ถูกพัฒนาขึ้นโดย Ross Quinlan [5] โดยพัฒนาขึ้นมาจากอัลกอริทึม ID3 ที่ใช้หลักการของ Information Entropy โดยที่ C4.5 พิจารณาความแตกต่างในเอนโทรปี (Entropy) สำหรับการแบ่งกลุ่มข้อมูลไปยังกลุ่มย่อย จากนั้นนำเอนโทรปีมาคำนวณหาค่าเกน (Gain) เพื่อนำค่าสูงสุดมาสร้างเป็นโหนดราก หลังจากนั้นอัลกอริทึม C4.5 จะคำนวณค่าสารสนเทศของการแบ่งแยก (Split Information) เพื่อนำมาเทียบอัตราส่วนกับค่าเกน แล้วทำการเทียบอัตราส่วนระหว่างสมการเพื่อคำนวณหาค่าอัตราส่วนเกน (Gain Ratio) โดยมีสมการคำนวณ ดังนี้

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (1)$$

หลังจากที่ได้ค่าอัตราส่วนเกน จากสมการจะนำค่าสูงสุดมาสร้างเป็นโหนดราก และจะคำนวณใหม่ โดยไม่นำโหนดที่เลือกไปแล้วมาคำนวณอีก เพื่อเลือกค่าสูงสุดมาสร้างเป็นโหนดถัดไป

2.3 เนอีฟเบย์ (Naïve Bayes)

คือ การใช้พื้นฐานจากทฤษฎีของเบย์ (Bayes theorem) [6] เพื่อคำนวณความน่าจะเป็น เพื่อใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปร สำหรับสร้างเงื่อนไขความน่าจะเป็นมาใช้ในการจำแนกข้อมูล ซึ่งต้องมีข้อมูลการเรียนรู้ก่อนหน้าใช้ในการวิเคราะห์

$$V_{NB} = argmax\, P(v_j) * \prod_{i=1}^{n} P(a_i|v_j) \quad (2)$$

โดย a_i คือแอตทริบิวต์ที่ลำดับที่ i
v_j คือค่าของแอตทริบิวต์ a ลำดับที่ j

2.4 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

คือ การหาระนาบในการตัดสินใจ โดยทำการแบ่งข้อมูลออกเป็นสองกลุ่ม ให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มให้มากที่สุด [7] เหมาะสำหรับข้อมูลที่มีมิติของข้อมูลสูง โดยเขียนมิติข้อมูลแทนด้วย (x_i, y_i) และมีสมการ ดังนี้

$$w * x + b > 0\,; y_i = 1 \quad (3)$$

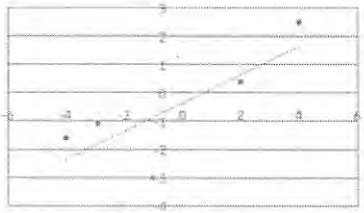
$$w * x + b < 0\,; y_i = -1 \quad (4)$$

โดย w คือค่าน้ำหนัก
b คือค่าความเอนเอียง

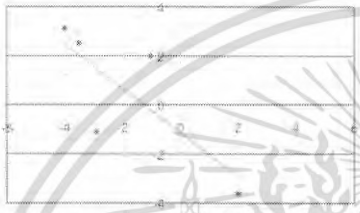
2.5 ค่าสหสัมพันธ์ (Correlation)

คือ สถิติที่ใช้หาความสัมพันธ์ระหว่างตัวแปรสองตัวขึ้นไป โดยค่าที่คำนวณได้เรียกว่า ค่าสัมประสิทธิสหสัมพันธ์ (Correlation coefficient) โดยค่าสัมประสิทธิสหสัมพันธ์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยมีทิศทางของความสัมพันธ์ 3 แบบคือ สหสัมพันธ์ทางบวก (Positive correlation) หมายถึงความสัมพันธ์ที่ตัวแปรมีทิศทางเดียวกัน สหสัมพันธ์ทางลบ (Negative correlation) หมายถึงความสัมพันธ์ที่

ตัวแปรมีทิศทางตรงข้ามกัน และสหสัมพันธ์เป็นศูนย์หมายถึงตัวแปรไม่มีความสัมพันธ์กัน แสดงในภาพที่ 1, 2 ดังนี้



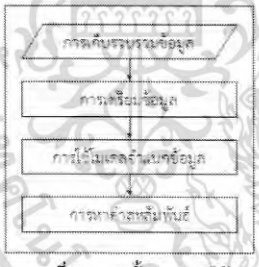
ภาพที่ 1: สหสัมพันธ์ทางบวก



ภาพที่ 2: สหสัมพันธ์ทางลบ

3. วิธีดำเนินงานวิจัย

งานวิจัยนี้ได้เลือกข้อมูลข่าวภาวะเศรษฐกิจและตลาดหุ้นในประเทศ จากเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์ [8] และมีขั้นตอนในการวิจัย แสดงในภาพที่ 3 ดังนี้



ภาพที่ 3: แสดงขั้นตอนการวิจัย

3.1 การเก็บรวบรวมข้อมูล

งานวิจัยนี้ได้รวบรวมข้อมูลข่าวจากเว็บไซต์ข่าวหุ้นธุรกิจออนไลน์ ในหมวดภาวะเศรษฐกิจและตลาดหุ้นในประเทศ มาเป็นข้อมูลต้นแบบสำหรับการทดลอง โดยมีการเก็บรวบรวมตั้งแต่วันที่ 1 เมษายน - 30 กันยายน 2556 ย้อนหลังเป็นเวลาทั้งสิ้น 6 เดือน โดยข้อมูลที่รวบรวมมาจะถูกเก็บอยู่ในรูปแบบแฟ้มข้อมูล Arff แสดงตัวอย่างข้อมูลดังนี้ "โบรกแนะเลือกซื้อ 11 บจ. อนาคตรุ่ง ดัชนีลุ้นฟื้นตัว สะสมหุ้นจ่ายปันผลระหว่างกาล ผู้สื่อข่าวรายงานว่า เช้านี้ ณ เวลา 9.39 น. ค่าเงินบาทอยู่ที่ 31.09 บาทต่อเหรียญสหรัฐ ขณะที่ตลาดหุ้นเอเชียมีทั้งปรับตัวอยู่ในแดนลบและแดนบวก นักวิเคราะห์คาดหุ้นไทยฟื้นตัว แต่ยังคงเปราะบางและผันผวน หากเป็นนักลงทุนรอรับต่ำกว่า 1400 จุด สะสมหุ้นที่จ่ายปันผลระหว่างกาล เลือกซื้อ 11 หุ้นเด่น ได้แก่ BGH, CPF, MALEE, SF, SAT, TNH, TICON, DRT, TICON, TCAP, PTTEP"

3.2 การเตรียมข้อมูล

หลังจากขั้นตอนการเก็บรวบรวมข้อมูลเรียบร้อยแล้ว จะต้องทำการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการทำเหมืองข้อมูล โดยในขั้นตอนการเตรียมข้อมูลมี 3 ขั้นตอนหลัก ดังนี้

การตัดคำ (Word Segmentation) [9] เนื่องจากภาษาไทยมีรูปแบบการเขียนข้อความติดกันเป็นสายยาว ไม่สามารถนำมาใช้ในการประมวลผลคำได้ ต่างจากภาษาอังกฤษที่มีรูปแบบการเขียนที่เว้นช่องว่างระหว่างคำ จึงมีความจำเป็นต้องใช้การตัดคำสำหรับข้อความภาษาไทย โดยการตัดคำนี้ได้ใช้เทคนิคของการตัดคำแบบไบนริดจ์ เพื่อให้ข้อความที่ตัดออกมามีใจความที่ถูกต้องมากที่สุด

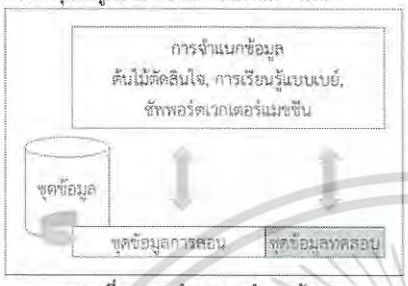
การกำจัดคำที่ไม่มีนัยสำคัญ (Word Removal) เป็นการกำจัดคำที่ไม่สำคัญออกไป โดยที่ไม่ทำให้ใจความในประโยคเสียไป เนื่องจากภาษาไทยมีการใช้คำที่มีความกำกวม

การแปลงข้อมูลเป็นปริภูมิเวกเตอร์ (Vector Space) เพื่อให้สามารถใช้เครื่องมือในการจำแนกประเภทข้อมูลสามารถวิเคราะห์ข้อมูลได้ โดยแปลงข้อมูลให้อยู่ในรูปแบบของเวกเตอร์ และใช้การแปลง TFIDF เพื่อให้ค่าน้ำหนักของคำในเอกสารมาใช้ในการจำแนกประเภทข้อมูล โดยผลของขั้นตอนการเตรียมข้อมูล จะแสดงในส่วนที่ 4 ผลการดำเนินงาน ต่อไป

3.3 การสร้างแบบจำลองการจำแนกข้อมูล

ข้อมูลที่ผ่านการเตรียมเรียบร้อยแล้ว จะใช้วิธีการจำแนกข้อมูล โดยใช้โปรแกรมโอเพ่นซอร์ส Weka ซึ่ง

สนับสนุนการสร้างการเรียนรู้แบบต้นไม้ตัดสินใจ (Decision Tree) การเรียนรู้แบบเบย์ (Naïve Bayes) และการเรียนรู้ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยแบ่งข้อมูลออกเป็น 2 ชุด คือชุดข้อมูลการสอน (Training data) และชุดข้อมูลทดสอบ (Testing data) เพื่อทดสอบชุดข้อมูลนำเข้า ซึ่งแสดงในภาพที่ 4 ดังนี้



ภาพที่ 4: แบบจำลองการจำแนกข้อมูล

การสร้างแบบจำลองการเรียนรู้ทั้งสามแบบ จะมีวิธีการประเมินความถูกต้องโดยใช้วิธีการตรวจสอบไขว้ (Cross Validation) [10] โดยข้อมูลจะถูกแบ่งออกเป็นจำนวน K กลุ่มเท่าๆ กันแบบสุ่ม โดยในการวิจัยนี้ได้กำหนดค่า K เท่ากับ 10 หมายความว่า ข้อมูลจะถูกทดสอบทั้งหมด 10 รอบโดยในแต่ละรอบจะแบ่งข้อมูลออกเป็น 2 ชุด คือชุดข้อมูลการสอน (Training data) และชุดข้อมูลทดสอบ (Testing data) ตัวอย่างเช่น K เท่ากับ 10 รอบที่ 1 จะเก็บข้อมูลกลุ่มที่ 1 ไว้สำหรับเป็นชุดข้อมูลทดสอบ ส่วนข้อมูลกลุ่มที่ 2 ถึง 10 จะนำไปเป็นชุดข้อมูลการสอน วนซ้ำจนครบ 10 รอบ

โดยจะนำผลการประเมินความถูกต้องมาคัดเลือกแบบจำลองการเรียนรู้ที่ให้ผลดีที่สุด เพื่อนำไปหาค่าสหสัมพันธ์ต่อไป โดยผลของขั้นตอนการสร้างแบบจำลองจะแสดงในส่วนที่ 4 ผลการดำเนินงาน ต่อไป

3.4 การหาค่าสหสัมพันธ์

หลังจากที่ได้แบบจำลองการเรียนรู้ จะนำผลของการจำแนกข้อมูล ที่เป็นอารมณ์หุ้น มาหาความสัมพันธ์กับเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของกลุ่มหลักทรัพย์ในแต่ละวัน เพื่อหาว่าอารมณ์หุ้นที่ได้จากการจำแนกข้อมูลข่าวมีผลอย่างไรกับราคาปิด โดยใช้ค่าสหสัมพันธ์ (Correlation) เป็นตัวชี้วัด หากตัวเลขที่ได้มีค่าเข้าใกล้ -1 หรือ 1 แสดงว่าข้อมูลทั้งสอง มีความสัมพันธ์กันอย่างหนาแน่น กล่าวได้ว่าถ้ามีค่าเป็นลบ จะมีความสัมพันธ์ในทิศทางตรงข้ามกัน ถ้ามีค่าเป็นบวก จะมีความสัมพันธ์ในทิศทางเดียวกัน แต่ถ้าตัวเลขที่ได้มีค่าเข้าใกล้ศูนย์ หรือเป็นศูนย์ แสดงว่าข้อมูลทั้งสอง มีความสัมพันธ์กันต่ำ หรือไม่มีความสัมพันธ์กันเลย โดยผลของขั้นตอนการหาค่าสหสัมพันธ์ จะแสดงในส่วนที่ 4 ผลการดำเนินงาน ต่อไป

4. ผลการดำเนินงาน

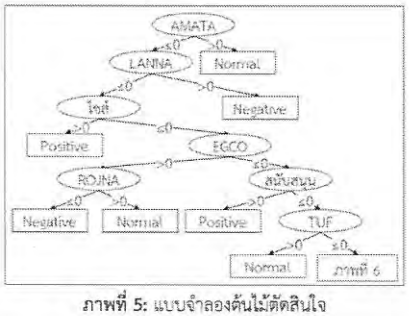
ผลการดำเนินงานในขั้นตอนการเตรียมข้อมูลให้ผลการตัดคำและการกำจัดคำที่ไม่มีนัยสำคัญออก ดังนี้ "แนะ เลือก ซื้อ 11 บจ. อนาคต รุ่ง คัชนี ลุ้น ฟื้นตัว สะสม หุ้น จ่าย ปัน ผล ระหว่าง กาล เช้า นี้ ค่า เงิน บาท อยู่ ที่ 31.09 บาท ต่อ เหรียญสหรัฐ ขณะ ที่ ตลาดหุ้น เอเชีย มี ทั้ง ปรับ ตัว อยู่ ใน แดน ลบ และ แดน บวก คาด หุ้น ไทย ฟื้นตัว แต่ ยัง คง เปราะ บาง และ ผันผวน หาก เป็น นัก ลงทุน รอ รับ ต่ำ กว่า 1400 จุด สะสม หุ้น ที่ จ่าย ปัน ผล ระหว่าง กาล เลือก ซื้อ 11 หุ้น เด่น ได้แก่ BGH CPF MALEE SF SAT TNH TICON DRT TICON TCAP PTTEP"

หลังจากนั้นทำการแปลงข้อมูลที่ผ่านการตัดคำและการกำจัดคำที่ไม่มีนัยสำคัญออกให้เป็นปริภูมิเวกเตอร์ แล้วให้น้ำหนักคำในเอกสารด้วยวิธีการแปลง TFIDF ซึ่งแสดงในตารางที่ 1 ดังนี้

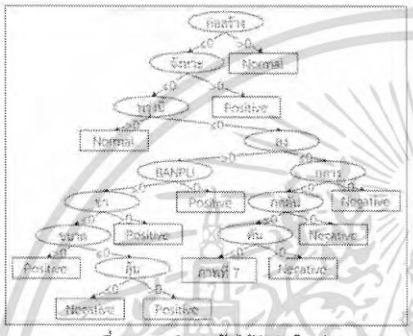
ตารางที่ 1: ตัวอย่างข้อมูลในรูปปริภูมิเวกเตอร์

No.	1: DATE Date	2: SENTIMENT Nominal	3: กลุ่ม Numeric
1	01-04-2013	Normal	0.0
2	02-04-2013	Positive	0.684
3	03-04-2013	Normal	0.684
4	03-04-2013	Normal	0.684
5	04-04-2013	Normal	0.0
6	05-04-2013	Negative	0.684
7	09-04-2013	Normal	0.0

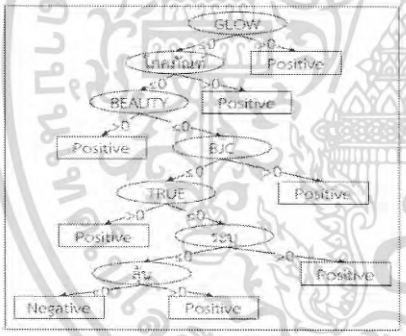
หลังจากนั้นทำการทดสอบการจำแนกข้อมูลด้วยการเรียนรู้ทั้ง 3 แบบ แสดงผลการทดลองเป็นต้นไม้ตัดสินใจจากการเรียนรู้แบบต้นไม้ตัดสินใจ ซึ่งแสดงในภาพที่ 5, 6 และ 7 ดังนี้



ภาพที่ 5: แบบจำลองต้นไม้ตัดสินใจ



ภาพที่ 6: แบบจำลองต้นไม้ตัดสินใจ(ต่อ)

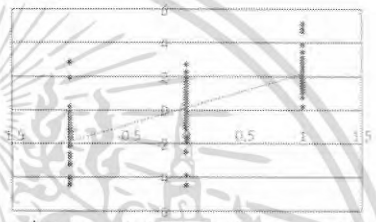


ภาพที่ 7: แบบจำลองต้นไม้ตัดสินใจ(ต่อ)

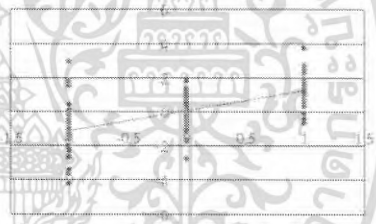
ตารางที่ 2: ผลเปรียบเทียบค่าความถูกต้อง

Model	Precision	Recall	F-Measure
Decision Tree	0.393	0.394	0.392
Naïve Bayes	0.537	0.547	0.529
SVM	0.498	0.504	0.500

จากตารางที่ 2 จะเห็นว่าการเรียนรู้แบบเบย์ ให้ค่าความถูกต้องสูงที่สุดคิดเป็นร้อยละ 54.47 ที่คุณลักษณะ 456 คำ จากนั้นนำอารมณ์หุ้นที่ได้มาคำนวณหาค่าสหสัมพันธ์ กับเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของหุ้น ในกลุ่มปิโตรเคมีและเคมีภัณฑ์ และกลุ่มพลังงาน ซึ่งแสดงในภาพที่ 8 และ 9 ดังนี้



ภาพที่ 8: กราฟสหสัมพันธ์หุ้นกลุ่มปิโตรเคมีและเคมีภัณฑ์



ภาพที่ 9: กราฟสหสัมพันธ์หุ้นกลุ่มพลังงาน

จากภาพที่ 8 และ 9 สรุปค่าสัมประสิทธิสหสัมพันธ์ระหว่างอารมณ์หุ้นของข่าวและเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของหุ้น ในกลุ่มปิโตรเคมีและเคมีภัณฑ์ พบว่าให้ค่า 0.74 และ 0.61 สำหรับหุ้นในกลุ่มพลังงาน จะเห็นได้ว่าค่าสัมประสิทธิสหสัมพันธ์มีค่าเข้าใกล้ 1 แสดงว่าข้อมูลทั้งสอง มีความสัมพันธ์ในทิศทางเดียวกัน

ดังนั้นจากการสร้างโมเดลการจำแนกข้อมูล เพื่อวิเคราะห์อารมณ์ของข่าวหุ้น และหาค่าสหสัมพันธ์

เปรียบเทียบ สรุปได้ว่าสามารถนำไปประกอบการวิเคราะห์ทิศทางของหุ้นได้ จากตัวอย่างของข่าวที่ได้กล่าวไปก่อนหน้านี้มีการสรุปอารมณ์ของข่าวเป็นกลาง ซึ่งเมื่อเปรียบเทียบกับเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของหุ้นในกลุ่มปิโตรเคมีและเคมีภัณฑ์ ในวันเดียวกันพบว่ามีเปอร์เซ็นต์การเปลี่ยนแปลง -0.22 ซึ่งพบว่ามีการเปลี่ยนแปลงเล็กน้อยเป็นไปตามการสรุปอารมณ์ของข่าว สามารถทำให้นักลงทุนวิเคราะห์แนวโน้ม ได้แคบลงและมีความแม่นยำในทิศทางมากขึ้น

5. สรุป

จากการวิเคราะห์อารมณ์ของข่าวภาวะเศรษฐกิจและตลาดหุ้นในประเทศ โดยการทำเหมืองข้อมูลและหาค่าสหสัมพันธ์ สามารถนำไปใช้ประกอบการวิเคราะห์ดูแนวโน้มหรือประกอบการซื้อ ขายหลักทรัพย์ ได้อย่างแม่นยำมากยิ่งขึ้น แม้เปอร์เซ็นต์ของการจำแนกข้อมูลจะยังไม่สูงมาก แต่การหาค่าสหสัมพันธ์พบว่าข้อมูลมีแนวโน้มที่ถูกต้องและเป็นไปในทิศทางเดียวกันกับเปอร์เซ็นต์การเปลี่ยนแปลงราคาปิดของหุ้น จากการทดลองนี้สามารถต่อยอดนำไปใช้ประกอบร่วมกับการวิเคราะห์ข้อมูลทางสถิติ เพื่อความถูกต้องและแม่นยำของการวิเคราะห์หุ้นมากขึ้น

เอกสารอ้างอิง

[1] อัมรินทร์ ก้อนแพง และสมจิตร อาจอินทร์, **การพยากรณ์ราคาข้าวเปลือกโดยใช้เทคนิคการทำเหมืองข้อมูล**, เอกสารประกอบการประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติ ครั้งที่ 23, มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน, 23-24 ธันวาคม 2554. หน้า 82-87.

[2] อรอุมา นองเนือง และณัฐวี อุตกลษฎ์, **ระบบช่วยวิเคราะห์บริการทางการเงินเพื่อกลุ่มลูกค้านิติบุคคลกรณีศึกษาธนาคารกสิกรไทย**, เอกสารประกอบการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 6, โรงแรมอิสติน, 3-5 มิถุนายน 2553. หน้า 970-975.

[3] Kushal Dave, Steve Lawrence and David M. Pennock "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proceeding of the 12th International Conference on World Wide Web, Budapest, Hungary, 20-24 May 2003. pp.519-528.

[4] พัชรนิกานต์ พงษ์ชนู, วรารัตน์ รุ่งวรวุฒิ, งามนิจ อาจอินทร์ และสมจิตร อาจอินทร์, **วิเคราะห์ความพึงพอใจของลูกค้าจากข้อความแนะนำโดยการทำเหมืองความคิดเห็น**, เอกสารประกอบการประชุมวิชาการ Knowledge and Smart Technologies ครั้งที่ 5, คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา, 31 มกราคม-1 กุมภาพันธ์ 2556. หน้า 53-60.

[5] จิโรจน์ ภาคคีรี และกาญจนา วิริยะพันธ์, **การวิเคราะห์รูปแบบการบุกรุกข้อมูลบนเครือข่าย โดยใช้เทคนิคดาต้าไมน์นิง**, วารสารเทคโนโลยีสารสนเทศ, ปีที่ 3, ฉบับที่ 6, กรกฎาคม-ธันวาคม 2550. หน้า 40-46.

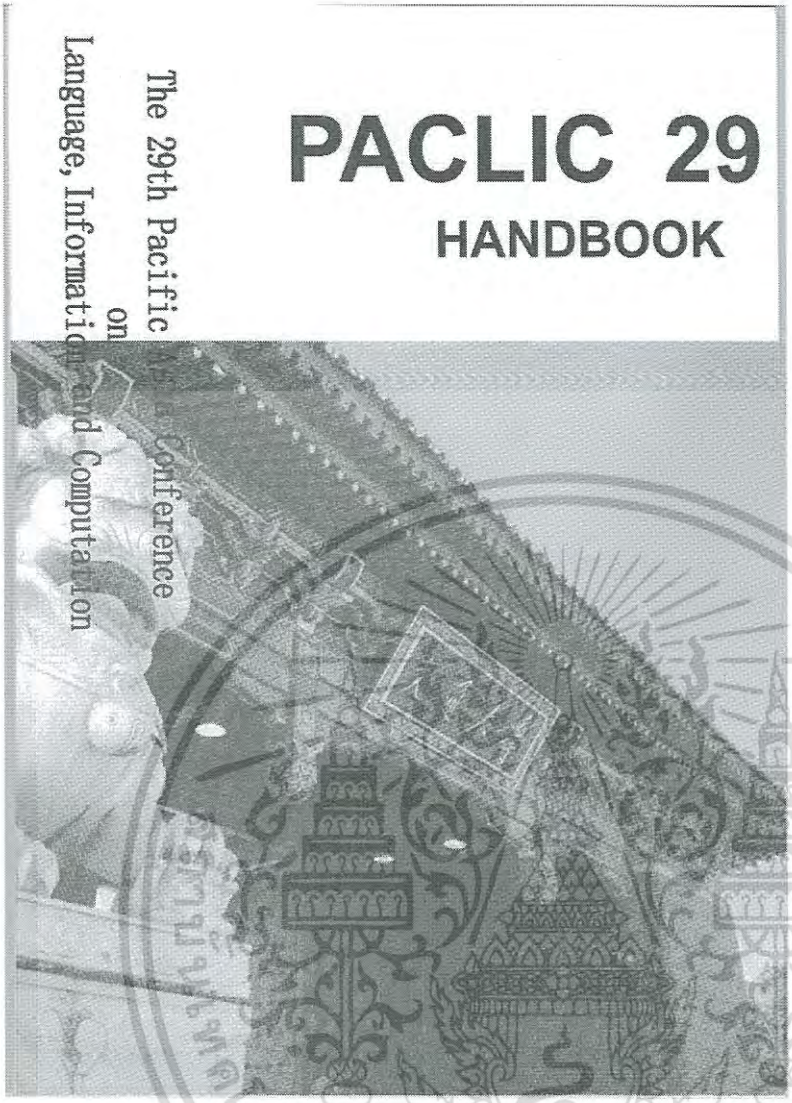
[6] นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ, **การจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน**, เอกสารประกอบการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 6, โรงแรมอิสติน, 3-5 มิถุนายน 2553. หน้า 89-92.

[7] นิเวศ จิระวิชิตชัย และนรินทร์ พนาวาส, **"Sentiment Classification Using Machine Learning Techniques."** Proceeding of 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE), Mahidol University, Nakhon Pathom, 11-13 May 2011.

[8] ข่าวหุ้นธุรกิจออนไลน์, **ข่าวภาวะเศรษฐกิจและตลาดหุ้นในประเทศ**. [Online]. Available: www.kaohoon.com [2013, September 30]

[9] อังศุมาลี สุทธภักดิ, **การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเบย์และซัพพอร์ตเวคเตอร์แมชชีน**. [Online]. Available: www.rc.ac.th/Library_web/doc/RC_RR/2553_ComBus_Angsumalee.pdf [2013, October 1]

[10] กรัณยา อำมฤครัตน์ และพยุง มีสัจ, **การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของโรคลมร้อนด้วยวิธีการทางเครือข่ายประสาทเทียม**, เอกสารประกอบการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 6, โรงแรมอิสติน, 3-5 มิถุนายน 2553. หน้า 116-121.



PACLIC 29

Thai Stock News Sentiment Classification using Wordpair Features

Apinan Chattupan
Knowledge Management and Knowledge Engineering Laboratory (KMAKE Lab)
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
S7606151@kmitl.ac.th

Ponrudee Netisopakul
Knowledge Management and Knowledge Engineering Laboratory (KMAKE Lab)
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
ponrudee@it.kmitl.ac.th

Abstract

Thai stock brokers issue daily stock news for their customers. One broker labels these news with plus, minus and zero sign to indicate the type of recommendation. This paper proposed to classify Thai stock news by extracting important texts from the news. The extracted text is in a form of a 'wordpair'. Three wordpair sets, manual wordpairs extraction (ME), manual wordpairs addition (MA), and automate wordpairs combination (AC), are constructed and compared for their precision, recall and f-measure. Using this broker's news as a training set and unseen stock news from other brokers as a testing set, the experiment shows that all three sets have similar results for the training set but the second and the third set have better classification results in classifying stock news from unseen brokers.

Keywords: Thai stock news, sentiment classification, text classification, wordpair features.

1 Introduction

Thai stock news are daily issued from many stock brokers. Thai stock news is an important source of information for stock traders to make a decision on stock trading. However, a usual Thai stock news has a long message and sometimes not easily to interpret or conclude. One stock broker makes it easier by labeling each news with plus (+), minus (-) and zero (0) sign, to indicate the type of news as positive, negative and neutral. This automatically classifies the news into three classes.

In this research, we assume that 'features' that can be used for classifying the news must be presented as text in the news. Although this assumption could be too strong in general, for our sole purpose of investigation, our focus here is on text form of the news. Therefore, we proposed to construct a set of these 'texts' to be used as features in order to classify Thai stock news into three sentiments: positive, negative and neutral classes, using known sentiment news as a training set and unseen news as a testing set.

Each feature is called a *wordpair*, since it is a pair of a keyword and a polarity word. A keyword is a word that signifies upcoming information. A polarity word is a word associated with a keyword and signifies a sentiment that related to the keyword. Following the classification from one broker, there are three sentiments: positive, negative, and neutral.

However, due to the flexibility of Thai language, the order of a keyword, a polarity word and a stock symbol may not be the same in the news. That is - a keyword may come before or after a polarity word. In addition, they may come before, after or between the stock symbols they intend to recommend.

29th Pacific Asia Conference on Language, Information and Computation pages 188 - 195
Shanghai, China, October 30 - November 1, 2015

There are two objectives of this paper. First, describing methods to construct these wordpairs collection. Second, utilizing them for constructing automatic classification models for classifying Thai stock news into three corresponding classes. This method can be very useful for general investors because investors can quickly obtain the information and make a decision in stock trading by following the trend of Thai stock news from the classification model.

The outline of this paper is as follows. Section 2 reviews related work. Section 3 describes stock news collection and wordpair construction. In subsection 3.1, we show an example stock news, signs, and stock symbols and compare the frequency of stock symbols in the training and testing set. In subsection 3.2, we propose three sets of wordpairs, which are used to classify Thai stock news sentiments. Section 4 gives details of two experimental designs. We also discuss an effect of varying window sizes for extracting wordpairs features. The results of stock news sentiment classification are also shown in this section. Section 5 analyzes misclassified stock news from the testing set. The last section gives a conclusion and our plan for imminent future work.

2 Related Work

There are two involving areas related to our work. First, the research involved language structure and processing research. Second, the research involved analyzing the stock and classification.

Tongtep and Theeramunkong (2010) mentioned the structural model for extracting patterns from Thai news documents. They focus on a pattern of unique name or noun such as person name, organization name, location, date and time. In addition, Sutheebanjard and Premchaiswadi (2010); Lertcheve and Aroonmanakun (2009) mentioned a similar extracting pattern. They extracted only the person name and only the product name respectively.

Taboada et al. (2011); Lertsuksakda et al. (2014) mentioned the types of word; such as a noun, a verb, an adjective and adverb; and polarity of the word. Taboada et al. (2011) discussed the types of word that have an emotional level and the negation of word that will affect to an emotional level. Lertsuksakda et al. (2014) discussed Thai sentiment terms by using the hourglass of emotion. They assigned an emotional level of a Thai word using two-way translation from English word corpus to Thai word. These techniques will be applied to extract wordpairs from Thai stock news.

Mittermayer (2004); Schumaker and Chen (2009); Chattupan and Netisopakul (2014) demonstrated stock trends prediction using text in the stock news. In addition, Lertsuksakda et al. (2015) discussed text mining techniques in Thai children stories. We will take above techniques and apply them to our work.

3 Stock News and Wordpairs

Section 3.1 describes data preparation including stock news collection and preliminary analysis. Section 3.2 describes wordpairs construction for stock news sentiment classification experiments.

3.1 Stock News Collection

The experiment collects Thai stock news from several brokers such as Bualuang securities (BLS), Thanachart securities (TNS), Krungsri securities (KSS), and so on. In this paper, we tag wordpairs from BLS stock news, hence, we use news from this set as a training set. Stock news from other brokers are combined and used as a testing set. Another important reason for using BLS as a training set is that the broker recommendation includes sentiment signs, such as +, 0, -, *. Therefore, wordpairs sentiments from this broker can be directly tagged using the sign sentiment. The example of stock news published by BLS with their signs are shown in Table 1. Note that some news contain more than one stock symbols.

Thai stock news from BLS (Bualuang Securities, 2015) was collected between 04/04/2014 to 27/05/2015. There are 1,381 stock news with totally 6,596 paragraph news containing stock symbols. Paragraph length is from 200 to 500 letters. Furthermore, Thai stock news under investigation was selected with the following condition. First, the selected stock news must have a stock symbol. This is used for obtaining stock statistics; percent price changes and trading volume. Second, stock news must contain at least one wordpairs. Hence, the stock news contains only a stock symbol and recommended price will not be selected.

Date	Stock news	Sign	Symbol
17/02/2015	MAKRO กำไรดีกว่าคาด คงกำไรปีนี้ / คงคำแนะนำถือ makro-kum-rai-dee-kwa-kard-kong-kum-rai-pee-nee-/-kong-kum-nae-num-thux 'Makro earn better than expected. This year continued profit. / Recommend hold.'	0	MAKRO
06/10/2014	กลุ่มโรงแรมท่องเที่ยวรายงาน... เรายังคงแนะนำซื้อ MINT CENTEL และ ERW koom-rong-ram-tong-tiew-rai-ngan-...-rao-young-kong-nae-num-shux-mint-centel-lae-erw 'The hotel and travel report... We recommend buy Mint, Centel and Erw.'	+	MINT CENTEL ERW
17/02/2015	เช้านี้เกาหลีใต้ คงดอกเบี้ย 2% Shao-nee-kao-lee-tai-kong-dok-bia-song-per-cent 'This morning, south korea fixed interest rate 2%.'	+	No

Table 1: The example of Thai stock news published by BLS

Thai stock news from other brokers (Stock News Online, 2015) were collected between 10/03/2015 to 03/07/2015 and has totally 3,489 paragraphs news. We used the same selecting criteria as the training set. However, we notice that this set has longer paragraph length from 500 to 1000 letters. We prepare this testing set from unseen data set in order to support for future unseen stock news.

Industry	BLS		Other Brokers	
	Freq.	Rank	Freq.	Rank
Agro	780	3	303	5
Consump	310	8	139	8
Fincial	514	5	345	3
Indus	359	7	189	7
Propcon	2512	1	1356	1
Resourc	629	4	265	6
Service	1094	2	562	2
Tech	398	6	330	4

Table 2: Comparing frequency of stock symbols grouped by types of industry in training and testing set

Table 2 shows the preliminary investigation of stock symbols frequencies grouped by types of industry, comparing the training set (BLS) and the testing set (other brokers). Notice that for both sets, the most and second most frequent stock symbols are in industry: Propcon and Service; while the least and second least frequent symbols are in industry: Consump and Indus, respectively. The total number of unique symbols in the training set is 296 comparing to 219 in the testing set. Hence, the average mentioned frequency for each symbol in the training and testing set are 51.74% and 38.28%, respectively. The total numbers of symbols grouped by types of industry are shown in Table 3.

Industry	Unique symbols in SET	Unique symbols in training set	Unique symbols in testing set
Agro	54	32	29
Consump	42	12	8
Fincial	61	32	27
Indus	87	32	17
Propcon	152	76	53
Resourc	36	28	20
Service	99	57	41
Tech	41	27	23
Summary	572	296	219

Table 3: The total number of unique symbols grouped by types of industry in SET, training set, and testing set

3.2 Wordpairs Construction

This paper proposed to use stock sentiment wordpairs to classify stock news into the positive, negative and neutral news. A wordpairs is a tuple of size 3, consists of a keyword, a polarity word and a sentiment. A keyword usually is a noun or a verb indicating characteristics of stock or business, such as *"profit, recommend, income, price, growth ..."* and so on. A polarity word is a verb or an

adjective or an adverb for the keyword above, such as *"good profit, recommend buy, steady income ..."* and so on. In this paper, we have only three sentiments: positive (1) if the news has + sign, negative (-1) if the news has – sign, and neutral (0) if the news has 0 sign. The examples of wordpairs are shown in Table 4.

Keyword	Polarity word	Sentiment
กำไร kum-rai 'profit'	ดี dee 'good'	+
คาด kard 'forecast'	กำไร kum-rai 'profit'	+
ปัจจัย pud-jai 'factor'	หนุน nun 'support'	+
แนะนำ nae-num 'recommend'	ถือ thux 'hold'	0
รายได้ rai-dai 'income'	ทรงตัว song-tua 'steady'	0
ราคา ra-ka 'price'	พักตัว puk-tua 'dormancy'	0
ผลกระทบ pon-kra-tob 'effect'	เชิงลบ chung-lop 'negative'	-
เศรษฐกิจ set-ta-kit 'economy'	ชะลอ cha-loor 'slow down'	-
ราคาหุ้น ra-ka-hoon 'stock price'	เสี่ยง sciyng 'risky'	-

Table 4: An example of wordpairs with a pattern {keyword, polarity word, sentiment}

We obtain the first set of wordpairs by hand – called *manual extraction* set (ME). Next, we add new wordpairs into the first set by duplicating the same keyword augmented with an opposite and a neutral polarity words. The second set is called manual wordpairs addition set or *manual addition* (MA), for short. For example, a keyword and polarity word ราคาหุ้น,ขึ้น ra-ka-hoon-,-khun 'stock price, up', the negation polarity word ราคาหุ้น,ลง ra-ka-hoon-,-lng 'stock price, down' and a neutral polarity word ราคาหุ้น,คงที่ ra-ka-hoon-,-kong-thi 'stock price, unchanged'. The MA set has only the positive and negative sense and does not have the neutral sense of 'stock price'. Therefore, this sense is added in the second set. Other examples are shown in Table 5.

The third set of wordpairs is automatically generated from the second set. Wordpairs with partial common keywords are assigned with the same polarity words and signs. For instance, the keywords ราคาหุ้น ra-ka-hoon 'stock price' and ราคา ra-ka 'price' have a common word 'ราคา – stock price'; hence, they will share the same set of polarity words and sentiments.

Keyword	Polarity word	Sentiment
กำไร kum-rai 'profit'	ทรงตัว song-tua 'settled'	0
	ขึ้น khun 'up'	+
	ลง lng 'down'	-
การลงทุน karn-lng-thun 'investment'	ฟื้นตัว fun-taw 'recover'	+
	ซบเซา sob-sea 'stagnant'	-
	คงที่ khong-thi 'stable'	0
แนะนำ nae-num 'recommend'	ขาย khai 'sell'	-
	ซื้อ sux 'buy'	+
	ถือ thux 'hold'	0

Table 5: The manual wordpairs addition with opposite and neutral polarity words

Keyword	Polarity word	Existing wordpairs	New wordpairs
ราคาหุ้น ra-ka-hoon 'stock price'	ขึ้น khun 'up'	ราคาหุ้น, ขึ้น ra-ka-hoon-,-khun 'stock price, up'	ราคา, ขึ้น ra-ka-,-khun 'price, up'
ราคาหุ้น ra-ka-hoon 'stock, price'	บวก bawk 'positive'	ราคาหุ้น, บวก ra-ka-hoon-,-bawk 'stock price, positive'	ราคา, บวก ra-ka-,-bawk 'price, postive'
ราคา ra-ka 'price'	ปรับลด prub-rod 'diluted'	ราคา, ปรับลด ra-ka-,-prub-rod 'price, diluted'	ราคาหุ้น, ปรับลด ra-ka-hoon-,-prub-rod 'stock price, diluted'

Table 6: The automate wordpairs combination

Examples of these crossovers are shown in Table 6. We called the third set automate wordpairs combination or *automate combination*

(AC), for short. The numbers of wordpairs for the three sets are 133, 277 and 331 respectively.

4 Experimental Design and Result

We hypothesized that wordpairs affecting the stock news sentiment can be located near the stock symbol. Hence, we design experiments, using the stock symbol as a center, with different window sizes varying from 20, 40, 60 and 80 letters. Figure 1 demonstrates effects of varying window sizes when extracting wordpairs features.

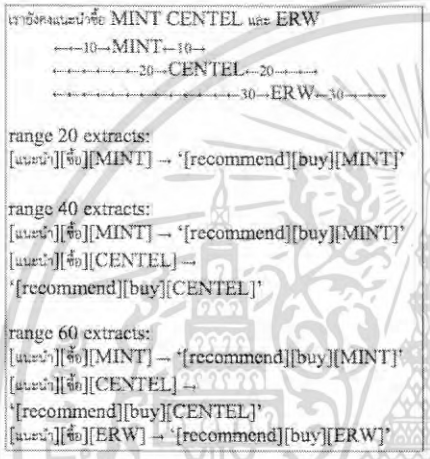


Figure 1: Effects of varying window sizes for extracting wordpairs features

We found that the optimal window size for extracting wordpairs features is 60, with average 1-3 wordpairs for each stock symbol, as shown in Figure 2. We also found that 80-letter window size sometimes extracts irrelevant wordpairs, such as wordpairs of the next symbol. For example, the stock news on Figure 2 has a stock symbol (1) and three wordpairs. Notice that a keyword and a polarity word of the second wordpairs (3) (forecast and profit) are not immediately adjacent to each other.

[1]MAKRO กำไร[2]ดี[3]กว่าคาด[3] คงกำไร[3]ปีนี้ / คงคำแนะนำ[4] ซื้อ[4]	
[1]MAKRO	→ symbol
[2]กำไร, ดี kum-rai-dee 'profit, good'	→ wordpairs
[3]คาด, กำไร kard-kum-rai 'forecast profit'	→ wordpairs
[4]แนะนำ, ซื้อ nae-num-thux 'recommend, hold'	→ wordpairs

Figure 2: An example of wordpair features for a stock

In short, there are 6 combinations of {symbol (S), keyword (K), polarity (P)} as shown in Table 7. In the training set, the most frequent patterns found is pattern#3 (K-P-S) with 335, 387, and 387 occurrences when extracted by ME, MA and AC set, respectively. The second most frequent is pattern#5 (P-K-S) with 215, 254, and 257 occurrences when extracted by ME, MA and AC set, respectively. The least frequent pattern found is pattern#4 with 47, 55 and 57 occurrences.

The most and the second most frequent patterns in the testing set found are similar to the training set with 599, 674, and 675 occurrences in pattern#3 and 577, 669, 669 occurrences in pattern#5.

There are two main experiments. The first experiment is designed to examine the effects of the three sets of wordpairs, manual extraction (ME), manual addition (MA) and automate combination (AC), as described in section 3 in training and testing set.

The second experiment is designed to examine the effects of S-K-P patterns on training and testing set. We use an open source software 'Weka' (Weka, 2013) to build classification model using decision tree and support vector machine.

For the first experiment, the results of decision tree and SVM classification models using ME, MA and AC wordpairs sets as features are shown in Table 8. We found that, for the training set, there are no significant differences for all three sets of wordpairs features.

Pattern#	{Symbol, keyword, polarity} combinations
1: SKP	[RATCH][แนะนำ] [ซื้อ] ราคาเป้าหมาย 64 บาท… [Symbol] [Keyword][Polarity] 'RATCH hold with a target price of 64 Thai baht…'
2: SPK	[SAMTEL] SAT [กำไร] ตาม [คาด]… [Symbol] [Polarity][Keyword] 'SAMTEL and SAT profit as expected…'
3: KSP	รายงานกลุ่ม Small Cap [คาด] [UNIQ] และ TRC มีโอกาสทำ [กำไร]… [Keyword][Symbol] [Polarity] 'small cap segment reported UNIQ and TRC are expected profitable opportunities…'
4: KPS	คงคำ [แนะนำ] [ซื้อ] [TP] 24 บาท… [Keyword][Polarity][Symbol] 'maintain buy with TP 24 Thai baht…'
5: PSK	เรา[ลด] [ราคา]เป้าหมาย และคำแนะนำ [SIM] และ SAMTEL ลงเหลือถือจากซื้อ… [Polarity] [Keyword] [Symbol] 'we lower our target price and recommendation SIM and SAMTEL to hold from buy'
6: PKS	เดือน สค. พบว่า[ปรับดีขึ้น] 0.5% และ +4.3% โดย [BBL] BAY TMB KBANK SCB รายงาน[สินเชื่อ]เติบโต [Polarity] [Symbol] [Keyword] 'In august, showed a rise of 0.5% and +4.3% BBL BAY TMB KBANK SCB loan growth…'

Table 7: The types of structure matching with the real Thai stock news

The first set – ME, gives a slightly better precision of 0.741 for decision tree model comparing to MA and AC with the precision of 0.722 and 0.721. The same is hold for SVM model, where ME give a slightly better precision result of 0.723 comparing to 0.712 for MA and AC. Every model has the same recall and F-measure of approximately 0.75 and 0.66 respectively. This is not surprising because wordpairs in the first set – ME – are extracted from the training set. Therefore, the first set of wordpairs, when use to classify the training set, should be the most accurate features.

Set	Decision tree		SVM	
ME	Precision	0.741	Precision	0.723
	Recall	0.758	Recall	0.755
	F-Measure	0.669	F-Measure	0.663
MA	Precision	0.722	Precision	0.712
	Recall	0.757	Recall	0.755
	F-Measure	0.669	F-Measure	0.665
AC	Precision	0.721	Precision	0.712
	Recall	0.757	Recall	0.754
	F-Measure	0.669	F-Measure	0.664

Table 8: The result of decision tree and SVM classification of each pattern

The resulting decision tree is partially shown in Figure 3.

Since the precision results of decision tree are better than SVM, we use these training models to classify the testing set. We found that, from 3,489 stock news (# of rows) in the testing set, ME, MA, and AC models predict the same outcome for 3,457 rows or 99.08%; predict the same outcomes two out of three times for 32 rows or 0.91%. There is no totally different outcome prediction – all three models predict different outcomes 0%.

We assume that if three classification models predict the same outcome, then there is no need to verify more. Now, we examine the 32 rows (same two out of three) by comparing the classification models outcomes with solutions provided by a human. The majority outcomes (two same outcomes) is correct for 12 times; while the minority outcome (one out of three) is correct for 16 times. The leftover 4 rows are those outcomes not matched to the solutions. These will be analyzed in section 5.

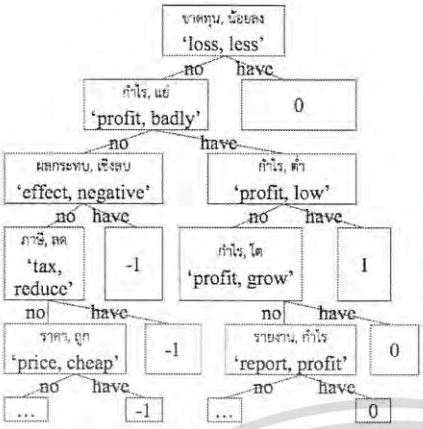


Figure 3: Partial decision tree model built from ME wordpairs features

Wordpair Features		ME			MA			AC		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Decision tree	P#1:SKP	0.68	0.69	0.66	0.77	0.80	0.77	0.77	0.80	0.77
	P#2:SPK	0.93	0.92	0.91	0.90	0.91	0.90	0.90	0.91	0.90
	P#3:KSP	0.63	0.64	0.62	0.58	0.61	0.58	0.58	0.61	0.58
	P#4:KPS	0.75	0.76	0.74	0.74	0.75	0.72	0.74	0.75	0.72
	P#5:PSK	0.65	0.74	0.69	0.63	0.69	0.66	0.68	0.71	0.69
	P#6:PKS	0.75	0.76	0.75	0.79	0.80	0.79	0.76	0.77	0.76
	average	**0.73**	**0.75**	**0.73**	**0.74**	**0.76**	**0.74**	**0.74**	**0.76**	**0.74**
SVM	P#1:SKP	0.70	0.71	0.70	0.81	0.83	0.81	0.79	0.81	0.80
	P#2:SPK	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
	P#3:KSP	0.70	0.71	0.70	0.69	0.70	0.70	0.69	0.70	0.69
	P#4:KPS	0.70	0.71	0.70	0.71	0.73	0.71	0.71	0.73	0.72
	P#5:PSK	0.77	0.78	0.77	0.72	0.72	0.72	0.75	0.77	0.76
	P#6:PKS	0.69	0.71	0.69	0.74	0.76	0.75	0.73	0.75	0.74
	average	**0.75**	**0.76**	**0.75**	**0.77**	**0.78**	**0.77**	**0.77**	**0.79**	**0.78**

Table 9: The result of decision tree and SVM classification of each pattern

For the second experiment, the result is shown in Table 9. Overall, the SVM models give a slightly better results than the decision tree models for all three wordpairs sets. Comparing among the SVM models, the first set – ME, gives the least average recall of 0.76. The second set – MA and the last set – AC give the second best average result of 0.78 and the best average result of 0.79, respectively. However, the result of pattern#2 (S-P-K) obtains very high results above 0.90 for all models. We will discuss insights for this pattern in the next section. The second best result is pattern#1 (S-K-P). It obtains recall of 0.83 and 0.81 when uses with MA and AC set respectively. The other patterns give the similar average results of 0.7 in SVM models.

5 Error Analysis and Discussion

In the first experiment, there are 4 stock news where no machine classification matches the human solutions. The investigation found that these 4 news appear during adjacent business days and have exact same text messages as shown in Figure 4.

กลุ่มรับเหมา... คาดกำไรไตรมาส 2/58 เติบโตดีอย่าง PTTGC	
koom-rab-mao-…-kard-kum-rai-nai-tri-mart-song-/-ha-sib-pad-teib-to-dee-yang-pttgc 'Contractor group… Earning expected in the quarter 2/58 growth as PTTGC'	
The wordpairs should be extracted.	เติบโต, ดี teib-to-dee 'growth, good'
The wordpairs is in the ME, MA, and AC sets.	การเติบโต, ดี karn-teib-to-dee 'growth, good'

Figure 4: Error analysis for stock news classification

From the figure 4, the wordpairs เติบโต, ดี teib-to-dee 'growth, good' and การเติบโต, ดี karn-teib-to-dee 'growth, good' have the same meaning. However, in all three wordpair sets, there is no เติบโต, ดี teib-to-dee 'growth, good' as a wordpair feature. Hence, the keyword is not extracted as a feature. The type of เติบโต tieb-to 'growth' is a verb, but การเติบโต karn-tieb-to 'growth' is a noun. In Thai language, the addition prefix of the word การ karn or ความ kwarm will change a type of a word from a verb to a noun. This investigation suggests that this factor should be considered in order to construct a better set of wordpair features for Thai language in the future.

In the second experiment, pattern#2 (S-P-K) has the highest precision, recall and f-measure. An investigation of the training set for pattern#2 (S-P-K) found that the set has only two sentiments: positive (1) and negative (-1). The training set contains no neutral sentiment. Therefore, there will be only two classes of classification results instead of three. The results suggest that the stock news sentiments classification for this pattern – (S-P-K)

can be performed with more accuracy than other patterns because it has only two polarities instead of three polarities.

6 Conclusion and Future Work

This paper proposes to classify stock news into three classes: positive, negative and neutral using only text in the news called wordpairs. Three sets of wordpairs are constructed. The first set is manually extracted from 1,381 stock news. It contains 133 wordpairs. The second set is manually added with opposite and neutral polarities and contains 277 wordpairs. The third set is automatically generated using partial keyword combined with existing polarities and contains 331 wordpairs.

Two experiments are conducted to test the effects of three wordpair sets. The result found no significant differences in the training set but found slightly improvement, for the second the third wordpair sets, when they are applied to unseen stock news (a testing set) from other brokers. Moreover, we found six combination patterns of a stock symbol, a keyword and a polarity (S-K-P) in stock news. The result from the second experiment shows that some pattern (S-P-K) has only two polarities, instead of three and therefore achieved the highest correct classification results.

For the future work, we will resolve the problem found and discussed in the error analysis section by consider adding and deleting a keyword's prefix.

References

Apinan Chattupan and Ponrudee Netisopakul. 2014. Stock sentiment analysis model using data mining (In Thai). In Knowledge and Smart Technology (KST), 2014. Proceeding of 6th National Conference on, Chonburi, Thailand.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2):267-307.

Marc-André Mittermayer. 2004. Forecasting intraday stock price trends with text mining techniques. In System Sciences, 2004. Proceeding of the 37th Annual Hawaii International Conference on. IEEE.

Nattadaporn Lertcheva and Wirote Aroonmanakun. 2009. A linguistic study of product names in Thai economic news. In Natural Language Processing, 2009. SNLP'09. Eight International Symposium on, 26-29. IEEE.

Nattapong Tongtep and Thanaruk Theeramunkong. 2010. Pattern-based extraction of named entities in thai news documents. Thammasat International Journal of Science and Technology, 15(1):70-81.

Phaisarn Sutheebanjard and Wichian Premchaiswadi. 2010. Disambiguation of Thai personal name from online news articles. In Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, 3:302-306. IEEE.

Rathawut Lertsuksakda, Kitsuchart Pasupa and Ponrudee Netisopakul. 2015. Sentiment analysis of Thai children stories on support vector machine. In Artificial Life and Robotics (AROB), 2015. Proceeding of the Twentieth International Symposium on. Beppu, Japan.

Rathawut Lertsuksakda, Ponrudee Netisopakul and Kitsuchart Pasupa. 2014. Thai sentiment terms construction using the Hourglass of Emotions. In Knowledge and Smart Technology (KST), 2014 6th International Conference on, 46-50. IEEE.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information System (TOIS), 27(2).

Bualuang Securities. Retrieved February 15, 2015. http://www.bualuang.co.th/th/index.php

Stock News Online. Retrieved May 15, 2015. http://www.kaohoon.com/online/content/category/13/ภาวะเศรษฐกิจและตลาดหุ้นในประเทศ

Weka 3.7.1. Retrieved October 1, 2013. http://www.cs.waikato.ac.nz/ml/weka

ประวัติผู้เขียน

ผู้เขียน	นาย อภินันท์ จัตตุพันธ์
วันเดือนปีเกิด	25 มกราคม 2535
สถานที่เกิด	จังหวัด กรุงเทพมหานครฯ
ปริญญา	2556 วิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
	2558 วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
งานวิจัยที่ตีพิมพ์	Apinan Chattupan, and Ponrudee Netisopakul. "Stock sentiment analysis model using data mining (In Thai)." *In Knowledge and Smart Technology (KST), 2014. Proceeding of 6th National Conference on.* Chonburi, Thailand.
	Apinan Chattupan, and Ponrudee Netisopakul. "Thai Stock News Sentiment Classification using Wordpair Features." *In Proceeding of 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29.* pp. 188-195. Shanghai, China. Oct. 30 – Nov. 1, 2015.