การวิเคราะห์ความรู้สึกผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา

นันทภัค สุทธิเลิศ

งานนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
พฤษภาคม 2560

คณะกรรมการควบคุมงานนิพนธ์และคณะกรรมการสอบงานนิพนธ์ได้พิจารณางานนิพนธ์ของ นางสาวนันทภัค สุทธิเลิศ ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมงานนิพนธ์

.....อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐนนท์ ลีลาตระกูล)

คณะกรรมการสอบงานนิพนธ์

.....ประธานกรรมการ
(ดร.ภารุจ รัตนวรพันธุ์)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุนิสา ริมเจริญ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ดร.ณัฐนนท์ ลีลาตระกูล)

คณะวิทยาการสารสนเทศ อนุมัติให้รับงานนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพา

..... คณบดีคณะวิทยาการสารสนเทศ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)
วันที่ 31 เดือน พฤษภาคม พ.ศ. 2560

กิตติกรรมประกาศ

งานนิพนธ์นี้ประสบความสำเร็จและลุล่วงไปด้วยนี้ เนื่องจากได้รับคำแนะนำและความช่วยเหลือเป็นอย่างดีจากบุคคลเหล่านี้มาโดยตลอด ผู้จัดทำงานนิพนธ์จึงขอขอบพระคุณบุคคลดังต่อไปนี้

อาจารย์ณัฐนนท์ ลีลาตระกูล ซึ่งเป็นอาจารย์ที่ปรึกษาที่ให้โอกาสได้จัดทำงานนิพนธ์เล่มนี้ และเสียสละเวลาอันมีค่าให้ความรู้ ชี้แนะแนวทางอันก่อให้เกิดความสำเร็จ และให้ความช่วยเหลือเป็นอย่างดีเสมอมา

อาจารย์เอกสิทธิ์ พัชรวงศ์ศักดา อาจารย์ประจำสาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ให้ความรู้การใช้งานโปรแกรม RapidMiner และงานที่เกี่ยวข้อง รวมถึงให้คำปรึกษาและช่วยเหลือทำให้ผู้จัดทำงานนิพนธ์นำความรู้ที่ได้มาใช้ประโยชน์ในงานนิพนธ์นี้

เพื่อน ๆ พี่ ๆ น้อง ๆ สาขาเทคโนโลยีสารสนเทศ (ป.โท) รุ่น 10 คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ที่ให้ความช่วยเหลือ ให้กำลังใจ และดูแลดั่งเป็นครอบครัวเดียวกัน

และสุดท้ายนี้ผู้จัดทำงานนิพนธ์ขอกราบขอบพระคุณบิดาและครอบครัว ที่ให้ความรักและกำลังใจ สนับสนุน ส่งเสริม ผลักดันจนงานนิพนธ์สำเร็จลุล่วงไปด้วยดี

นันทภัค สุทธิเลิศ

57920638: สาขาวิชา: เทคโนโลยีสารสนเทศ; วท.ม. (เทคโนโลยีสารสนเทศ)
คำสำคัญ: Sentiment Analysis / Opinion mining

นันทภัค สุทธิเลิศ: การวิเคราะห์ความรู้สึกผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา (Sentiment Analysis for US Airline Passengers) คณะกรรมการควบคุมงานนิพนธ์: ณัฐนนท์ ลีลาตระกูล, Ph.D., 39 หน้า. ปี พ.ศ. 2560.

งานนิพนธ์นี้มีวัตถุประสงค์เพื่อสร้างกระบวนการวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็น (Sentiment Analysis) ของผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา เพื่อจำแนกความคิดเห็นเชิงลบ (Negative), ความคิดเห็นเชิงบวก (Positive) และความคิดเห็นเป็นกลาง (Neutral) โดยข้อมูลที่ผ่านการวิเคราะห์จะถูกนำมาใช้สำหรับปรับปรุงการให้บริการสายการบิน

หลังจากได้แบบจำลอง (หรือโมเดลการวิเคราะห์) แบบจำลองนี้จะช่วยบริษัทลดระยะเวลาการวิเคราะห์ความคิดเห็นของผู้โดยสารอื่น ๆ ภายหลังได้ โดยงานนิพนธ์นี้ได้ศึกษาและสร้างแบบจำลองสำหรับใช้วิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็นผู้ใช้บริการสายการบินประเทศสหรัฐอเมริกา ด้วยข้อมูลจาก http:// www.kaggle.com ซึ่งเป็นข้อมูลเดือนกุมภาพันธ์ 2015 มีจำนวนข้อมูลทั้งสิ้น 14,640 ข้อความ เพื่อนำแบบจำลองที่สร้างขึ้นจำแนกข้อมูลแสดงความคิดเห็นที่เกิดขึ้นใหม่

นอกจากนี้ หลังจากบริษัททราบว่าลูกค้ามีความรู้สึกอย่างไรต่อสินค้าหรือบริการ หากลูกค้ามีความคิดเห็นเชิงลบ (Negative) มากกว่าความคิดเห็นเชิงบวก (Positive) บริษัทก็สามารถปรับปรุงแก้ไขได้ทันเวลา และยังนำผลจากการวิเคราะห์มาใช้สำหรับกำหนดกลยุทธ์เพื่อสร้างความพึงพอใจเพื่อใช้ขับเคลื่อนธุรกิจได้ด้วยระยะเวลาน้อยลง

57920638: MAJOR : INFORMATION TECHNOLOGY ; M.Sc.
(INFORMATION TECHNOLOGY)
คำสำคัญ: SENTIMENT ANALYSIS / OPINION MINING

NUNTAPAK SUTTHILERT : US AIRLINE SENTIMENT ANALYSIS
ADVISORY COMMITTEE : NUTTHANON LEELATHAKUL, Ph.D., 39 P. 2017.

This research is to analyze the sentiments of US airlines' passengers. We attempt to classify their Twitter comments (into negative, positive and neutral comments). The derived analyzing model could help decrease the time needed for analyzing any new comments in order to capture negative comments in time to improve the airlines' services. We obtained the US airline Twitter data from www.kaggle.com, which is the information in February 2015 and consists of 14,640 records in total.

Using our derived model, the airlines will know the service quality in time. Should have customers negative comments more than positive, the airline could, as early as possible, approach the corresponding issue appropriately. The airlines also can use this analysis model to set up the strategies to improve their customer relationship.

สารบัญ

สารบัญ (ต่อ)

\

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็น (Sentiment Analysis) ที่มีในสื่อสังคมออนไลน์ (Social Media) ของผู้ใช้สินค้าหรือบริการได้รับความนิยมเป็นอย่างมาก เนื่องเพราะรูปแบบการดำเนินชีวิตและการดำเนินธุรกิจในปัจจุบันถูกขับเคลื่อนด้วยข้อมูล ผู้ใช้เพียงสมัครและเข้าใช้งานผ่านแอพพลิเคชั่นต่าง ๆ ที่เปิดให้บริการในสื่อสังคมออนไลน์ (Social Media) ก็สามารถใช้บริการข้อมูลต่าง ๆ ได้ เช่น ตรวจสอบสภาพอากาศ จองตั๋วโดยสาร ซื้อสินค้า หรือศึกษาข้อมูลของสินค้าหรือบริการที่สนใจได้จากโพสต์และข้อความแสดงความคิดเห็นของผู้ที่เคยใช้สินค้าหรือบริการ (Review) นั้นมาก่อน ซึ่งข้อมูลเหล่านี้ส่วนช่วยให้ผู้ที่สนใจตัดสินใจง่ายขึ้น ทุก ๆ วินาทีจะมีข้อมูลแสดงความคิดเห็นเกิดขึ้นมาใหม่ โดยที่มีปริมาณข้อมูลเพิ่มขึ้นอย่างมหาศาล ในยุคก่อนหากบริษัทต้องการสำรวจความพึงพอใจเพื่อปรับปรุงการให้บริการ ธุรกิจจำเป็นต้องอาศัยแบบสอบถามในการรวบรวมข้อมูล ซึ่งมีค่าใช้จ่ายในการจัดทำสูงและยุ่งยากในการบริหารจัดการ ดังนั้นการเลือกนำข้อความแสดงความคิดเห็น (Sentiment Analysis) จากสื่อออนไลน์มาวิเคราะห์เพื่อหาความพึงพอใจของผู้ใช้บริการจึงได้รับความนิยมเพิ่มขึ้น เนื่องจากเป็นข้อมูลที่ลูกค้าแสดงความคิดเห็นด้วยตนเอง แต่เนื่องจากข้อมูลเหล่านี้มีจำนวนมาก และไม่มีโครงสร้างที่แน่นอน การจัดการข้อมูลด้วยคนจึงทำได้ยากและต้องใช้ระยะเวลานาน ดังนั้นหากสามารถสร้างกระบวนการวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็น (Sentiment Analysis) ได้อย่างมีประสิทธิภาพ หน่วยงานธุรกิจจะสามารถทราบได้ว่าลูกค้ามีความรู้สึกอย่างไรเมื่อใช้สินค้าหรือบริการ หากลูกค้ามีความคิดเห็นเชิงลบ (Negative) มากกว่าความคิดเห็นเชิงบวก (Positive) ธุรกิจก็สามารถข้อมูลที่ผ่านการวิเคราะห์มาปรับปรุงแก้ไขได้ทันเวลา และยังนำผลจากการวิเคราะห์มาใช้นำกำหนดกลยุทธ์สร้างความพึงพอใจเพื่อใช้ขับเคลื่อนธุรกิจได้โดยใช้ระยะเวลาน้อยลง

งานนิพนธ์นี้ได้นำข้อมูล Twitter จากเว็บไซต์ http://www.kaggle.com ซึ่งได้รวบรวมความคิดเห็นของผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา ในเดือนกุมภาพันธ์ ค.ศ.2015 มีข้อมูลที่รวบรวมเอาไว้ทั้งสิ้น 14,640 เรคคอร์ด และผู้จัดทำงานนิพนธ์ได้ศึกษาเครื่องมือ เทคนิค และการบวนการดำเนินการ เพื่อสร้างกระบวนการวิเคราะห์ความรู้สึก และนำแบบจำลองที่สร้างขึ้นไปใช้กับวิเคราะห์ข้อมูลที่ได้จัดเตรียมไว้ และ นำมาทดลองใช้กับกลุ่มของข้อความแสดงความคิดเห็นที่ถูกแบ่งไว้สำหรับทดสอบ (Test Data)

1.2 วัตถุประสงค์ของการศึกษา

1. เพื่อสร้างกระบวนการวิเคราะห์ข้อความแสดงความคิดเห็น
2. เพื่อศึกษาและเลือกเทคนิคสร้างแบบจำลองที่ให้ค่าความถูกต้องที่ดี
3. เพื่อวิเคราะห์ผลลัพธ์ที่ได้

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา

1. ได้กระบวนการวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็นที่ช่วยลดระยะเวลาการวิเคราะห์ความคิดเห็นที่เกิดขึ้นใหม่ ต่อบริการของสายการบินของประเทศสหรัฐ
2. ได้อัลกอริทึมที่เหมาะสมสำหรับวิเคราะห์ข้อความแสดงความคิดเห็น
3. ได้ข้อมูลที่ผ่านการวิเคราะห์ เพื่อใช้สำหรับปรับปรุงการให้บริการสายการบิน

1.3 ขอบเขตของการศึกษา

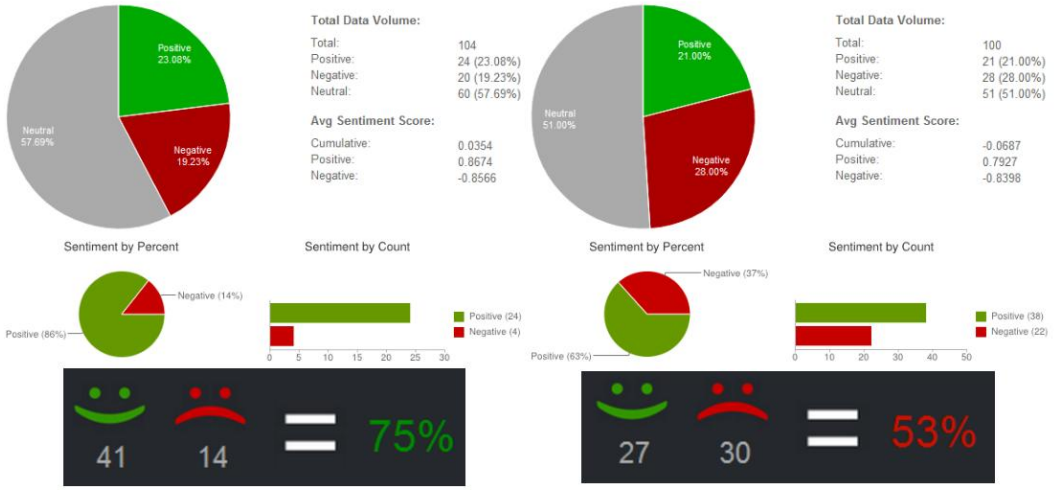
ศึกษาและสร้างแบบจำลองสำหรับใช้วิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็นผู้ใช้บริการสายการบินประเทศสหรัฐอเมริกา ด้วยข้อมูลจาก Kaggle.com ซึ่งเป็นข้อมูลเดือนกุมภาพันธ์ 2015 มีจำนวนข้อมูลทั้งสิ้น 14,640 ข้อความ เพื่อนำแบบจำลองที่สร้างขึ้นจำแนกข้อมูลแสดงความคิดเห็นที่เกิดขึ้นใหม่ทุก ๆ วันได้

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 Sentiment Analysis

Sentiment Analysis หรือบางครั้งอาจเรียกว่า Opinion mining, Sentiment mining คือ การประมวลภาษาธรรมชาติ โดยมีข้อมูลเป็นข้อความแสดงความคิดเห็นของผู้ใช้สินค้าหรือบริการที่ได้แสดงความคิดเห็นไว้ในที่ต่าง ๆ ทั้งในกระดานสนทนา (Web board) หรือในสื่อสังคมออนไลน์ เช่น Facebook, Twitter โดยนำข้อมูลมาวิเคราะห์และจำแนกประเภทของความรู้สึก เช่น ความรู้สึกเชิงบวก (Positive) ความรู้สึกเชิงลบ (Negative) หรือความรู้สึกเป็นกลาง (Neutral)

บริษัทสามารถนำผลที่ได้จากการวิเคราะห์ มาใช้เพื่อปรับปรุงสินค้าหรือบริการให้ดีขึ้น ดังภาพที่ 2-1 ที่รวบรวมข้อมูลและแสดงผลออกมาในรูปแบบกราฟที่สวยงาม

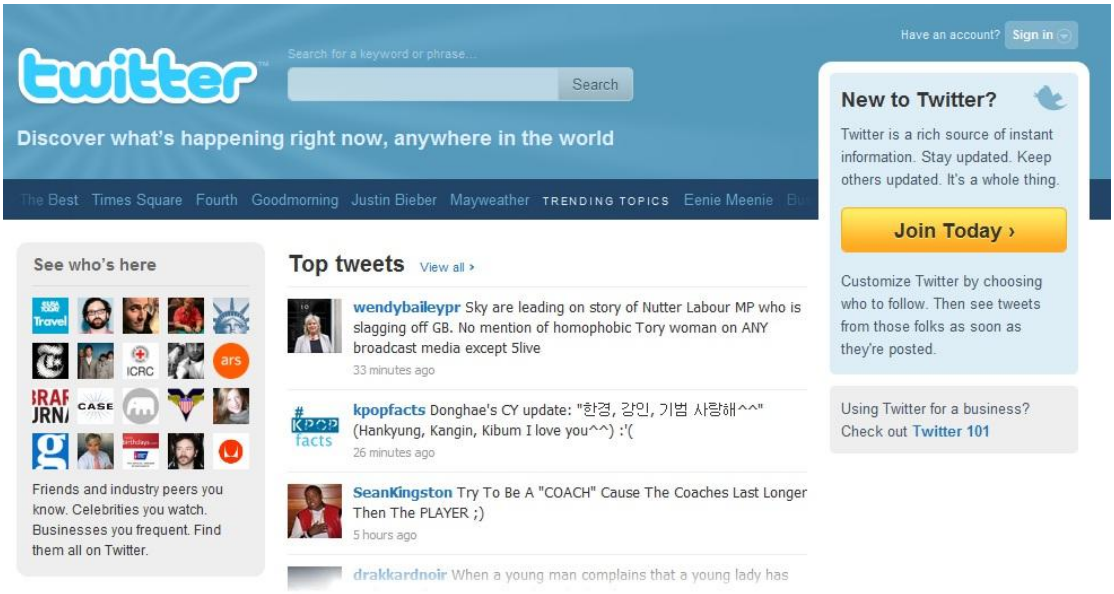


ภาพที่ 2-1 ทัศนคติของผู้ใช้สินค้าและบริการ

สืบค้นจาก http://meta-guide.com/bots-agents-assistants/chatbots

2.2 Twitter

Twitter เป็นสื่อสังคมออนไลน์ (Social Media) ที่ให้บริการสำหรับส่งข่าวสารและติดต่อสื่อสารประเภทข้อความสั้น (Short Message) ถูกพัฒนาขึ้นเมื่อเดือนมีนาคม ปี คศ. 2006 โดย Jack Dorsey, Noah Glass, Biz Stone, Evan William และเริ่มเปิดให้บริการเมื่อเดือนกรกฎาคม ปี คศ. 2012 สำนักงานใหญ่ตั้งอยู่ที่ เมืองซานฟรานซิสโก ประเทศสหรัฐอเมริกา โดยชื่อของทวิสเตอร์ตั้งมาจากเสียงร้องของนก

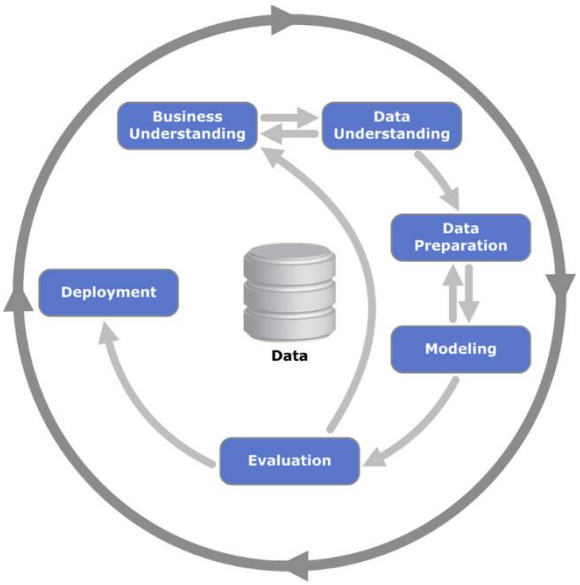


ภาพที่ 2-2 หน้าจอการใช้งาน Twitter สืบค้นจาก http://keywordsuggest.org

ผู้ใช้สามารถโพสต์และตอบกลับข้อความได้ แต่จำนวนตัวอักษรต้องไม่เกิน 140 อักษร ผู้ใช้สามารถเรียกใช้บริการผ่านหน้าเว็บไซต์ของทวิสเตอร์ หรือผ่าน Appication บนอุปกรณ์สื่อสารได้ และทวิสเตอรยังเปิดส่วนเชื่อมต่อ (Application programming interface: API) เพื่อให้บุคคลสามารถเชื่อมต่อและนำข้อมูลจาก Twitter มาใช้งานได้ (ปัจจุบัน Twitter ได้ความนิยมไปทั่วโลก ใน 1 วัน มีผู้ส่งข้อความถึง 340 ล้านข้อความ)

2.3 Data Mining

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนค้นหาความรู้ที่ถูกซ่อนอยู่ในข้อมูลจำนวนมาก ที่ถูกจัดเก็บไว้ในคลังข้อมูล ด้วยการใช้หลักการทางคณิตศาสตร์และสถิติ ซึ่งความรู้ที่ได้จากการทำเหมืองข้อมูลจะเป็นองค์ความรู้ใหม่ที่ไม่เคยเกิดขึ้นมาก่อน โดยขั้นตอนการทำเหมืองข้อมูลประกอบด้วย 6 ขั้นตอน คือ การทำความเข้าใจปัญหา, การทำความเข้าใจข้อมูล, การเตรียมข้อมูล, การสร้างแบบจำลอง, การประเมินผล และ การนำผลลัพธ์ที่ได้ไปใช้งาน การทำเหมืองข้อมูลมีเทคนิคการค้นหาความรู้ได้หลายรูปแบบ โดยที่สามารถเลือกใช้ได้ตามลักษณะของปัญหาที่ต้องการนำการทำเหมืองข้อมูลไปใช้แก้ไข เช่น การจำแนกประเภทของข้อมูล (Classification), การค้นหาความสัมพันธ์ของข้อมูล (Association rule) หรือการแบ่งกลุ่มข้อมูล (Clustering)



ภาพที่ 2-3 แสดงกระบวนการมาตรฐาน CRISP-DM ในการทำเหมืองข้อมูล

สืบค้นจาก https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_ for_ Data_Mining

การทำเหมืองข้อมูลนิยมนำไปใช้ในงานหลายด้าน เช่น การนำไปใช้จัดกลุ่มลูกค้าสินเชื่อของธนาคาร การแยกประเภทโรคของพืช หรือการหาความสัมพันธ์ของสินค้าที่มักถูกซื้อร่วมกัน งานนิพนธ์นี้ได้นำการจำแนกประเภทของข้อมูล (Classification) มาประยุกต์ใช้จำแนกประเภทการแสดงความคิดเห็นเพื่อให้ทราบว่าข้อความแสดงความคิดเห็นจากลูกค้า ว่ามีความคิดในเชิงบวก (Positive) หรือ ลบ (Negative) ต่อการใช้บริการ

โดยผู้จัดทำงานนิพนธ์ได้เลือกใช้ 3 อัลกอริทึมของการจำแนกประเภทของข้อมูล (Classification) มาใช้ในงานนิพนธ์ ได้แก่ Random forest, Neural Network และอัลกอริทึม Naive Bayes

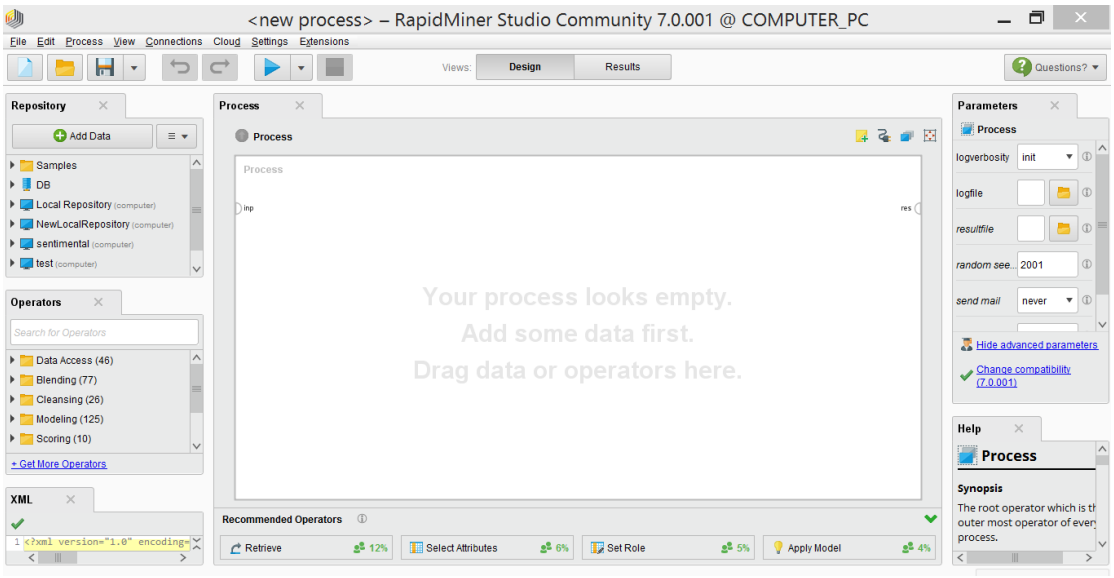
2.3.1 อัลกอริทึม Random forest เป็นอัลกอริทึมที่ถูกพัฒนาต่อมาจากต้นไม้ตัดสิน (Decision Tree) หลักการทำงานของ Random Forest จะทำการสุ่มแอตทริบิวต์และสุ่มข้อมูลตัวอย่างจากข้อมูลที่นำมาทำ Training data และแบ่งออกมาเป็นหลาย ๆ ชุด จากนั้นจะทำการสร้างเป็นต้นไม้ตัดสิน (Decision Tree) หลาย ๆ ต้น โดยจะเลือกเอาผลลัพธ์จากต้นไม้ตัดสินใจ ที่ให้ผลลัพธ์ที่ดีที่สุดมาเป็นคำตอบ (Majority Vote)

2.3.2 อัลกอริทึม Neural Network วิธีนี้เป็นการจำลองการทำงานของสมองมนุษย์ โดยใช้การคำนวณค่าน้ำหนัก (weight) ของเส้นที่เชื่อมระหว่างแต่ละโหนด โครงสร้างของ Neural Network ประกอบด้วย 3 ชั้น (layer) ชั้นที่ 1 Input layer , ชั้นที่ 2 Hidden layer และชั้นที่ 3 Output layer เมื่อนำเข้าข้อมูลในชั้น Input layer แล้วชั้น Hidden layer จะนำค่าของโหนดต่าง ๆ คูณกับค่าน้ำหนักในแต่ละเส้นเชื่อมเพื่อหาผลรวม หลังจากนั้นจะส่งออกไปยังชั้น Output layer โดยถ้าผลลัพธ์จากการคูณแตกต่างจากคำตอบ Neural Network จะปรับค่าน้ำหนักไปเรื่อย ๆ จนผลลัพธ์ใกล้เคียงคำตอบมากที่สุด

2.3.3 อัลกอริทึม Naive Bayes วิธีนี้จะใช้ความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น โดยนำสมการที่เรียกว่า Bayes Theorem หรือทฤษฎีของเบย์ มาใช้หาคำตอบ เช่น หากเกิดเหตุการณ์ที่หนึ่งขึ้นแล้ว ความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สองด้วยนั้นมีความน่าจะเป็นคิดเป็นกี่เปอร์เซ็นต์ และนำค่าความน่าจะเป็นมาคูณกัน โดยเลือกเอาผลลัพธ์ที่ได้ค่ามากที่สุดเป็นคำตอบ

2.4 RapidMiner

RapidMiner เป็นโปรแกรมสำหรับวิเคราะห์ข้อมูล ด้วยวิธี Data Mining, Text Mining, Predictive Analytics, Machine Learning, และ Business Analytics ที่ได้รับความนิยมโปรแกรมหนึ่ง ถูกพัฒนาขึ้นจากบริษัทที่ชื่อว่า Rapid-I ในประเทศเยอรมนีเมื่อปี คศ.2013 ต่อมาเปลี่ยนชื่อเป็น RapidMiner



ภาพที่ 2-4 ตัวอย่าง GUI ของโปรแกรม RapidMiner

โปรแกรม RapidMiner มีความสามารถรองรับการใช้งานไฟล์ได้หลายประเภท มีรูปแบบการแสดงผลเป็นกราฟที่สวยงาม และสามารถบันทึกผลออกมาเป็นไฟล์ภาพได้หลายนามสกุล ผลสำรวจจากเว็บไซต์ KDnuggets ในปี 2014 พบว่า RapidMiner มีผู้ตอบรับการใช้งานในการวิเคราะห์ข้อมูลมากเป็นอันดับ 1 และบริษัท Gartner ที่ทำวิจัยและให้คำแนะนำด้าน IT ในสหรัฐอเมริกาได้จัดให้โปรแกรม Rapidminer อยู่ในกลุ่ม Leadder สำหรับ Software ในการวิเคราะห์ข้อมูลปี คศ. 2015

บทที่ 3
วิธีการดำเนินการวิจัย

ในบทนี้นำเสนอขั้นตอนการดำเนินงานกระบวนการวิเคราะห์ความรู้สึก (Sentiment Analysis) ของผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา ซึ่งมีกระบวนการทั้งหมด มีรายละเอียดดังต่อไปนี้

1. การวิเคราะห์ข้อมูล
2. การเตรียมข้อมูล
3. การสร้างแบบจำลองและวัดประสิทธิภาพ

3.1 การวิเคราะห์ข้อมูล

งานนิพนธ์นี้ได้นำเสนอกระบวนการวิเคราะห์ความรู้สึก (Sentiment Analysis) ของผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา ในเดือนกุมภาพันธ์ ค.ศ.2015 โดยผู้ทำงานนิพนธ์ได้นำข้อมูล Twitter จากเว็บไซต์ http://www.kaggle.com ซึ่งได้รวบรวมความคิดเห็นของผู้โดยสาร ซึ่งมีข้อมูลที่รวบรวมเอาไว้ทั้งสิ้น 14,640 เรคคอร์ด ภายในข้อมูลถูกแบ่งเป็นประเภทข้อมูลเชิงลบ (Negative) มีจำนวน 9,178 เรคคอร์ด ข้อมูลเชิงบวก (Positive) มีจำนวน 2,363 เรคคอร์ด และข้อมูลที่มีลักษณะความคิดเห็นเป็นกลาง (Neutral) มีจำนวน 3,099 เรคคอร์ด

ลักษณะข้อมูลจะเป็นข้อมูลที่เป็นข้อความสั้น (short message) เนื่องจาก twitter จำกัดให้ผู้ใช้งานสามารถส่งข้อความยาวได้ไม่เกิน 140 ตัวอักษรต่อการส่งข้อความ 1 ครั้ง ภายในข้อความผู้ใช้ส่วนใหญ่จะระบุเที่ยวบิน สถานที่ที่ต้องการเดินทาง และความรู้สึกต่อการใช้บริการ ข้อความอาจมีสัญลักษณ์แสดงอารมณ์ (emoticon) เช่น การใช้ :) เพื่อใช้สำหรับแทนการยิ้ม สัญลักษณ์ Hashtag (#) เพื่อรวมคำที่เป็นคำสำคัญ (Keyword) ที่ต้องการเน้นความรู้สึกบางครั้งผู้โพสต์ข้อความอาจพิมพ์คำสองคำติดกันโดยไม่มีวรรค รวมถึงใช้คำศัพท์ย่อ (Abbreviation) เพื่อลดจำนวนตัวอักษรในการพิมพ์ ให้สามารถส่งข้อความที่ต้องการได้ครบถ้วนและไม่เกินจำนวน 140 ตัวอักษรตามที่ twitter ได้กำหนดไว้

ตารางที่ 3-1 ตัวอย่างข้อความภายในชุดข้อมูล

ID	Class	Tweet
14127	Negative	we need a miracle @slcairport please help flight 1228 get out of here by 640 #FingersCrossed #pleasegod #missmykids

จากข้อความแสดงให้เห็นได้ว่าผู้ใช้บริการสายการบินที่ มีความรู้สึกในเชิงลบ (Negative) ต่อการใช้บริการเที่ยวบิน 1228 โดยได้บรรยายความรู้สึก และนำสัญลักษณ์ Hashtag มาใช้เพื่อเน้นย้ำข้อความให้ดูเด่น รวมถึงเน้นย้ำความรู้สึกขณะโพสต์

3.2 การเตรียมข้อมูล

การเตรียมข้อมูลผู้จัดทำงานนิพนธ์ทำการ Download ข้อมูลจากเว็บไซต์ http://www.kaggle.com ในชุดข้อมูลมีจำนวน attribute มีจำนวน 15 attribute ประกอบด้วย Tweet_id, Airline_sentiment, Airline_sentiment_confidence, Negativereason_confidence, Text Negativereason, Airline, Airline_sentiment_gold, Name, Negativereason_gold, Tweet_location Retweet_count, Tweet_coord, Tweet_created และ User_timezone ผู้จัดทำงานนิพนธ์ได้เลือก Attribute name ที่มีชื่อว่า Airline_sentiment และได้ทำการเปลี่ยนชื่อเป็น Class เพื่อใช้สำหรับระบุประเภทความรู้สึกของข้อความว่าเป็นประเภท Negative, Positive หรือ Neutral

จากนั้นและได้เลือก Attribute ที่ชื่อว่า Text แล้วเปลี่ยนชื่อเป็น Tweet เพื่อใช้สำหรับแสดงข้อความที่ผู้ใช้ได้แสดงความคิดเห็น นอกจากนี้ผู้จัดทำงานนิพนธ์ได้สร้าง attribute โดยตั้งชื่อว่า ID เพื่อใช้บอกลำดับข้อมูลและตรวจสอบข้อมูลที่ใช้ในการประมวลผล

ตารางที่ 3-2 ข้อมูลที่ได้ Download จากเว็บไซต์ http://www.kaggle.com โดยมีจำนวนข้อมูลทั้งหมด 15 attribute

Tweet id	**airline sentiment**	**airline sentiment confidence**	**negative reason**	**negative reason confidence**	**airline**	**airline sentiment gold**	**name**	**negative reason gold**	**retweet count**	**text**	**tweet coord**	**tweet created**	**tweet location**	**user timezone**
5.7E+17	neutral	1			Virgin America		cairdin		0	What @dhepburn said.		24/2/2015 11:35		Eastern Time (US & Canada)
5.7E+17	positive	0.3486		0	Virgin America		jnardino		0	plus you've added commercials to the experience... tacky.		24/2/2015 11:15		Pacific Time (US & Canada)
5.7E+17	neutral	1			Virgin America		kyle_romanoff		0	what happened to Doom?!		23/2/2015 10:58:43		

ตารางที่ 3-3 ข้อมูล attribute ที่ถูกเลือกเพื่อใช้ในการประมวลผล

ID	Class	Tweet
1	neutral	What @dhepburn said.

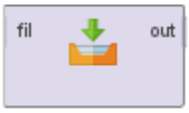
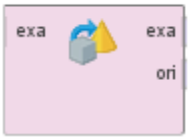
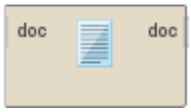
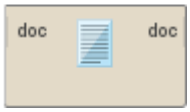
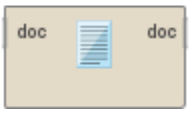
3.2.1 ขั้นตอนการเตรียมข้อมูลโดยใช้ RapidMiner

ผู้จัดทำงานนิพนธ์ได้จัดเตรียมข้อมูลทั้งหมด 2 ชุด โดยใช้โปรแกรม RapidMiner และ Excel ข้อมูลชุดแรกเป็นข้อมูลชุดที่ไม่ได้ผ่านการวิเคราะห์คำ วิเคราะห์สัญลักษณ์และแยกคำที่อยู่ติดกันภายในข้อความ ข้อมูลชุดที่สองเป็นข้อมูลที่ผ่านการวิเคราะห์คำ วิเคราะห์สัญลักษณ์และแยกคำแล้ว

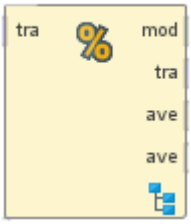
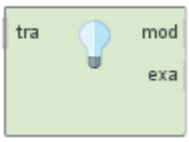
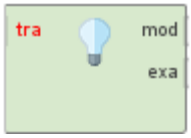
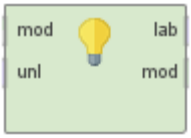
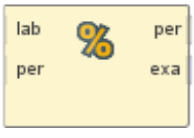
3.2.1.1 เตรียมข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์ที่เกิดขึ้นภายในข้อความ

ในขั้นตอนนี้ผู้จัดทำงานนิพนธ์แบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน โดยแบ่งข้อมูลส่วนแรกเป็น 70% (จากข้อมูลทั้งหมด) เก็บไว้ไฟล์ Trainning.xls เพื่อใช้ในการสร้างโมเดล และแบ่งข้อมูลส่วนที่สองเป็น 30% (จากข้อมูลทั้งหมด) เก็บไว้ไฟล์ Testing.xls เพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลองโดยข้อมูลทั้งสองส่วนไม่ได้ผ่านการ วิเคราะห์สัญลักษณ์และตัดคำเพิ่มเติม ตารางที่ 3-4 แสดงถึงรายละเอียดโอเปอเรเตอร์ที่เกี่ยวข้องกับการเตรียมข้อมูลทั้งสองชุด

ตารางที่ 3-4 โอเปอเรเตอร์ที่เกี่ยวข้องในการประมวลผลข้อมูล

ชื่อโอเปอเรเตอร์	คำอธิบาย
Read Excel 	ใช้อ่านไฟล์ข้อมูลประเภท Excel มาใช้งาน
Transform Cases 	ใช้สำหรับเปลี่ยนรูปแบบตัวอักษรภาษาอังกฤษตัวพิมพ์และและตัวพิมพ์ใหญ่ให้อยู่ในรูปแบบเดียวกัน
Nominal to Text 	ใช้สำหรับเปลี่ยนข้อมูลที่มีค่ามากกว่า 2 ค่าให้เป็นข้อมูลประเภทข้อความ
Tokenize 	ใช้สำหรับตัดประโยคข้อความให้แยกออกเป็นคำ
Filter Stopwords 	ใช้สำหรับตัดคำเชื่อมหรือคำที่ไม่จำเป็นทิ้ง
Generate n-Grams 	ใช้สำหรับการค้นหาเกิดขึ้นร่วมกัน

ตารางที่ 3-4 โอเปอเรเตอร์ที่เกี่ยวข้องในการประมวลผลข้อมูล (ต่อ)

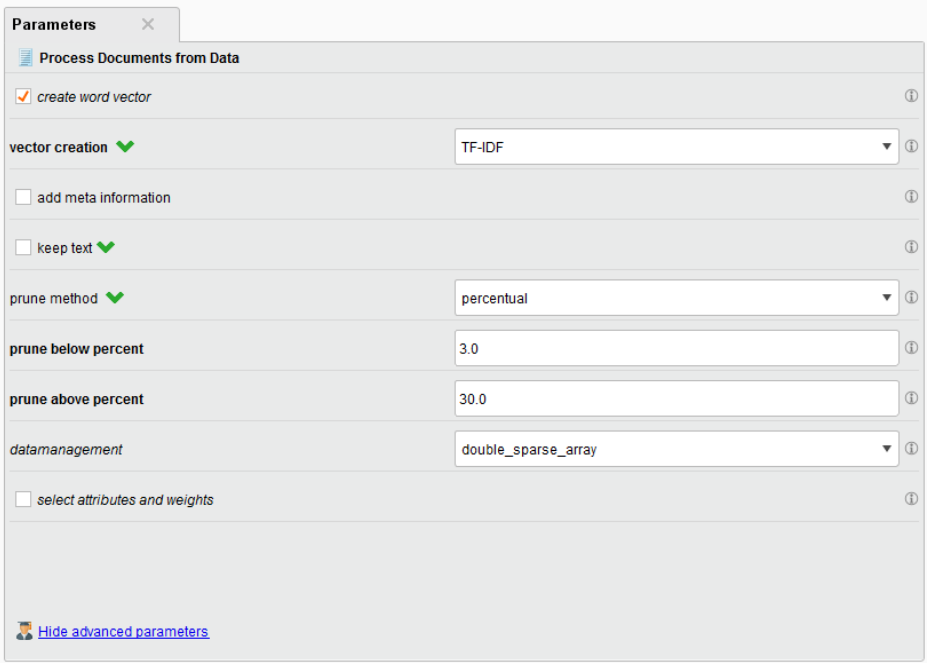
ชื่อโอเปอเรเตอร์	คำอธิบาย
Validation 	ใช้สำหรับแบ่งข้อมูลสำหรับสร้างแบบจำลองและทำการทดสอบ แบบ Cross – Validation
Naive Bayes 	ใช้สำหรับสร้างแบบจำลอง Naïve Bayes
Neural Net 	ใช้สำหรับสร้างแบบจำลอง Neural Net
Random Forest 	ใช้สำหรับสร้างแบจำลอง Random Forest
Apply Model 	ใช้สำหรับนำแบบจำลอง (Classification model) ไปใช้ทำนาย (Predict) ข้อมูลใหม่
Performance 	ใช้สำหรับวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้น

ขั้นตอนใช้งาน RapidMiner และการกำหนดค่าโอเปอเรเตอร์สำหรับข้อมูลชุดที่ 1 มีดังต่อไปนี้

1. อ่านข้อมูลจากไฟล์ที่ชื่อ Train imbalance.xlsx ด้วยโอเปอเรเตอร์ Read excel

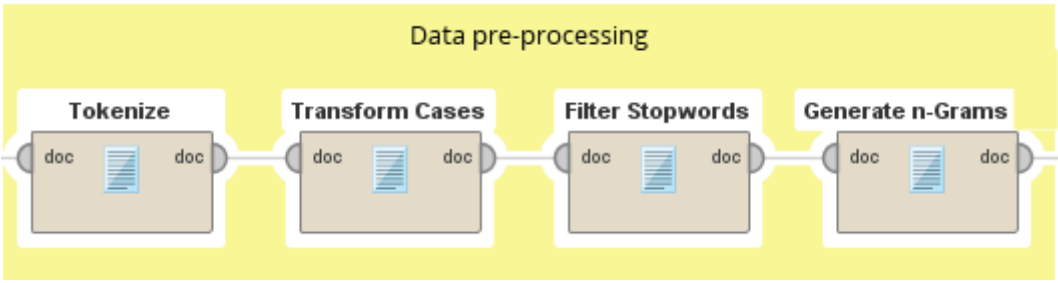
2. เพิ่มโอเปอเรเตอร์ Nominal to text และทำการเชื่อม port ด้านซ้ายกับ โอเปอเรเตอร์ Read excel เพื่อทำการแปลงข้อมูลทั้งหมดที่มีค่ามากกว่า 2 ค่าให้กลายเป็นข้อมูลประเภท Text

3. Click ที่โอเปอเรเตอร์ Process document from data เพื่อทำการจัดการข้อมูลประเภทข้อความที่ได้อ่านเข้ามาจากไฟล์ Excel กำหนดส่วนของ Vector creation ให้เป็นแบบ TF-IDF และเลือก Prune Method ให้เป็น Percentual เพื่อจำกัดคำที่มีค่าความถี่ของการเกิดขึ้นต่ำกว่าหรืออยู่ระหว่าง 3% - 30%



ภาพที่ 3-1 การกำหนดค่าโอเปอเรเตอร์ Process document from data

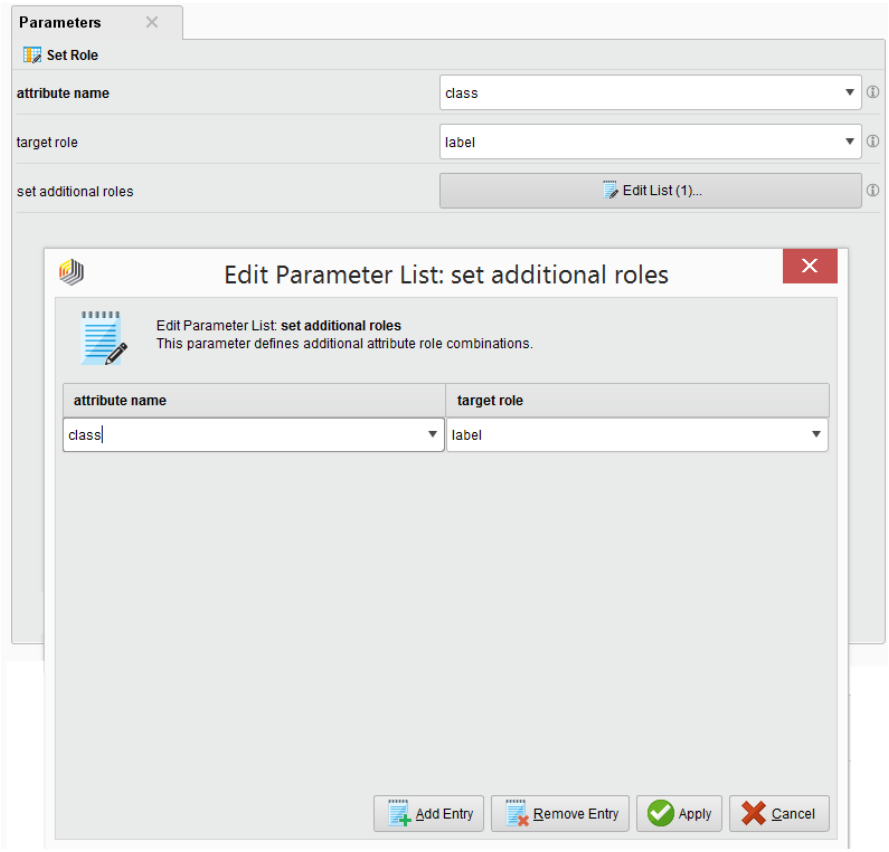
3.1 Double Click ที่โอเปอเรเตอร์ Process document from data อีกครั้งเพื่อเพิ่มโอเปอเรเตอร์ที่ใช้สำหรับจัดการข้อมูลประเภทข้อความ ประกอบไปด้วย โอเปอเรเตอร์ Tokenize , Transform case , Filter Stopwords และโอเปอเรเตอร์ Generate n-Grams ดังภาพที่ 3-2



ภาพที่ 3-2 กระบวนการจัดการข้อมูลประเภทข้อความ

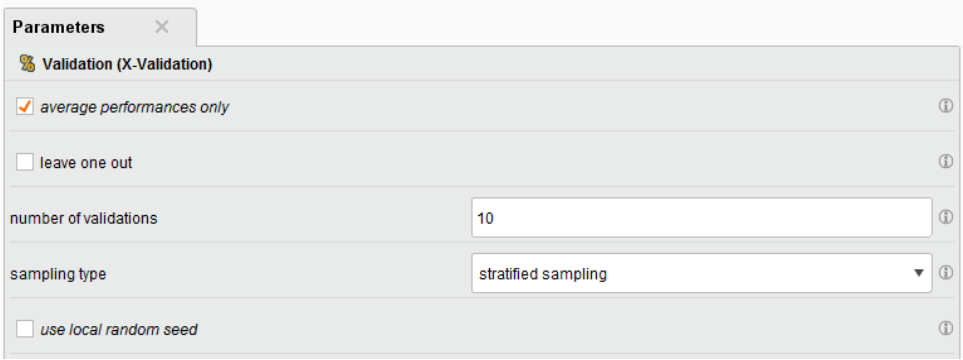
การกำหนดค่าโอเปอเรเตอร์ Tokenize ผู้จัดทำงานนิพนธ์ได้เลือกการตัดคำในประโยคโดยเลือกประเภท Regular Expression โดยกำหนดสัญลักษณ์พิเศษในการแทนค่า เป็น S/+ เพื่อให้โอเปอเรเตอร์ Tokenize ตัดคำที่ White space หรือตัดจากช่องว่างระหว่างคำเพื่อเก็บสัญลักษณ์ที่มีอยู่ทั้งหมดเอาไว้ และทำการเชื่อมต่อ Port ด้านซ้ายกับ port ที่ชื่อ Doc ในส่วน port ด้านขวาเชื่อมต่อกับ Port ของโอเปอเรเตอร์ Transform case กำหนดค่าโอเปอเรเตอร์ Transform case เลือก Lower case เพื่อเปลี่ยนตัวอักษรทั้งหมดให้เป็นตัวอักษรภาษาอังกฤษพิมพ์เล็ก จากนั้นทำการเชื่อมต่อโอเปอเรเตอร์ Filter Stopwords (English) เพื่อใช้ตัดคำเชื่อมหรือคำที่ไม่จำเป็นทิ้ง เช่น is, a, in ขั้นตอนสุดท้ายเชื่อมต่อโอเปอเรเตอร์ Generate n-grams (Terms) กำหนดค่า Max length ให้เท่ากับ 2 เพื่อค้นหาคำที่มักมีการเกิดร่วมกัน 2 คำ

4. เพิ่มโอเปอเรเตอร์ Set Role เพื่อกำหนดค่าที่ใช้เป็นคำตอบสำหรับ Training Data โดยเลือก Attribute name เป็น Attribute ที่มีชื่อว่า Class และ target role ให้เลือกเป็นประเภท label



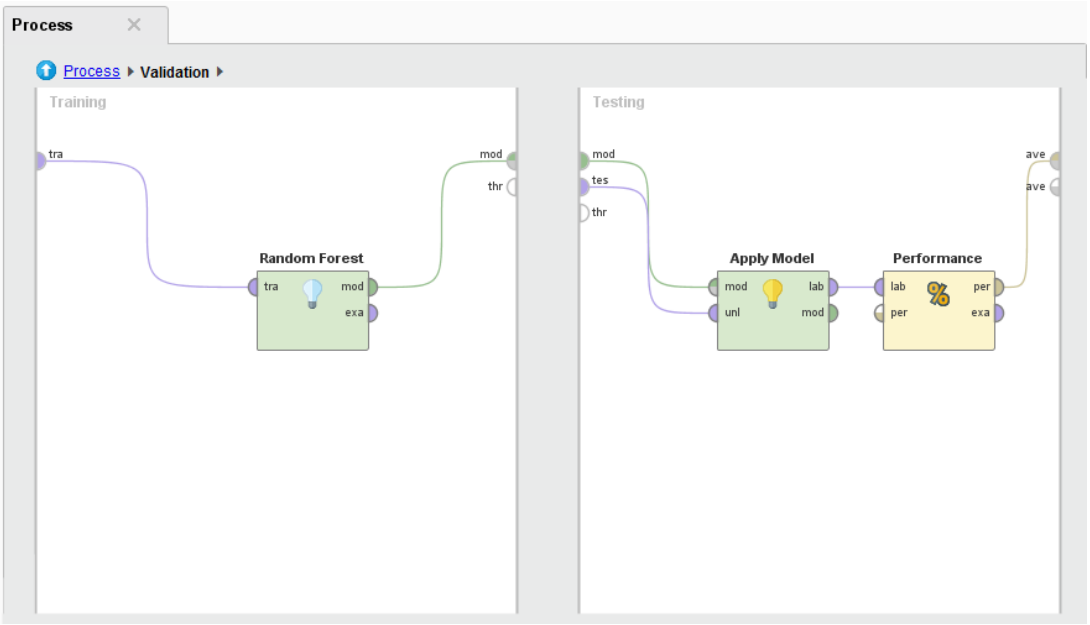
ภาพที่ 3-3 การกำหนดค่าโอเปอเรเตอร์ Set Role

5. เพิ่มโอเปอเรเตอร์ Validation เพื่อแบ่งข้อมูลสำหรับสร้างแบบจำลองและทำการทดสอบแบบจำลอง แบบ Cross – Validation โดยกำหนด number of validations เท่ากับ 10



ภาพที่ 3-4 การกำหนดค่าโอเปอเรเตอร์ Validation

5.1 Double Click ที่โอเปอเรเตอร์ Validation อีกครั้ง การทำงานภายในโอเปอเรเตอร์ Validation จะถูกแบ่งออกเป็น 2 ฝั่ง ในฝั่งซ้ายจะใช้สำหรับสร้างแบบจำลองโดยเลือกโอเปอเรเตอร์สำหรับสร้างเป็นแบบจำลองนำมาใส่ ผู้จัดทำงานนิพนธ์เลือกแบบจำลอง Random Forest มาใช้ฝั่งขวาเพิ่มโอเปอเรเตอร์ Apply Model และโอเปอเรเตอร์ Performance ใช้สำหรับวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้นมา



ภาพที่ 3-5 กระบวนการทำงานภายในโอเปอเรเตอร์ Validation

6. หลังจากได้แบบจำลองในขั้นตอนที่ 5 เพิ่มโอเปอเรเตอร์ Read excel ให้อ่านข้อมูลจากไฟล์ที่ชื่อ Test imbalance.xlsx โดยผู้จัดทำงานนิพนธ์จากนั้นเพิ่มโอเปอเรเตอร์ Nominal to text , Process document from data, Tokenize Transform case, Filter Stopwords (English), Generate n-grams (Terms) และเพิ่มโอเปอเรเตอร์ Apply Model เพื่อใช้นำโมเดล (Classification model) ไปใช้ทำนาย (Predict) ข้อมูลใหม่ ความแม่นยำของแบบจำลองดังกล่าวถูกแสดงดังภาพ 3-6

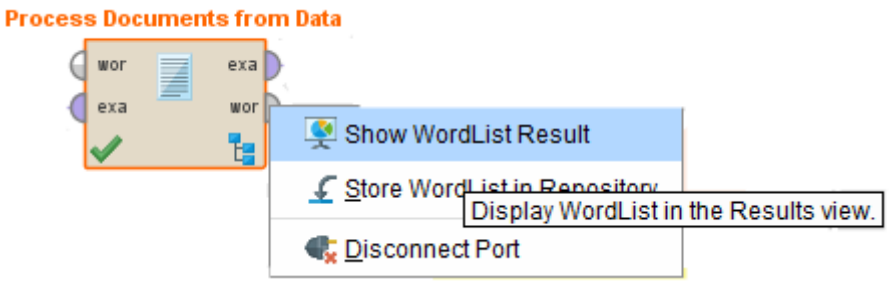
accuracy: 62.69% +/- 0.03% (mikro: 62.69%)

	true neutral	true positive	true negative	class precision
pred. neutral	0	0	0	0.00%
pred. positive	0	0	0	0.00%
pred. negative	3099	2363	9178	62.69%
class recall	0.00%	0.00%	100.00%	

ภาพที่ 3-6 ค่าความแม่นยำ (Accuracy) การประมวลผลด้วย Random Forest ข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์

3.1.2.2 การเตรียมข้อมูลแบบการวิเคราะห์คำ สัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ

จากการที่ผู้จัดทำงานนิพนธ์ได้เลือก Attribute ที่สำคัญสำหรับทำ Classification และใช้กระบวนการจัดการข้อมูลประเภทข้อความจากโปรแกรม Rapid miner ผู้จัดทำงานนิพนธ์สังเกตเห็นค่าความแม่นยำ (Accuracy) ที่ได้จากการประมวลผลยังไม่ดีเท่าที่ควร ผู้จัดทำงานนิพนธ์จึงได้วิเคราะห์คำ และวิเคราะห์สัญลักษณ์โดยใช้รายการคำ (Wordlist) ดังแสดงในภาพที่ 3-8 โดยรายการคำประกอบด้วยคำ (หรือสัญลักษณ์) ความถี่ในการเกิดขึ้นของคำ (หรือสัญลักษณ์) นั้น และจำนวนของคำ (หรือสัญลักษณ์) ที่เกิดขึ้นในแต่ละประเภทความคิดเห็น ทำให้ทราบได้ว่า คำหรือสัญลักษณ์ใดบ้างที่มีความสำคัญที่สามารถแยกประเภทความคิดเห็นเชิงบวก เชิงลบ หรือความคิดเห็นที่มีลักษณะเป็นกลางได้อย่างชัดเจน ผู้ใช้ RapidMiner สามารถเลือกดูผลลัพธ์จากรายการคำ (Wordlist) โดยกด Click ขวาที่ Port ชื่อว่า Wor ที่อยู่ด้านล่างฝั่งขวาของโอเปอเรเตอร์ Process document from data และเลือกคำสั่ง Show wordlist Result ดังภาพที่ 3-7



ภาพที่ 3-7 การเลือกดูผลลัพธ์จากรายการคำ (Wordlist)

WordList (Process Documents from Data)

Word	Attribute Name	Total Occurences	Document Occurences	neutral	positive	negative
!	!	48	44	6	19	23
!!	!!	22	22	1	5	16
!!!	!!!	11	11	3	5	3
!!!!	!!!!	3	3	0	2	1
!!!!!	!!!!!	3	3	2	0	1
!!!!!!	!!!!!!	1	1	0	0	1
!=	!=	1	1	0	1	0
!?	!?	2	2	1	1	0
!?!?	!?!?	1	1	0	0	1
!cancelled	!cancelled	1	1	0	0	1
!we	!we	1	1	0	0	1
"	"	6	5	0	2	4
"#flt	"#flt	1	1	0	0	1
".	".	1	1	0	0	1

ภาพที่ 3-8 หน้าจอผลลัพธ์ของรายการคำ (Wordlist)

จากภาพที่ 3-8 แสดงการนำข้อมูลจากรายการคำ (Wordlist) มาวิเคราะห์ความถี่ที่เกิดขึ้นของคำ (Total Occurrence , Document Occurrence) และจำนวนคำที่ถูกแบ่งอยู่ในแต่ละประเภทความคิดเห็น (Negative, Positive, Neutral) ทำให้ทราบได้ว่า คำหรือสัญลักษณ์เกิดขึ้นในประเภทความคิดเห็นใดบาง และเกิดขึ้นมากหรือน้อยเพียงใด มีประโยชน์เป็นอย่างมาก เพราะทำให้สามารถเห็นความสำคัญของแต่ละคำได้อย่างชัดเจน ซึ่งจะช่วยเพิ่มประสิทธิภาพให้แบบจำลอง มีความแม่นยำในการทำนายข้อมูลใหม่ได้มากขึ้น ผู้จัดทำงานนิพนธ์ได้นำข้อมูลที่ได้จากรายการคำ (Wordlist) มาทำการคำนวณอีกครั้งด้วยโปรแกรม Microsoft Excel เพื่อหาผลรวมของคำหรือสัญลักษณ์ โดยคิดเป็นค่าเฉลี่ยซึ่งทำการคำนวณจากข้อมูลทั้งหมด

ตารางที่ 3-5 แสดงการคำนวณคำ เพื่อหาค่าเฉลี่ยจากจำนวนข้อมูลทั้งหมด

คำ	จำนวนที่พบในเอกสาร	Neutral	Positive	Negative	Neutral คิดเป็นเปอร์เซ็นต์	Positive คิดเป็นเปอร์เซ็นต์	Negative คิดเป็นเปอร์เซ็นต์	ค่าเฉลี่ย Neutral	ค่าเฉลี่ย Positive	ค่าเฉลี่ย Negative	ผลรวม	ประเภท
(jblu)	10	10	0	0	0	0	0	1	0	0	1	neutral
$	347	31	15	301	0.1	0	0.2	0.2	0.2	0.6	1	negative
:(	89	12	4	73	0	0	0	0.3	0.2	0.5	1	negative
:)	155	29	106	20	0.1	0.3	0	0.2	0.8	0	1	neutral
great	268	436	445	1608	0.026	0.6	0.036	0.038	0.9	0.05	1	positive

ข้อมูลจากตารางที่ 3-5 พบว่ามีสัญลักษณ์ที่มีความสามารถในการแบ่งแยกประเภทของข้อมูลได้อย่างชัดเจน เช่น สัญลักษณ์รูปยิ้ม :) ซึ่งเกิดขึ้นในข้อมูลความคิดเห็นเชิงบวก (Positive) เป็นส่วนมาก แต่ในกระบวนการจัดการข้อมูลประเภทข้อความ RapidMiner อาจกำจัดสัญลักษณ์ที่มีความสำคัญออกไปก่อนจะถึงการนำข้อมูลเข้าไปประมวลผลในแบบจำลองที่ได้สร้างไว้

ผู้จัดทำงานนิพนธ์จึงได้ทำการแทนคำ (Replace) เพื่อป้องกันไม่ให้สัญลักษณ์ที่มีความสำคัญถูกกำจัดออกไป และได้เปลี่ยนสัญลักษณ์และตัวเลขบางกลุ่มให้กลายเป็นคำศัพท์ที่มีความหมายและง่ายต่อการประมวลผลข้อมูลของ RapidMiner ดังตาราง 3-6 ที่แสดงคำที่นำมาแทนสัญลักษณ์และกลุ่มของตัวเลขบางกลุ่ม

ตารางที่ 3-6 คำที่นำมาแทนสัญลักษณ์และกลุ่มของตัวเลขบางกลุ่ม

คำและสัญลักษณ์ที่ถูกแทน	**คำที่ใช้ในการแทนค่า**
;) , :)	smileyface
: (	angryface
Using	use
hrs. , hr	hour
Mins	minute
Luv	love
Nyc	NewYork
plz , pls	please
Sat	Saturday
Thx	Thank
Tmrw	Tomorrow
w/	With
w/o	With out
U	you
Ur	you are
$	money

ตารางที่ 3-6 คำที่นำมาแทนสัญลักษณ์และกลุ่มของตัวเลขบางกลุ่ม (ต่อ)

คำและสัญลักษณ์ที่ถูกแทน	**คำที่ใช้ในการแทนค่า**
weren't	were not
won't	will not
aren't	are not
can't , cant	can not
couldn't	could not
didn't	did not
don't	do not
doesn't	does not
aa. (American Airlines)	myairlines
lax (Los Angeles International Airport)	myairport
lga (LaGuardia Airport)	myairport
phx (Phoenix Sky Harbor International Airport)	myairport
ord (Chicago O'Hare International Airport)	myairport
กลุ่มตัวเลขที่แสดงถึงวัน เช่น Feb 21	mydate
กลุ่มข้อมูลตัวเลขที่แสดงถึงเที่ยวของสายการบิน เช่น AA1359 ,UA3417	myflight
กลุ่มตัวเลขที่อยู่หลังสัญลักษณ์ $ และ € เช่น $200 , €600	mymoney
กลุ่มตัวเลขที่แสดงถึงหมายเลขโทรศัพท์ เช่น (631)891-5722 , 310-795-2210	myphone
กลุ่มตัวเลขที่ที่แสดงถึงเวลา เช่น 11:30 pm , 3 hours. ,	mytime
กลุ่มตัวเลขที่แสดงถึงปี ค.ศ. เช่น oscars 2015 ,	myyear

ตารางที่ 3-7 ผลลัพธ์จากการแทนคำ (Replace)

ID	class	tweet
11174	negative	On phone hold for **mytime minute** trying to speak to an agent. **Can not** change reservation online. Sigh. #badcustomerexperience

เมื่อทำการเปลี่ยนแปลงข้อมูลส่วนหนึ่งด้วยการแทนคำ (Replace) แล้ว ผู้จัดทำงานนิพนธ์ยังสังเกตได้ว่าข้อมูล Twitter มีกลุ่มคำอีกส่วนหนึ่งที่อยู่หลังสัญลักษณ์ Hashtag ซึ่งเป็นกลุ่มคำที่สำคัญ (Keyword) ที่ผู้ใช้บริการสายการบินต้องการสื่อถึงความรู้สึกที่มีต่อการใช้บริการ ซึ่งกลุ่มคำเหล่านี้ก็มีความสามารถในการแบ่งแยกอารมณ์ได้ชัดเจน โดยมีจำนวนกลุ่มคำ 2,059 กลุ่มคำ นอกจากกลุ่มคำที่อยู่หลังสัญลักษณ์ Hashtag แล้วยังมีคำที่ผู้ใช้งานพิมพ์ข้อความติดกับสัญลักษณ์เพื่อลดจำนวนตัวอักษรในการสื่อสาร และการตัดคำด้วยโอเปอเรเตอร์ Tokenize ไม่สามารถแยกคำเหล่านี้ออกได้ ผู้จัดทำงานนิพนธ์จึงได้ทำการแยกสัญลักษณ์ Hashtag และสัญลักษณ์อื่น ๆ ที่ติดอยู่กับคำ รวมถึงทำการเว้นวรรคระหว่างคำ ให้เกิดเป็นคำศัพท์ที่มีความหมายและประมวลผลได้ง่าย ตารางที่ 3-8 แสดงผลจากการแยกสัญลักษณ์ในประโยคเดียวกับประโยคในตารางด้านบน

ตารางที่ 3-8 ผลลัพธ์ที่ได้จากการแทนคำ (Replace) และแยกสัญลักษณ์

ID	Class	Tweet
11174	negative	On phone hold for **mytime minute** trying to speak to an agent **. Can not** change reservation online **.** Sigh **. # bad customere xperience**

หลังจากมีการแทนคำและแยกสัญลักษณ์ และนำเข้าสู่การประมวลผลด้วยโปรแกรม RapidMiner อีกครั้งผลลัพธ์จากการประมวลผลได้ค่าความแม่นยำ (Accuracy) สูงขึ้นจากเดิมอย่างชัดเจน สาเหตุของการเพิ่มขึ้นของค่าความแม่นยำ (Accuracy) ส่วนหนึ่งมาจากแยกสัญลักษณ์ Hashtag และการแยกคำที่พิมพ์ติดกับสัญลักษณ์อื่น ๆ เช่น (#fail, #badservice, select??? (hopefully), *not*) คำส่วนใหญ่เป็นคำสำคัญ (Keyword) มีแม่นยำสูงเมื่อถูกแยกออกมาแล้ว หากคำๆ นั้นไปตรงกับคำที่ไม่ใช้สัญลักษณ์ก็จะทำให้ค่าผลรวมของคำ ๆ นั้นสูงขึ้น ทำให้ค่าความแม่นยำ (Accuracy)โดยรวมดีขึ้น

ตารางที่ 3-9 ตัวอย่างข้อมูลในตาราง Wordlist ของประโยคด้านล่าง
(ที่ได้ผ่านการตัดคำของ RapidMiner แล้ว)

Bad service #Badservice (service) 11:30 hour

ตารางที่ 3-10 ตัวอย่าง Wordlist ของประโยค Bad service #Badservice (service) 11:30 hour

Word	**Negative**	**Positive**	**Neutral**
Bad	1	0	0
service	1	0	0
#Badservice	1	0	0
(service)	1	0	0
11:30	1	0	0
Hour	1	0	0

ตารางที่ 3-11 ตัวอย่างข้อมูลในตาราง Wordlist ของประโยคด้านล่าง
(ที่ได้ผ่านการแทนคำและแยกคำแล้ว)

Bad service # Bad service (service) <u>mytime</u> hour

ตารางที่ 3-12 ตัวอย่าง Wordlist ของประโยค Bad service # Bad service (service)

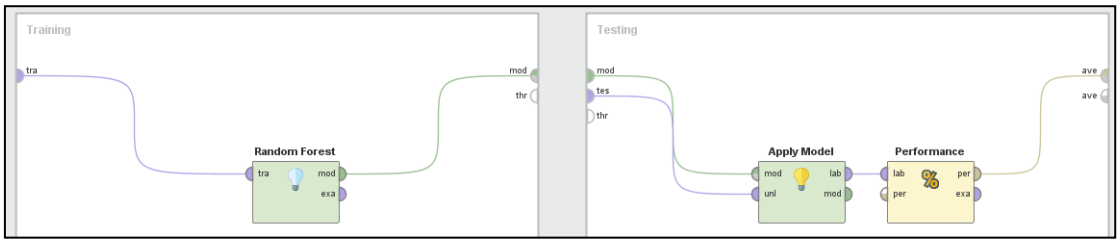
mytime hour

Word	Negative	Positive	Neutral
Bad	**2**	0	0
service	**3**	0	0
#	1	0	0
(	1	0	0
)	1	0	0
mytime	1	0	0
Hour	1	0	0

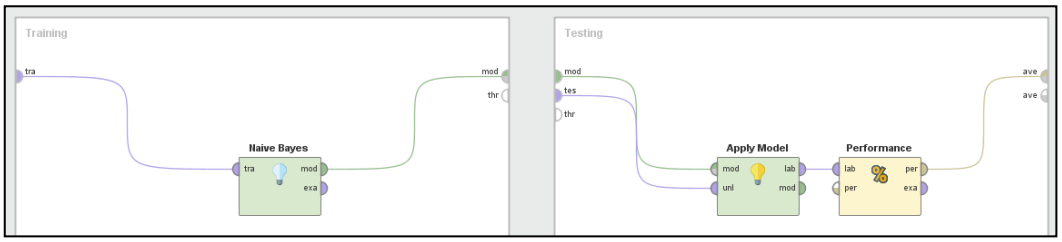
ข้อมูลจากตารางที่ 3-12 จะเห็นได้ว่าการแยกประโยคและการแทนคำ ทำให้ค่าความถี่ของคำว่า Bad และ Service ซึ่ง 2 คำนี้เป็นคำที่อยู่ในประเภทความคิดเห็นเชิงลบ (Negative) เพิ่มขึ้นอย่างชัดเจน ทำให้ค่าความแม่นยำ (Accuracy) ในการทำนายข้อมูลใหม่ถูกต้องมากขึ้น

3.3 การสร้างแบบจำลองและวัดประสิทธิภาพ

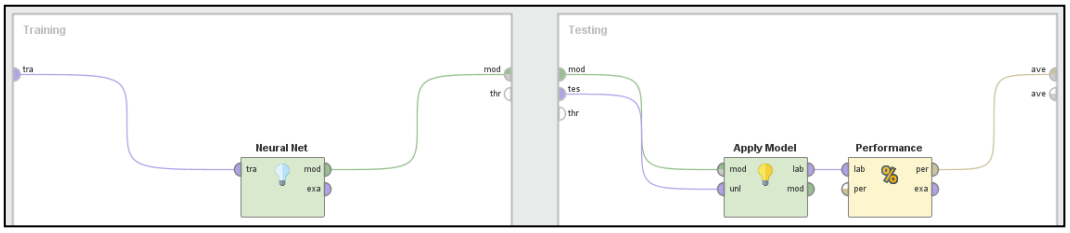
การสร้างแบบจำลอง สำหรับการสร้างแบบจำลองการเรียนรู้ ผู้จัดทำงานนิพนธ์ได้ทำการเลือกอัลกอริทึมการจำแนกเป็น 3 วิธีคือ Random forest ดังภาพที่ 3-9 อัลกอริทึม Naïve Bayes ดังภาพที่ 3-10 และอัลกอริทึม Neural Network ดังภาพที่ 3-11 (โดยอ้างอิงขั้นตอนวิธีการทำจากข้อที่ 5 หัวข้อ 3.2.1.1 เตรียมข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์ที่เกิดขึ้นภายในข้อความ)



ภาพที่ 3-9 การสร้างแบบจำลองโดยใช้อัลกอริทึม Random forest



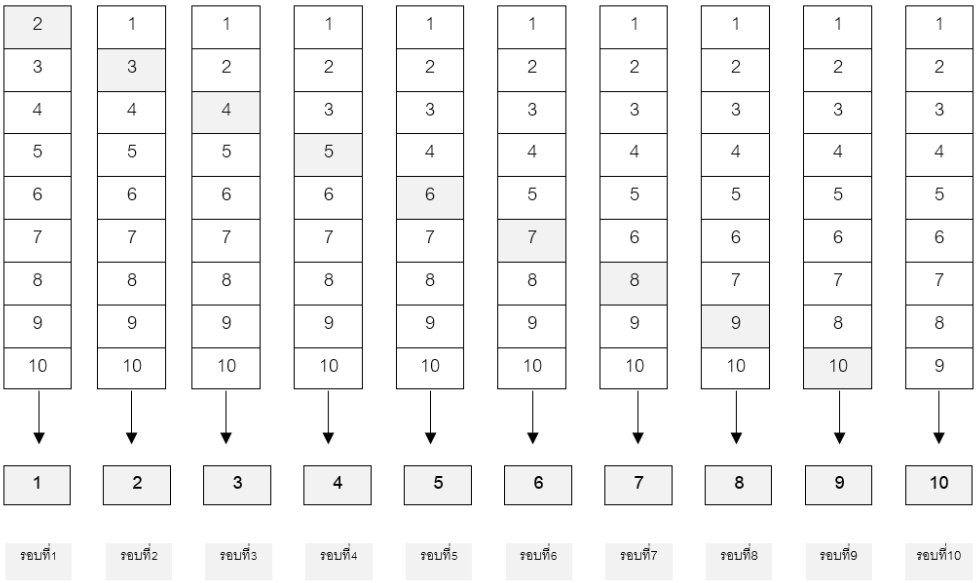
ภาพที่ 3-10 การสร้างแบบจำลองโดยใช้อัลกอริทึม Naïve Bayes



ภาพที่ 3-11 การสร้างแบบจำลองโดยใช้อัลกอริทึม Neural Network

3.3.1 การทดสอบและวัดประสิทธิภาพของแบบจำลอง

ในส่วนของการทดสอบข้อมูลผู้จัดทำงานนิพนธ์ได้เลือกวิธีการแบ่งข้อมูลทดสอบแบบ Cross - validation Test เพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลอง เนื่องจากเป็นวิธีที่มีความนิยมนำไปใช้ในการทำวิจัย และผลลัพธ์จากการทดสอบที่ได้มีความน่าเชื่อถือ การแบ่งข้อมูลทดสอบแบบ Cross-validation Test จะทำการแบ่งข้อมูลออกเป็นหลายส่วน โดยมักจะแทนด้วยค่า K โดยงานนิพนธ์นี้ใช้การกำหนดรอบของการทดสอบเท่ากับ 10 หรือ 10-fold cross-validation ซึ่งจะแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลอง การทดสอบจะทดสอบวนรอบจนครบจำนวนที่แบ่งไว้คือ 10 รอบการทดสอบ



ภาพที่ 3-12 วิธีการแบ่งข้อมูลทดสอบแบบ Cross - validation Test จำนวน 10 รอบ

จากภาพที่ 3-12 จะทำการแบ่งข้อมูลที่ใช้เป็นข้อมูลเรียนรู้ (Training data) โดยแบ่งข้อมูลออกเป็น 10 ส่วน มีจำนวนเท่ากันหลังจากนั้นทำการทดสอบประสิทธิภาพของแบบจำลองจำนวน 10 ครั้ง ดังนี้

การทดสอบรอบที่ 1 ใช้ข้อมูลส่วนที่ 2,3,4,5,6,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 1 เพื่อทำการทดสอบ

การทดสอบรอบที่ 2 ใช้ข้อมูลส่วนที่ 1,3,4,5,6,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 2 เพื่อทำการทดสอบ

การทดสอบรอบที่ 3 ใช้ข้อมูลส่วนที่ 1,2,4,5,6,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 3 เพื่อทำการทดสอบ

การทดสอบรอบที่ 4 ใช้ข้อมูลส่วนที่ 1,2,3,5,6,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 4 เพื่อทำการทดสอบ

การทดสอบรอบที่ 5 ใช้ข้อมูลส่วนที่ 1,2,3,4,6,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 5 เพื่อทำการทดสอบ

การทดสอบรอบที่ 6 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,7,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 6 เพื่อทำการทดสอบ

การทดสอบรอบที่ 7 ใช้ข้อมูลส่วนที่ 1, 2,3,4,5,6,8,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 7 เพื่อทำการทดสอบ

การทดสอบรอบที่ 8 ใช้ข้อมูลส่วนที่ 1, 2,3,4,5,6,7,9 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 8 เพื่อทำการทดสอบ

การทดสอบรอบที่ 9 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,7,8 และ 10 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 9 เพื่อทำการทดสอบ

การทดสอบรอบที่ 10 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,7,8,9 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 10 เพื่อทำการทดสอบ

หลังจากการแบ่งข้อมูลทดสอบแบบ Cross - validation Test เพื่อทดสอบแบบจำลองในการวิเคราะห์ความคิดเห็นของผู้ใช้บริการสายการบิน ผู้จัดทำงานนิพนธ์ได้เลือกวัดประสิทธิภาพด้วยการวัดค่าความแม่นยำ (Accuracy) เพื่อประเมินความสามารถของแบบจำลอง การวัดค่าความแม่นยำ (Accuracy) จะวัดจากการทำนายจำนวนข้อมูลที่แบบจำลองทำนายถูกต้อง จากทุกประเภทความคิดเห็นทั้งความคิดเห็นที่เป็น Negative Positive และ Neutral

ID	1	2	3	4	5	6	7	8	9
ประเภท	Positive	Negative	Positive	Positive	Negative	Positive	Neutral	Positive	Negative
ผลการทำนาย	Positive	Negative	Positive	Positive	Negative	Positive	Positive	Positive	Neutral
	✓	✓	✓	✓	✓	✓	×	✓	×

ภาพที่ 3-13 ผลการทำนายของแบบจำลอง

สูตรที่ใช้การคำนวณค่า Accuracy

$$\frac{\text{True Negative + True Positive + True Neutral}}{\text{True Negative + True Positive + True Neutral+ False Negative + False Positive + False Neutral}}$$

ภาพที่ 3-14 การคำนวณค่า Accuracy

จากรูปที่ 3-14 แสดงสูตรที่ใช้การคำนวณค่า Accuracy อ้างอิงจากหนังสือ (*AN INTRODUCTION TO DATA MINING TECHNIQUES (THAI VERSION)*. กรุงเทพมหานคร: เอเชีย ดิจิตอลการพิมพ์) โดยแบบจำลองทำนายประเภทข้อมูลถูกต้อง 7 ข้อมูล และทำนายประเภทของข้อมูลผิดจำนวน 2 ข้อมูล จากจำนวนข้อมูลทั้งหมด 9 ข้อมูล ซึ่งจะเท่ากับจำนวนข้อมูลที่ทำนายถูกหารด้วยจำนวนข้อมูลทั้งหมดแล้วนำมาคูณด้วย 100 เพื่อคิดเป็นเปอร์เซ็นต์หรือเท่ากับ 7 หาร 9 คูณด้วย 100 ได้ค่าความแม่นยำ (Accuracy) เท่ากับ 77.77 เปอร์เซ็นต์

บทที่ 4
ผลการศึกษา

4.1 ผลจากการเตรียมข้อมูล

ผลการทดสอบความถูกต้องของการวิเคราะห์ความรู้สึก (Sentiment Analysis) ของผู้โดยสารที่ใช้บริการสายการบินด้วยแบบจำลองที่สร้างขึ้น ในการทดลองผู้จัดทำงานนิพนธ์ได้จัดเตรียมข้อมูลเพื่อนำเข้ากระบวนการประมวลผลทั้งสิ้น 4 รูปแบบ ดังนี้

1. การเตรียมข้อมูลแบบไม่ได้มีการวิเคราะห์คำ และ สัญลักษณ์ที่เกิดขึ้นภายในข้อความ (ใช้ข้อความและสัญลักษณ์ที่ถูก Download มาเพื่อสร้างแบบจำลองทั้งหมด)

2. การเตรียมข้อมูลแบบตัดสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมายออกทั้งหมด (ตัดข้อมูลที่เป็นสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมายออกทั้งหมดก่อนสร้างแบบจำลอง)

3. การเตรียมข้อมูลแบบแยกคำที่ติดกันใน Hashtag และแทนสัญลักษณ์ด้วยคำเฉพาะ (ตัดคำที่ติดกันใน Hashtag ออกมาเป็นคำ ๆ และ แทนสัญลักษณ์ต่าง ๆ ด้วยคำฉพาะ เช่น แทน :=) ด้วยคำว่า SmileyFace)

4. การเตรียมข้อมูลแบบการวิเคราะห์คำ สัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ (คล้ายแบบที่สาม แต่แยกสัญลักษณ์ที่ติดกับตัวอักษรออก)

ผู้จัดทำงานนิพนธ์ได้เลือกใช้ 3 อัลกอริทึม ได้แก่ Naive Bayes และ Random forest Neural Network เพื่อวัดประสิทธิภาพค่าความแม่นยำ (Accuracy) สำหรับการวิเคราะห์ความรู้สึก

ผลจากการทดลอง ชี้ให้เห็นว่า การเตรียมข้อมูลด้วยในรูปแบบที่ 4 การวิเคราะห์คำ สัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ และ การสร้างแบบจำลองโดยใช้อัลกอริทึม Neural Network ให้ผลลัพธ์และค่าความแม่นยำ (Accuracy) มากที่สุด โดยให้ค่าความแม่นยำ 98.97 % มากกว่า Naive Bayes ที่ให้ค่าความแม่นยำ (Accuracy) 82.34 % และ Random forest ที่ให้ค่าความแม่นยำ (Accuracy) 82.69 %

4.1.1 ผลการเตรียมข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์ที่เกิดขึ้นภายในข้อความ

ตารางที่ 4-1 ข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์ที่เกิดขึ้นภายในข้อความ

ID	Class	Tweet
14473	Negative	you have to run the engine to troubleshoot an issue before boarding the plane!?! How about another plane? #aa2227 #miatoiah

ตารางที่ 4-2 ค่าความแม่นยำ (Accuracy) ของ 3 อัลกอริทึม ข้อมูลแบบไม่ได้มีการวิเคราะห์คำและสัญลักษณ์ที่เกิดขึ้นภายในข้อความ

Class	Naive Bayes		Random forest		Neural Network	
	Recall	Precision	Recall	precision	Recall	Precision
Positive	63.77 %	34.10 %	0.00 %	0.00 %	41.90 %	63.06 %
Negative	47.10 %	87.33 %	100 %	62.69 %	91.00 %	73.00 %
Neutral	50.60 %	29.75 %	0.00 %	0.00 %	24.94 %	47.45 %
Overall Accuracy	50.53 %		62.69 %		69.09 %	

ตารางที่ 4-1 แสดงข้อมูลที่ได้จากการ Download มาใช้ในกระบวนการโดยยังไม่ได้มีการเปลี่ยนแปลงรูปแบบของข้อมูล และ ตารางที่ 4-2 แสดงค่าความแม่นยำ (Accuracy) ของ 3 อัลกอริทึมโดย Neural Network ให้ค่าความแม่นยำ (Accuracy) มากที่สุด โดยให้ค่าความแม่นยำ 69.09 % เนื่องจากผลลัพธ์จากแบบจำลองยังให้ค่าความแม่นยำ (Accuracy) การทำนายข้อมูลใหม่ไม่ดี ผู้นิพนธ์จึงได้เลือกปรับปรุงที่ข้อมูลก่อนนำเข้าประมวลผล โดยการปรับปรุงข้อมูลในแต่ละรูปแบบจะถูกนำมาแสดงในตารางที่ 4-3 ตารางที่ 4-5 และตารางที่ 4-7 (ตารางที่ 4-3, 4-5 และ 4-7 แสดงข้อความเดียวกับข้อความในตารางที่ 4-1 หลังจากที่ได้รับการปรับปรุงแล้ว)

4.1.2 ผลการเตรียมข้อมูลแบบตัดสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมาย

ตารางที่ 4-3 ข้อมูลแบบตัดสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมายออกทั้งหมด

ID	Class	Tweet
14473	negative	run engine troubleshoot issue before boarding plane about another plane

ตารางที่ 4-4 ค่าความแม่นยำ (Accuracy) ของ 3 อัลกอริทึม ข้อมูลแบบตัดสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมายออกทั้งหมด

Class	Naive Bayes		Random forest		Neural Network	
	Recall	Precision	Recall	precision	Recall	precision
Positive	37.51 %	48.87 %	29.96 %	33.29 %	40.64 %	55.23 %
Negative	69.76 %	41.02 %	39.99 %	33.32 %	56.41 %	44.59 %
Neutral	23.74 %	44.66 %	29.95 %	33.28 %	39.84 %	38.89 %
Overall Accuracy	43.67 %		33.30 %		45.63%	

จากตารางที่ 4-3 ผู้จัดทำงานนิพนธ์ได้เตรียมข้อมูลแบบตัดสัญลักษณ์ ตัวเลข คำย่อ และคำที่ไม่มีความหมายตามพจนานุกรมออกทั้งหมด โดยหวังว่า จะทำให้แบบจำลองเกิดความเข้าใจข้อมูลและประมวลผลง่ายขึ้น สัญลักษณ์ที่ผู้จัดทำงานนิพนธ์ตัดออกจากข้อความ เช่น (# !! . , / ( ) - & $ € ™ € @) ,ตัวเลข 1234567890 , RT, NYC, LUV

หลังการนำข้อมูลเข้าไปในกระบวนการประมวล ค่าความแม่นยำ (Accuracy) ลดลงอย่างเห็นได้ชัด จากการวิเคราะห์ข้อมูลด้วยรายการคำ (WordList) จากโปรแกรม RapidMiner ดังแสดงในรูปที่ 4-1 พบว่าสาเหตุส่วนหนึ่งการลดลงของค่าความแม่นยำ (Accuracy) มาจากค่าความถี่ของคำในแต่ละประเภทความคิดเห็นทั้งประเภท Possitive Negative และ Neutral มีค่าความถี่การเกิดใกล้เคียงกัน (โดยข้อมูลไม่มีสัญลักษณ์หรือคำย่อบางคำที่มีความแม่นยำสูงในการแบ่งแยกประเภทเข้ามาช่วยจำแนก) ทำให้แบบจำลองทำนายได้ยากมากขึ้น ค่าของการทำนายข้อมูลรูปแบบที่ 2 จึงลดลง

WordList (Process Documents from Data)

Word	Attribute Name	Total Occurences	Document Occurences	negative	neutral	positive
booking	booking	320	315	119	99	102
seat	seat	389	380	159	116	114
make	make	312	307	94	97	121
gate	gate	435	417	124	184	127
check	check	361	355	145	87	129
fly	fly	474	467	162	173	139
please	please	422	411	102	167	153
change	change	443	430	133	156	154
plane	plane	638	608	226	241	171
airline	airline	574	513	240	156	178
time	time	801	766	359	257	185
delay	delay	668	617	193	281	194
phone	phone	452	436	92	165	195
baggage	baggage	740	712	304	240	196

ภาพที่ 4-1 ผลจากรายการคำ (Wordlist) ที่นำมาช่วยวิเคราะห์ข้อมูล

4.1.3 ผลการเตรียมข้อมูลแบบแยก Hashtag และแทนคำ

ตารางที่ 4-5 ข้อมูลแบบแยก Hashtag และแทนคำ

ID	Class	Tweet
14473	negative	you have to run the engine to troubleshoot an issue before boarding the plane!?! How about another plane? **# myflight # miatoiah**

ตารางที่ 4-6 ค่าความแม่นยำ (Accuracy) ของ 3 อัลกอริทึม ข้อมูลแบบแยก Hashtag และแทนคำ

Class	Naive Bayes		Random forest		Neural Network	
	Recall	**Precision**	**Recall**	**precision**	**Recall**	**precision**
Positive	46.68 %	79.81 %	37.16 %	80.77 %	55.01 %	74.67 %
Negative	48.84 %	60.93 %	45.07 %	40.45 %	45.96 %	61.74 %
Neutral	75.33 %	46.68 %	54.85 %	38.47 %	72.41 %	47.67 %
Overall Accuracy	56.95 %		45.69 %		57.79 %	

ข้อมูลตารางที่ 4-5 เป็นการเตรียมข้อมูลแบบแยก Hashtag และแทนคำเป็นการปรับปรุงการเตรียมข้อมูลจากหัวข้อที่ 4.2 จากการทดลองพบว่าการตัดสัญลักษณ์ ตัวเลข และคำที่ไม่มีความหมายออกไปจากข้อมูล มีผลทำให้ค่าความแม่นยำ (Accuracy) ลดลง ผู้จัดทำงานนิพนธ์จึงได้ใช้รายการคำ (Wordlist) วิเคราะห์คำ และวิเคราะห์สัญลักษณ์ ดังรูปที่ 3-8 ในบทที่ 3 และแยกคำที่อยู่ติดกันหลังสัญลักษณ์ Hashtag ออก จากนั้นแทนคำที่มีความหมายเพื่อใช้เปลี่ยนแทนสัญลักษณ์และคำย่อบางคำที่ต้องการเก็บไว้ ดังตารางที่ 4-5 เพื่อป้องกันไม่ให้สัญลักษณ์ที่มีความสำคัญถูกกำจัดออกไป

แต่เนื่องจากจำนวนของคำที่อยู่หลังสัญลักษณ์ Hashtag ที่ถูกแยกออกและคำที่ถูกแทนอาจมีจำนวนไม่มากเพียงพอที่จะช่วยในแบบจำลองทำนายข้อมูลได้ดีขึ้น ค่าความแม่นยำ (Accuracy) จึงยังไม่เพิ่มขึ้น

4.1.4 ผลการเตรียมข้อมูลแบบการวิเคราะห์คำ สัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ

ตารางที่ 4-7 ข้อมูลแบบการวิเคราะห์คำ สัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ

ID	Class	Tweet
14473	negative	you have to run the engine to troubleshoot an issue before boarding the plane ! ? ! How about another plane ? # myflight # mia to iah

ตารางที่ 4-8 ค่าความแม่นยำ (Accuracy) ของ 3 อัลกอริทึม ข้อมูลแบบการวิเคราะห์คำ สัญลักษณ์และแยกคำที่เกิดขึ้นภายในข้อความ

Class	Naive Bayes		Random forest		Neural Network	
	Recall	Precision	Recall	precision	Recall	precision
Positive	87.13 %	90.03 %	91.28 %	96.99 %	99.49 %	99.28 %
Negative	70.59 %	76.41 %	76.05 %	97.77 %	97.97 %	98.93 %
Neutral	89.29 %	80.57 %	97.16 %	75.85 %	99.45 %	98.70 %
Overall Accuracy	82.34 %		88.17 %		98.97 %	

ตารางที่ 4-8 แสดงให้เห็นว่า หลังจากมีการแทนคำและแยกสัญลักษณ์ และนำเข้าประมวลผลด้วยโปรแกรม RapidMiner อีกครั้ง ผลลัพธ์จากการประมวลผลได้ค่าความแม่นยำ (Accuracy) สูงขึ้นจากเดิมอย่างชัดเจน

สาเหตุของการเพิ่มขึ้นของค่าความแม่นยำ (Accuracy) ส่วนหนึ่งน่าจะมาจากแยกสัญลักษณ์ Hashtag และการแยกคำที่พิมพ์ติดกับสัญลักษณ์อื่น ๆ เช่น (#fail, #badservice, select???, (hopefully), *not*) เนื่องจาก คำส่วนใหญ่เป็นคำที่มีความสำคัญ (Keyword) และมีแม่นยำสูงเมื่อถูกแยกออกมาแล้ว หากคำ ๆ นั้นเมื่อถูกแยกออกมาแล้วและไปตรงกับคำที่ไม่ใช้สัญลักษณ์ก็จะทำให้จำนวนรวมของคำ ๆ นั้นสูงขึ้น (นั่นคือ มีอิทธิพลต่อการจำแนกอารมณ์มากขึ้น) ช่วยให้ค่าความแม่นยำ (Accuracy) โดยรวมดีขึ้น

สาเหตุที่แบบจำลองการเรียนรู้จากอัลกอริทึม Neural Network ให้ค่าความแม่นยำในการทำนายข้อมูลถูกต้องมากที่สุด เนื่องจาก Neural Network มีโครงสร้างภายในที่อยู่ในชั้น Hidden layer ที่ซับซ้อน และมีการส่งค่าย้อนกลับหากคำนวน เพื่อแก้ไขค่าที่มีความผิดพลาดที่แตกต่างจากผลคำตอบ (Label) และจะถูกปรับค่าจนแล้วว่าค่าผลลัพธ์ไม่ตรงกับคำตอบ โดยที่จะส่งค่ากลับไปปรับใหม่ตนได้ค่าที่ใกล้เคียงกับคำตอบมากที่สุดจึงจะส่งออกมายังชั้น Output Layer อัลกอริทึม Neural Network จึงทำงานที่มีความซับซ้อนของข้อมูลมาก ๆ ได้ดี

บทที่ 5
สรุปและอภิปรายผล

5.1 ผลการดำเนินงาน

งานนิพนธ์นี้ได้นำเสนอกระบวนการวิเคราะห์ความรู้สึก (Sentiment Analysis) โดยใช้ข้อมูลของผู้โดยสารที่ใช้บริการสายการบินของบริษัทในประเทศสหรัฐอเมริกา ในเดือนกุมภาพันธ์ ค.ศ.2015 โดยผู้ทำงานนิพนธ์ได้นำข้อมูล Twitter จากเว็บไซต์ http://www.kaggle.com ซึ่งได้รวบรวมความคิดเห็นของผู้โดยสาร โดยมีข้อมูลทั้งสิ้น 14,640 เรคคอร์ด เป็นข้อมูลเชิงลบ (Negative) จำนวน 9,178 เรคคอร์ด ข้อมูลเชิงบวก (Positive) จำนวน 2,363 เรคคอร์ด และข้อมูลความคิดเห็นที่เป็นกลาง (Neutral) จำนวน 3,099 เรคคอร์ด

ผู้จัดทำงานนิพนธ์ได้ดำเนินการทั้งหมด 3 ขั้นตอน เริ่มจาก การวิเคราะห์ข้อมูล ตามด้วย การเตรียมข้อมูล และ สิ้นสุดที่ การสร้างแบบจำลองและวัดประสิทธิภาพ

จากการทดลอง ผู้จัดทำงานนิพนธ์พบว่า หลังจากเตรียมข้อมูลด้วยการวิเคราะห์คำสัญลักษณ์และการแยกคำที่เกิดขึ้นภายในข้อความ แบบจำลองที่ถูกสร้างขึ้นโดยใช้อัลกอริทึม Neural Network ให้ผลลัพธ์และค่าความแม่นยำ (Accuracy) มากที่สุด โดยให้ค่าความแม่นยำ 98.97 % มากกว่า Naive Bayes ที่ให้ค่าความแม่นยำ (Accuracy) 82.34 % และ Random forest ที่ให้ค่าความแม่นยำ (Accuracy) 82.69%

5.2 ปัญหาและอุปสรรคของงานวิจัย

ลักษณะข้อมูลที่นำมาใช้ในงานนิพนธ์เป็นข้อมูลที่เป็นข้อความสั้น (Short messages) เนื่องจาก Twitter จำกัดให้ผู้ใช้งานสามารถส่งข้อความยาวได้ไม่เกิน 140 ตัวอักษรต่อการส่งข้อความ 1 ครั้ง ผู้ใช้จึงต้องพิมพ์ข้อความที่ต้องการส่งรวมกับสัญลักษณ์แสดงอารมณ์ (Emoticon) และสัญลักษณ์อื่นร่วมด้วยเพื่อลดจำนวนตัวอักษร ดังนั้นสัญลักษณ์และคำที่ติดกันโดยไม่มีวรรคในข้อความจึงมีเป็นจำนวนมาก การเตรียมข้อมูลโดยเปลี่ยนสัญลักษณ์ให้เป็นคำพูดเฉพาะ และ ตัดคำที่ติดกันจึงต้องใช้เวลามาก แต่จำเป็น เนื่องเพราะ หากเราสร้างแบบจำลองโดยใช้เพียงตัวอักษรในข้อความ (ตัดสัญลักษณ์ออก) และไม่ตัดคำที่ติดกันออกเป็นคำ ๆ แบบจำลองที่ได้จะมีความแม่นยำไม่สูงเท่าที่ควร

บรรณานุกรม

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิงเบื้องต้น*. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์.

Kaggle. *Twitter US Airline Sentiment* . เข้าถึงได้จาก https://www.kaggle.com/crowdflower/twitter-airline-sentiment

RapidMiner GmbH. *RapidMiner 7 Operator Reference Manual* . (2016). Available from: http://www.rapidminer.com.

Xiaotong Duan, Tianshu Ji, Wanyi Qian. (2016). *Twitter US Airline Recommendation Prediction. In CS 229 Machine Learning Final Projects, Spring 2016*. Departments of Chemical Engineering, Stanford University.