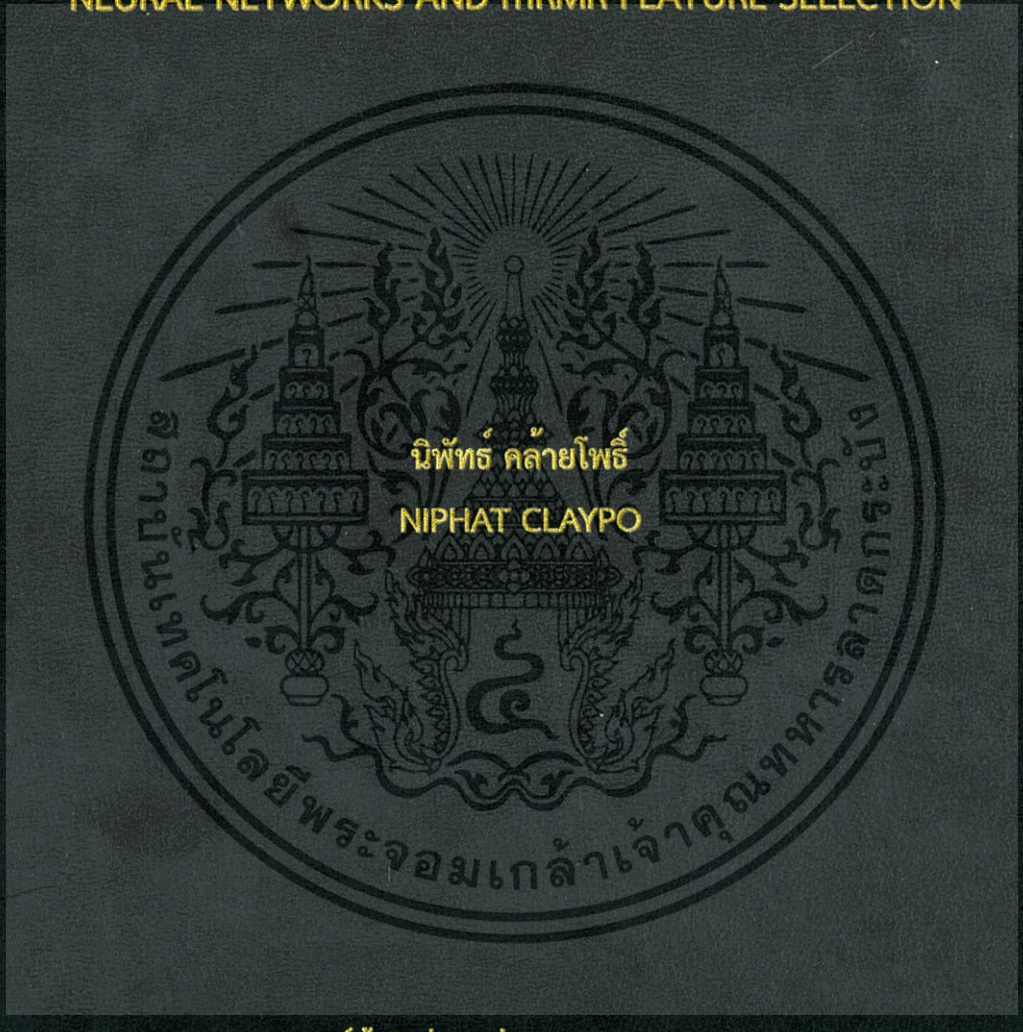
การทำเหมืองความคิดเห็นสำหรับร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียม และการคัดเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์

OPINION MINING FOR THAI RESTAURANT REVIEWS USING NEURAL NETWORKS AND mRMR FEATURE SELECTION



นิพัทธ์ คล้ายโพธิ์
NIPHAT CLAYPO

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2558
KMITL-2015-SC-M-002-048

การทำเหมืองความคิดเห็นสำหรับร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียม และการคัดเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์

OPINION MINING FOR THAI RESTAURANT REVIEWS USING NEURAL NETWORKS AND mRMR FEATURE SELECTION

นิพัทธ์ คล้ายโพธิ์

NIPHAT CLAYPO

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-SC-M-002-048

OPINION MINING FOR THAI RESTAURANT REVIEWS USING NEURAL NETWORKS AND mRMR FEATURE SELECTION

NIPHAT CLAYPO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2014

KMITL-2015-SC-M-002-048

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ "การทำเหมืองความคิดเห็นสำหรับร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียมและการคัดเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์"
"OPINION MINING FOR THAI RESTAURANT REVIEWS USING NEURAL NETWORKS AND mRMR FEATURE SELECTION"

ชื่อนักศึกษา นายนิพัทธ์ คล้ายโพธิ์
รหัสประจำตัว 566050805
ปริญญา วิทยาศาสตรมหาบัณฑิต (สาขาวิชาวิทยาการคอมพิวเตอร์)
ภาควิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.สายชล ใจเย็น
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม -

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
ผศ.ดร.อนันตพร หรรษคุณาฒัย ประธานกรรมการ	
ดร.สันทิฏฐ์ นรินิน อาจารย์บัณฑิตประจำ (ในสาขาวิชาที่เกี่ยวข้อง)	
ดร.ธนัสนี เพียรตระกูล ผู้ทรงคุณวุฒิจากภายนอกสถาบันฯ	
ดร.สายชล ใจเย็น อาจารย์ที่ปรึกษาวิทยานิพนธ์	

วัน/ เดือน/ ปี ที่สอบ วันที่ 30 มิถุนายน พ.ศ.2558
สถานที่สอบ ห้อง 306 อาคารปฏิบัติการใหม่ ชั้น 3

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 28 **เดือน** ก.ค. **พ.ศ.** 58

หัวข้อวิทยานิพนธ์ การทำเหมืองความคิดเห็นสำหรับร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียมและการคัดเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์

ชื่อนักศึกษา นายนิพัทธ์ คล้ายโพธิ์

รหัสประจำตัว 55650805

ปริญญา วิทยาศาสตรมหาบัณฑิต

ภาควิชา วิทยาการคอมพิวเตอร์

พ.ศ. 2558

อาจารย์ที่ปรึกษาวิทยานิพนธ์ อาจารย์ ดร.สายชล ใจเย็น

บทคัดย่อ

ในปัจจุบันร้านอาหารไทยเป็นที่นิยมของคนทั่วโลก บนเครือข่ายทางสังคม ลูกค้ามีการแสดงความคิดเห็นเกี่ยวกับสินค้าหรือบริการของร้านอาหารผ่านทางเว็บไซต์เป็นจำนวนมาก ส่วนใหญ่ความคิดเห็นเหล่านี้จะสามารถพบได้บนเว็บไซต์แนะนำการท่องเที่ยวและความคิดเห็นที่มีเป็นจำนวนมากนี้ ทำให้ยากต่อการวิเคราะห์และแยกแยะความคิดเห็นของลูกค้าที่มีต่อสินค้าหรือบริการ การทำเหมืองความคิดเห็นจะช่วยจำแนกประเภทความคิดเห็นของลูกค้า เพื่อนำมาใช้ปรับปรุงสินค้าและบริการของธุรกิจได้ ดังนั้น ในงานวิจัยนี้ จึงมีวัตถุประสงค์หลักเพื่อนำเสนอวิธีการทำเหมืองความคิดเห็นโดยใช้โครงข่ายประสาทเทียมเพื่อจำแนกประเภทความคิดเห็นและใช้การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์เพื่อเลือกชุดของคำสำคัญที่เหมาะสมต่อการจำแนกความคิดเห็น การเลือกเฉพาะชุดของคำสำคัญที่เหมาะสม จะทำให้ข้อมูลมีขนาดลดลงและส่งผลให้เวลาในการประมวลผลของโครงข่ายประสาทเทียมลดลงด้วย จากการทดลองพบว่า การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์และการใช้โครงข่ายประสาทเทียมสามารถจำแนกประเภทของความคิดเห็นของร้านอาหารไทยได้อย่างมีประสิทธิภาพและแม่นยำโดยใช้เวลาในการเรียนรู้ที่น้อยลง

คำสำคัญ: การจำแนกประเภท คัดเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น ซัพพอร์ตเวคเตอร์แมชชีน เรเดียลเบสิสฟังก์ชัน

Thesis Title	Opinion Mining for Thai Restaurant Reviews using Neural Networks and mRMR Feature Selection
Student Name	Mr.Niphat Claypo
StudentID	55650805
Degree	Master of Science
Department	Computer Science
Year	2015
Thesis Advisor	Dr. Saichon Jaiyen

ABSTRACT

Currently, Thai restaurants are popular around the world. There are tons of reviews related to foods and services in social networking web sites. These customer reviews can be mostly found on travel information web sites and the great number of them is difficult to analyze and classify the opinions of customers towards the foods and services. In business perspectives, the models of opinion mining are utilized for classifying attitudes of the customers in order to improve their products and services. Therefore, a main objective of this research is to propose the opinion mining based on the artificial neural network for classifying the positive and negative reviews, and on the mRMR feature selection for selecting appropriate sets of keywords for the classification. Especially, this selection reduces the number of features in the data set. Therefore, computational times of learning algorithms used by neural networks are also reduced. From the experimental results, they have shown that the mRMR feature selection with the neural networks is an effective model for classifying Thai restaurant reviews with high accuracy and time efficiency.

Keywords: Classification, mRMR Feature selection, multilayer perceptron (MLP), Support Vector Machine (SVM), Radial Basis Function (RBF)

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้มีอาจจะสำเร็จลุล่วงไปด้วยดี หากมิได้รับความช่วยเหลือแนะนำ คำชี้แจงความรู้ และความเอาใจใส่จาก ดร.สายชล ใจเย็น ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งได้สละเวลามาให้กับข้าพเจ้าอย่างเต็มที่ จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ. ดร.อนันตพร หรรษคุณาฒัย ดร. สันทิฏฐ์ นรบิน และดร. ธนัสนี เพียรตระกูล คณะกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำตลอดจนชี้แนะจนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ขอขอบพระคุณ อาจารย์ผู้สอนรายวิชาต่างๆให้กับข้าพเจ้าทำให้ระหว่างศึกษาในรายวิชา ต่างๆนั้น ได้สามารถนำมาปรับและประยุกต์ใช้ในงานวิจัยชิ้นนี้ได้ โดยเฉพาะ ผศ. ดร.อนันตพร หรรษคุณาฒัย ที่ได้สอนทฤษฎีการทำเหมืองข้อมูลให้แก่ข้าพเจ้า

ขอขอบพระคุณบิดา มารดา และญาติพี่น้องที่ได้ให้กำลังใจ และให้การสนับสนุนในด้านต่างๆระหว่างศึกษาจนสำเร็จลุล่วงไปด้วยดี

สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกๆท่าน ตลอดจนพี่น้อง เพื่อนๆ และน้องๆทุกคน

นิพัทธ์ คล้ายโพธิ์

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันร้านอาหารไทยเป็นที่นิยมของคนทั่วโลก และลูกค้าที่เข้ามาใช้บริการร้านอาหารสามารถแสดงความคิดเห็นเกี่ยวกับสินค้าหรือบริการของร้านอาหารผ่านทางเว็บไซต์ได้ บนเว็บไซต์แนะนำการท่องเที่ยวลูกค้าที่เคยมาใช้บริการของแต่ละร้านอาหารสามารถเข้ามาแนะนำอาหารหรือบริการของทางร้านได้ ความคิดเห็นของลูกค้าที่เคยมาใช้บริการของทางร้านแล้วนั้นเป็นข้อมูลอย่างดีที่จะช่วยให้ผู้ที่สนใจในร้านอาหารนั้นตัดสินใจที่จะมาใช้บริการของทางร้าน เพราะจะได้ทราบข้อมูลเกี่ยวกับร้านอาหารนั้นๆ ว่าดีหรือไม่ดีซึ่งจะช่วยในการตัดสินใจมาใช้บริการของร้านอาหารที่สนใจ ในทางตรงกันข้ามความคิดเห็นของลูกค้าที่มาใช้บริการของทางร้านอาหารสามารถนำมาใช้ปรับปรุงสินค้าและบริการของทางร้านได้ จากการแสดงความคิดเห็นของลูกค้าบนเว็บไซต์แนะนำการท่องเที่ยวทำให้มีความคิดเห็นเป็นจำนวนมากซึ่งยากต่อการวิเคราะห์ความคิดเห็น

การทำเหมืองความคิดเห็น (Opinion Mining) เป็นวิธีการที่ช่วยในการวิเคราะห์ความคิดเห็น เนื่องจากความคิดเห็นมีเป็นจำนวนมากและเป็นเรื่องยากที่ต้องอ่านความคิดเห็นทุกความคิดเห็นเพื่อทำการวิเคราะห์ [1] แต่ละความคิดเห็นนั้นมีรูปแบบที่ไม่แน่นอนสั้นบ้างยาวบ้าง บางประโยคก็ไม่เกี่ยวกับสิ่งที่แสดงความคิดเห็น บางความคิดเห็นก็ประกอบไปด้วยตัวอักษรที่ไม่มีความหมาย ดังนั้นงานวิจัยนี้จึงได้เสนอวิธีการทำเหมืองข้อมูลความคิดเห็นโดยใช้โครงข่ายประสาทเทียม (Neural Network) และการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์เพื่อใช้ในการประมวลผลข้อความ ข้อมูลที่ใช้ในการทดลองเป็นความคิดเห็นเกี่ยวกับร้านอาหารไทย แต่ละความคิดเห็นผ่านขั้นตอนการเตรียมข้อมูลและการแปลงข้อมูลเพื่อให้ข้อมูลมีความเหมาะสมกับวิธีการที่จะใช้ในการจำแนกประเภทความคิดเห็น

1.2 วัตถุประสงค์

1) เพื่อพัฒนาวิธีการคัดเลือกคำสำคัญแบบอัตโนมัติจากข้อความในความคิดเห็นเพื่อนำมาใช้เป็นคุณลักษณะของข้อมูล

2) เพื่อจำแนกประเภทความคิดเห็นของร้านอาหารไทยด้วยการทำเหมืองข้อมูลความคิดเห็นโดยใช้โครงข่ายประสาทเทียมร่วมกับเทคนิคการคัดเลือกคุณลักษณะ

1.3 ขอบเขตการวิจัย

1) วิทยานิพนธ์นี้พัฒนาเครื่องมือในการทำเหมืองความคิดเห็นจากข้อมูลความคิดเห็นของร้านอาหารไทย

2) ใช้โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้นในการจำแนกประเภทความคิดเห็น

3) เทคนิคการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์เพื่อนำไปใช้ในการคัดเลือกคุณลักษณะของชุดข้อมูล

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถจำแนกประเภทความคิดเห็นของร้านอาหารไทยได้อย่างถูกต้อง

2) สามารถวิเคราะห์ความคิดเห็นของลูกค้าที่มีต่อร้านอาหารจำนวนมากได้

3) สามารถนำไปประยุกต์ใช้กับความคิดเห็นของสินค้าประเภทอื่นได้

4) สามารถทราบถึงความคิดเห็นของลูกค้าที่เคยมาใช้บริการว่าเป็นไปทางบวกหรือลบ

5) สามารถเป็นเครื่องมือที่ช่วยในการตัดสินใจเลือกใช้บริการร้านอาหารได้

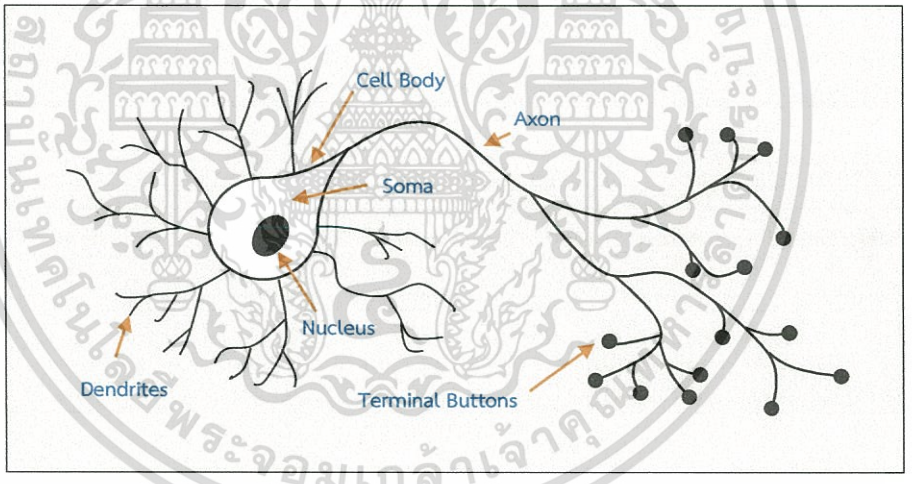
6) สามารถนำไปประยุกต์ใช้เพื่อพัฒนาอุตสาหกรรมการท่องเที่ยวได้

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

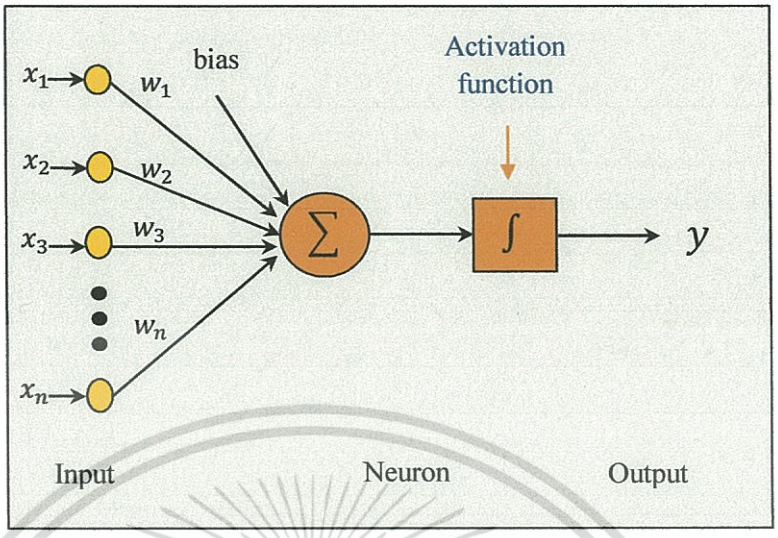
2.1 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นสาขาหนึ่งในสาขาปัญญาประดิษฐ์ (Artificial Intelligent) ซึ่งเป็นโมเดลทางคณิตศาสตร์ที่จำลองการทำงานของสมองมนุษย์เพื่อทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้เหมือนมนุษย์ได้ โครงข่ายประสาทเทียมเป็นวิธีการที่นิยมนำมาแก้ปัญหาการจำแนกประเภทข้อมูล (Classification Problems) ที่สามารถจำแนกข้อมูลได้อย่างถูกต้องและมีประสิทธิภาพ การเรียนรู้ของโครงข่ายประสาทเทียมนั้นแบ่ งออกเป็น 3 ประเภท ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมแรง (Reinforcement Learning)



รูปที่ 2.1 โครงสร้างเซลล์ประสาท

โครงข่ายประสาทเทียมประกอบไปด้วยโหนด (Node) จำนวนมากเชื่อมต่อกัน โดยที่โหนดจำลองมาจากเซลล์ประสาทในสมองมนุษย์ประกอบด้วย ไซแนป (Synapse) เดนไดร์ท (Dendrite) ตัวเซลล์ (Soma) และแอกซอน (Axon) ดังรูปที่ 2.1 โดยในเซลล์ประสาทเทียมจะมีฟังก์ชันกระตุ้น (Activation Function or Transfer Function) เป็นตัวกำหนดสัญญาณส่งออก [2] ดังรูปที่ 2.1



รูปที่ 2.2 โครงสร้างภายในของโครงข่ายประสาทเทียม

ในรูปที่ 2.2 แสดงโครงสร้างของโครงข่ายประสาทเทียมซึ่งประกอบด้วย

1. ข้อมูลนำเข้า (Input) เป็นข้อมูลที่เป็นตัวเลขที่ถูกป้อนให้กับโครงข่ายประสาทเทียม

2. ข้อมูลส่งออก (Output) เป็นผลลัพธ์ที่ได้จากการเรียนรู้ของโครงข่ายประสาทเทียม

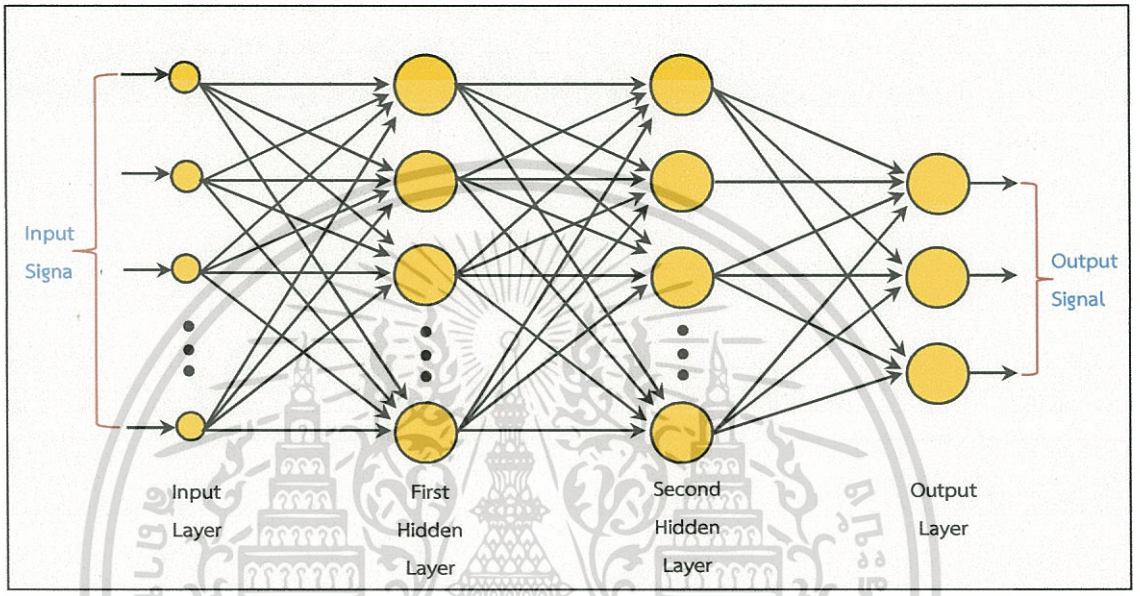
3. ค่าน้ำหนัก (Weights) หรือเรียกอีกอย่างหนึ่งว่า ค่าความรู้ (Knowledge) เป็นค่าที่ได้จากการเรียนรู้ของโครงข่ายประสาทเทียม ค่าน้ำหนักนั้นจะกำกับบนเส้นเชื่อมโยงทุกเส้นเพื่อเป็นทักษะที่นำไปใช้กับข้อมูลอื่นๆ

4. ฟังก์ชันผลรวม (Summation Function) เป็นฟังก์ชันที่เป็นผลรวมของผลคูณระหว่างข้อมูลนำเข้าและค่าน้ำหนัก

5. ฟังก์ชันกระตุ้น (Activation Function) เป็นฟังก์ชันที่ทำการแปลงค่าผลลัพธ์จากฟังก์ชันผลรวมในแต่ละโหนดให้อยู่ในช่วงที่กำหนด

2.1.1 โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น

โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron Neural Network: MLP) เป็นโครงข่ายประสาทเทียบแบบมีโครงสร้างหลายชั้น



รูปที่ 2.3 โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น

จากรูปที่ 2.3 โครงสร้างภายในของโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นนั้นประกอบด้วย ชั้นข้อมูลนำเข้า (Input Layer) ชั้นข้อมูลส่งออก (Output Layer) และชั้นที่อยู่ระหว่างชั้นข้อมูลนำเข้าและชั้นข้อมูลส่งออกคือชั้นซ่อน (Hidden Layer) ในโครงข่ายประสาทเทียมประกอบด้วยโหนด (Node) หลายๆ โหนดเชื่อมต่อกันเป็นโครงข่าย กลุ่มของโหนดเรียกว่าชั้น (Layer) โดยที่แต่ละโหนดจะมีเส้นเชื่อมกันทุกโหนดรวมกันเป็นโครงข่าย แต่ละเส้นเชื่อมของโครงข่ายจะมีค่าน้ำหนัก (Weight value) กำกับอยู่ทุกเส้นเชื่อม ชั้นซ่อนของโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นจะประประกอบไปด้วยโหนดหลายๆ โหนดที่เรียกว่า โหนดชั้นซ่อน (Hidden node) โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นจะเหมาะสำหรับงานที่มีความซับซ้อนและงานที่ต้องการความถูกต้องสูง โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นนั้นมีกระบวนการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือการเรียนรู้แบบแพร่กระจายย้อนกลับ (Backpropagation) [3] โดยที่ข้อมูลจะถูกแบ่งออกเป็นสองชุด คือ ชุดข้อมูลฝึกสอน (Training Set) และ ชุดข้อมูลทดสอบ (Testing Set)

โดยชุดข้อมูลฝึกสอนประกอบด้วยข้อมูลตัวอย่าง และคำตอบที่ต้องการ เพื่อใช้ในการสอนโครงข่ายประสาทเทียม [3]

2.1.2 การเรียนรู้แบบแพร่กระจายย้อนกลับ

สำหรับกระบวนการเรียนรู้แบบแพร่กระจายย้อนกลับ (Backpropagation) [4, 5] ประกอบด้วยการทำงานสองส่วนย่อย คือ การส่งผ่านไปข้างหน้า (Forward Pass) และ การส่งผ่านย้อนกลับ (Backward Pass) กระบวนการทำงานของการเรียนรู้แบบแพร่กระจายย้อนกลับมีดังต่อไปนี้

การส่งผ่านไปข้างหน้า (Forward Pass)

1. สุ่มค่าน้ำหนัก (Weight) และค่าไบแอส (bias)
2. รับข้อมูลฝึกสอนจากชุดของข้อมูลฝึกสอน $D = \{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$ โดยที่ $\mathbf{x}_i$ คือ ข้อมูลนำเข้า d_i คือ ผลลัพธ์ที่ต้องการ
3. คำนวณค่าผลลัพธ์ (Output) ของแต่ละโหนดในชั้นซ่อนโดยใช้สมการ 2.1

$$y_j = \varphi_j(v_j) \tag{2.1}$$

โดยที่ $\varphi_j(v_j) = \frac{1}{1+e^{-v_j}}$

$v_j = \sum_{i=1}^{N} x_i w_{ij}$

w_{ij} คือ ค่าน้ำหนักของเส้นเชื่อมจากโหนด i ไปยังโหนด j

4. คำนวณค่าผลลัพธ์ (Output) ของแต่ละโหนดในชั้นข้อมูลส่งออก โดยใช้สมการ 2.2

$$y_k = \varphi_k(v_k) \tag{2.2}$$

โดยที่ $v_k = \sum_{j=1}^{N} y_j w_{jk}$

การส่งผ่านย้อนกลับ (Backward Pass)

5. คำนวณค่า Error Gradient ของแต่ละโหนดข้อมูลในชั้นข้อมูลส่งออกโดยใช้สมการ 2.3

$$\delta_k = (d_k - y_k) y_k (1 - y_k) \tag{2.3}$$

โดยที่ y_k คือ ค่าผลลัพธ์ของโหนดที่ k ในชั้นข้อมูลส่งออก

6. คำนวณค่า Error Gradient ของแต่ละโหนดในชั้นซ่อนโดยใช้สมการ 2.4

$$\delta_j = y_j(1 - y_j) \sum_k \delta_k w_{jk} \tag{2.4}$$

โดยที่ y_j คือ ค่าผลลัพธ์ของโหนดที่ j ในชั้นซ่อน

7. ปรับค่าน้ำหนักของทุกเส้นเชื่อมโดยใช้สมการที่ 2.5

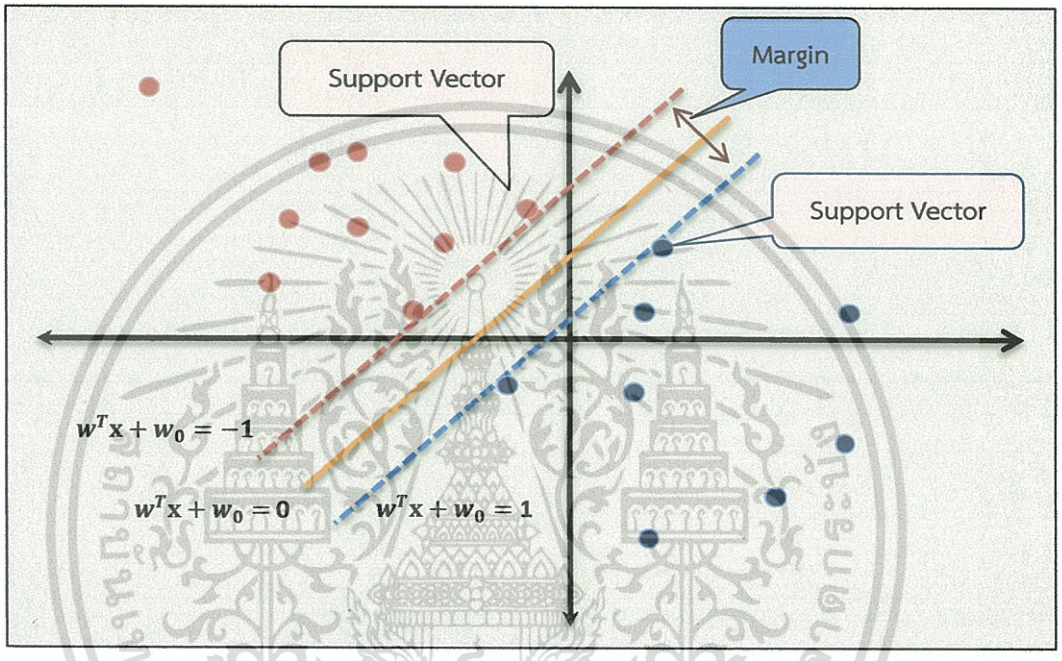
$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{2.5}$$

โดยที่ $\Delta w_{ij} = \eta \delta_j y_i$ และ η คือ อัตราการเรียนรู้ (Learning Rate)

การปรับเปลี่ยนค่าน้ำหนักจะเริ่มจากชั้นข้อมูลส่งออกแพร่ย้อนกลับไปจนถึงชั้นข้อมูลนำเข้า ทำซ้ำขั้นตอนที่ 2 ถึง 7 จนกระทั่งโครงข่ายประสาทเทียมสามารถจำแนกข้อมูลตัวอย่างได้อย่างถูกต้อง

2.2 ซัพพอร์ตเวคเตอร์แมชชีน

ซัพพอร์ตเวคเตอร์แมชชีน *Support Vector Machine* (SVM) เป็นเทคนิคหนึ่งในเทคนิคการเรียนรู้ของเครื่องที่ใช้ในการแก้ปัญหาการจำแนกประเภทข้อมูล SVM มีวิธีการทำงานโดยการหา Optimal Hyperplane ที่สามารถแบ่งข้อมูลได้ดีที่สุดภายใต้เงื่อนไข [3] ดังรูปที่ 2.3



รูปที่ 2.4 Support Vector Machine

จากรูปที่ 2.4 วิธีการของ SVM นั้นจะทำการหาสมการเชิงเส้นที่อยู่ตรงกลางระหว่างเส้นแบ่งพอดีเรียกสมการเชิงเส้นนี้ว่า Optimal Hyperplane โดยการเรียนรู้จากชุดข้อมูลฝึกสอน $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$ โดยที่ $\mathbf{x}_i$ คือชุดข้อมูลฝึกสอน และ d_i คือคลาสคำตอบ โดยสมมติให้คลาสคำตอบแทนด้วย $d_i = +1$ สำหรับคลาส +1 และ $d_i = -1$ สำหรับคลาส -1 ดังสมการที่ 2.6

$$\mathbf{w}^T\mathbf{x} + w_0 = \mathbf{0} \tag{2.6}$$

โดยที่ $\mathbf{w}$ คือ เว็กเตอร์ค่าน้ำหนัก

w_0 คือ ค่า bias

จากสมการที่ 2.6 เราสามารถเขียนได้ดังนี้

$$\mathbf{w}^T\mathbf{x} + w_0 \geq 1 \text{ สำหรับ } d_i = +1$$

$$\mathbf{w}^T\mathbf{x} + w_0 < 1 \text{ สำหรับ } d_i = -1 \quad (2.7)$$

SVM จะทำการหา Optimal Hyperplane ที่มีระยะห่างของขอบ (Margin) กว้างที่สุดโดยดูจากระยะทางระหว่าง Support Vector และ Optimal Hyperplane โดยความกว้างของขอบคำนวณได้จากสมการที่ 2.8

$$2r = \frac{2}{\|\mathbf{w}\|} \quad (2.8)$$

โดยที่ r คือ ระยะทางระหว่าง Support Vector และ Optimal Hyperplane

เนื่องจากการหาความกว้างมากสุดของขอบสมมูล (equivalent) กับการหาค่าขนาดของเวกเตอร์น้ำหนักน้อยสุด ดังนั้น SVM จะทำการหา Optimal Hyperplane โดยใช้ Lagrangian Function ของขนาดของเวกเตอร์น้ำหนักภายใต้เงื่อนไขของสมการที่ 2.7 ดังแสดงในสมการที่ 2.9

$$J(\mathbf{w}, w_0, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i[d_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1] \quad (2.9)$$

โดยที่ α_i คือ Lagrange Multipliers

จากสมการที่ 2.9 เราสามารถหาค่าน้ำหนักที่เหมาะสมที่สุดได้ดังสมการที่ 2.10

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \mathbf{x}_i \quad (2.10)$$

สำหรับข้อมูลที่ไม่สามารถแบ่งแบบเชิงเส้นได้ในปริภูมิข้อมูลนำเข้า SVM จะหา Optimal Hyperplane โดยใช้ Kernel Function ดังสมการที่ 2.11

$$\sum_{i=1}^{N} \alpha_i d_i k(\mathbf{x}, \mathbf{x}_i) = 0 \quad (2.11)$$

สำหรับ Kernel Function ที่นิยมใช้กันมากได้แก่ Kernel Function แบบ Radial Basis Function ดังสมการ 2.12

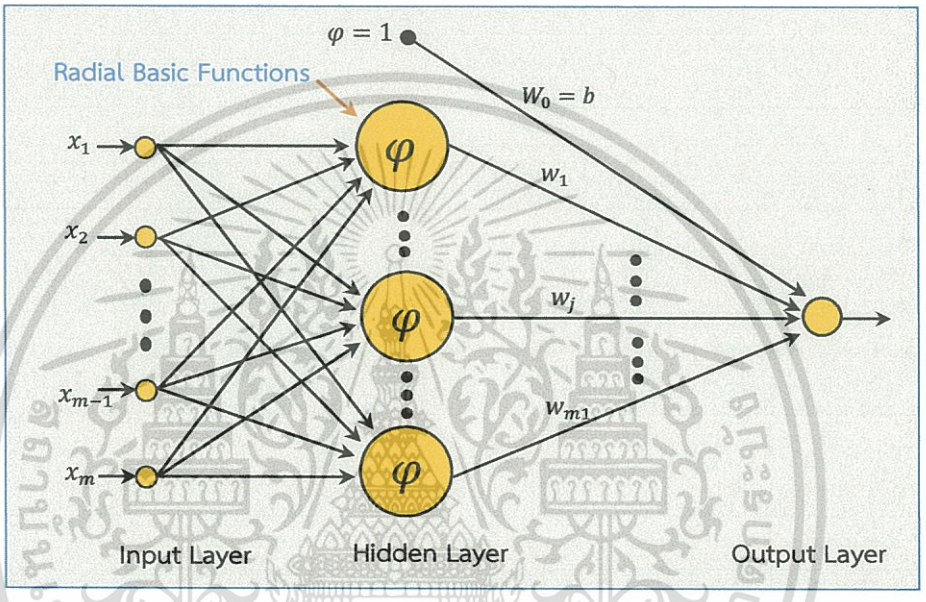
$$k(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{1}{2\sigma^2} \| \mathbf{x} - \mathbf{x}_i \|^2\right) \quad (2.12)$$

โดยที่ σ คือ ค่าพารามิเตอร์ความกว้างเชิงรัศมี

SVM แก้ปัญหาการหาค่าน้ำหนักโดยการหาค่า Lagrange Multipliers แทนการปรับค่าหนักเพื่อหาค่าน้ำหนักที่เหมาะสมจึงทำให้ SVM ทำงานได้รวดเร็ว ลดปัญหา Over Fitting และ Hyperplane ที่ได้ยังเป็น Hyperplane ที่เหมาะสมที่สุดอีกด้วย

2.3 เรเดียลเบสิสฟังก์ชัน

เรเดียลเบสิสฟังก์ชัน *Redial Basis Function Neural Network* (RBF) เป็นโครงข่ายประสาทเทียมที่ใช้ Redial Basis Function เป็นฟังก์ชันกระตุ้น (Activation Functions) โครงข่ายประสาทเทียมแบบ RBF นั้นนิยมใช้ในการจำแนกประเภทข้อมูล ซึ่งโครงสร้างของ RBF [5] แสดงดังรูปที่ 2.5

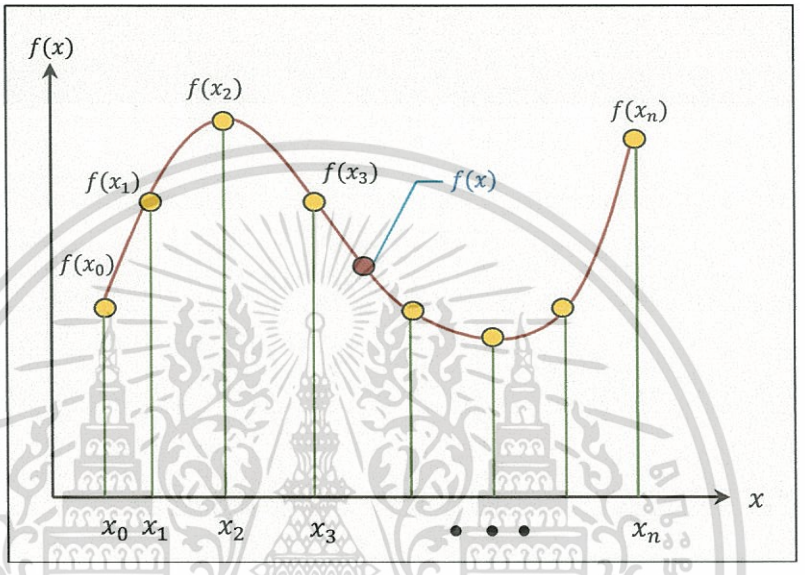


รูปที่ 2.5 Redial Basis Function Neural Network

จากรูปที่ 2.5 โครงสร้างของ RBF ประกอบด้วย ชั้นข้อมูลนำเข้า (Input Layer) ชั้นข้อมูลส่งออก (Output Layer) และชั้นที่อยู่ระหว่างชั้นข้อมูลนำเข้าและชั้นข้อมูลส่งออกคือชั้นซ่อน (Hidden Layer) ที่มี Gaussion Function เป็นฟังก์ชันกระตุ้น โดยที่เส้นเชื่อมระหว่างชั้นข้อมูลนำเข้าและชั้นข้อมูลส่งออกของ RBF จะไม่มีค่าน้ำหนักกำกับอยู่บนเส้นเชื่อมโยง โครงข่ายประสาทเทียมแบบ RBF นั้นมีแนวคิดมาจากปัญหาการประมาณค่าในช่วง

2.3.1 การประมาณค่าในช่วง

ปัญหาการประมาณค่าในช่วง (Interpolation Problem) เป็นปัญหาการหาฟังก์ชันที่สามารถลากผ่านทุกจุดข้อมูลได้ดังรูปที่ 2.6



รูปที่ 2.6 การประมาณค่าในช่วง

รูปที่ 2.6 แสดงการประมาณค่า $f(x)$ ที่ลากผ่านชุดข้อมูล โดยการประมาณค่าจะทำการสร้างเส้นโค้งที่สามารถลากผ่านจุดทุกจุดในช่วงข้อมูล และประมาณค่าจุดบนเส้นโค้งนั้นโดยใช้ฟังก์ชัน $f(x)$

2.3.2 โครงข่ายประสาทเทียมแบบ RBF

โครงข่ายประสาทเทียมแบบ RBF มีขั้นตอนการเรียนรู้ดังนี้

1. รับชุดข้อมูลฝึกสอน $D = \{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$ โดยที่ $\mathbf{x}_i$ คือ ข้อมูลนำเข้า และ d_i คือ ผลลัพธ์ที่ต้องการ
2. คำนวณค่าผลลัพธ์ในชั้นซ่อนโดยใช้ Gaussian Function ดังสมการที่ 2.13

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_i \varphi(\| \mathbf{x} - \mathbf{x}_i \|) \tag{2.13}$$

โดยที่ $\varphi(r) = exp\left(-\frac{r^2}{2\sigma^2}\right)$

$\|.\|$ คือ นอร์มแบบยุคลิด (Euclidean Norm)

$\mathbf{x}_i$ คือ จุดศูนย์กลางของ RBF

จากสมการที่ 2.13 จะได้เมทริกซ์ของข้อมูลที่ผ่าน Gaussion Function ดังสมการที่ 2.14

$$\Phi = [\varphi_{ij}]N \times N \tag{2.14}$$

เรียกเมทริกซ์ของ Φ ว่า Interpolation matrix ที่ประกอบด้วยสมาชิก φ_{ij} โดยเมทริกซ์นี้จะถูกนำมาคูณกับเวกเตอร์ค่าน้ำหนักเพื่อหาคำตอบดังสมการที่ 2.15

$$\begin{bmatrix} \varphi_{11}\ \varphi_{12} \cdots \varphi_{1N} \\ \varphi_{21}\ \varphi_{22} \cdots \varphi_{2N} \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ \varphi_{N1}\ \varphi_{N2} \cdots \varphi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ . \\ d_N \end{bmatrix} \tag{2.15}$$

โดยที่ $\varphi_{ji} = \varphi(\| \mathbf{x_j} - \mathbf{x}_i \|), (i,j) = 1, 2, \ldots N$ (2.16)

3. คำนวนหาค่าน้ำหนักของแต่ละเส้นเชื่อมระหว่างชั้นซ่อนและชั้นข้อมูลส่งออกโดยสมการที่ 2.17

$$\Phi \mathbf{w} = \mathbf{d} \tag{2.17}$$

โดยที่ $\mathbf{d} = [d_1,\ d_2, \ldots, d_N]^T$ เป็นเวกเตอร์ของคำตอบ

$\mathbf{w} = [w_1,\ w_2, \ldots, w_N]^T$ เป็นเวกเตอร์ค่าน้ำหนัก

N คือ ขนาดของชุดข้อมูลฝึกสอน

4. คำนวนหาค่าน้ำหนัก $\mathbf{w}$ ดังสมการที่ 2.18

$$\mathbf{w} = \Phi^{+}\mathbf{d} \tag{2.18}$$

โดยที่ Φ^{+} คือ Pseudo-inverse ของเมทริกซ์ Φ

คำนวณค่าผลลัพธ์ของโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชันได้ ดังสมการที่ 2.19

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_i \varphi(\| \mathbf{x} - \mathbf{x}_i \|) \quad (2.19)$$

2.4 การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์

การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ (mRMR Feature Selection) เป็นวิธีการที่นิยมใช้เลือกคุณลักษณะกับข้อมูลที่มีคุณลักษณะ (Features) จำนวนมาก [6] โดยวิธี mRMR จะเลือกคุณลักษณะที่มีค่าความสัมพันธ์กับคลาสคำตอบที่สูงและมีค่าความซ้ำซ้อนกันระหว่างคุณลักษณะด้วยกันน้อย เรียกว่า วิธี Max-Relevance and Min-Redundancy (mRMR) โดยวิธีการหาความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบ $I(f_i, c)$ และ คุณลักษณะกับคุณลักษณะ $I(f_i, f_j)$ นั้นอาศัยค่า Mutual Information (MI) ในสมการที่ 2.20

$$I(x, y) = \iint p(x, y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (2.20)$$

โดยที่ $p(x, y)$ คือ ค่าความน่าจะเป็นร่วมของ x และ y

$p(x)$ คือ ค่าความน่าจะเป็นของ x

$p(y)$ คือ ค่าความน่าจะเป็นของ y

โดยค่า MI จะเป็นค่าที่บอกปริมาณความขึ้นต่อกันของตัวแปรสองตัวแปร เป็นค่าที่นิยมใช้กันในการเลือกคุณลักษณะ ขั้นตอนการเลือกคุณลักษณะประกอบด้วย 2 ขั้นตอนหลักดังนี้

2.4.1. ขั้นตอนการหาความสัมพันธ์ระหว่างคุณลักษณะที่มีความซ้ำซ้อนกัน

วิธีการนี้เรียกว่า Min-Redundancy โดยหาความสัมพันธ์ระหว่างคุณลักษณะกับคุณลักษณะ $I(f_i, f_j)$ ดังสมการที่ 2.21

$$\min R(S), R = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j). \quad (2.21)$$

โดยที่ $I(f_i, f_j)$ คือ ค่า Mutual Information ของคุณลักษณะ f_i และ f_j

f_i คือ คุณลักษณะที่ i ของข้อมูล

S คือ เซตของคุณลักษณะ

$|S|$ คือ จำนวนของคุณลักษณะทั้งหมดใน S

ในขั้นตอนนี้จะทำการเลือกคุณลักษณะที่มีค่าของการขึ้นต่อกันที่มีค่าน้อยทำให้ได้คุณลักษณะที่ไม่ซ้ำซ้อนกันออกมา

2.4.2 ขั้นตอนการหาความสัมพันธ์ระหว่างคุณลักษณะและคลาสคำตอบ

วิธีการนี้เรียกว่า Max-Relevance โดยใช้ค่า MI เพื่อหาความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบ $I(f_i, c)$ โดย Max-Relevance จะเลือกคุณลักษณะที่มีค่า MI ที่มีค่ามากที่สุดดังสมการที่ 2.22

$$\max R(S, c), D = \frac{1}{|S|} \sum_{f_i \in S} I(f_i, c). \quad (2.22)$$

โดยที่ $I(f_i, c)$ คือ ค่า Mutual Information ของคุณลักษณะ f_i และคลาสคำตอบ c

$|S|$ คือ จำนวนของคุณลักษณะทั้งหมดใน S

คุณลักษณะที่เลือกเป็นคุณลักษณะที่มีความสัมพันธ์กับคลาสคำตอบ (Max-Dependency) ในขั้นตอนนี้จะทำการเลือกคุณลักษณะที่มีค่าของการขึ้นต่อกันมากที่สุดทำให้ได้คุณลักษณะที่เกี่ยวข้องกับคลาสคำตอบออกมา หลังจากหาค่าความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบและความซ้ำซ้อนของแต่ละคุณลักษณะแล้ว จะทำการรวมสองข้อจำกัดข้างต้นเข้าด้วยกันเรียกว่า Max-Relevance and Min-Redundancy (mRMR) โดยการรวมสมการ D และ R จากการเลือกคุณลักษณะข้างต้นเพื่อหาชุดของคุณลักษณะที่เหมาะสมที่สุดดังสมการที่ 2.23

$$\max \Phi\ (D, R), \Phi = D - R \quad (2.23)$$

โดยที่ D คือ Max-Relevance

R คือ Min-Redundancy

จากสมการเป็นการเลือกชุดของคุณลักษณะที่เหมาะสมระหว่างคุณลักษณะที่มีความสัมพันธ์กับคลาสคำตอบสูงและมีความซ้ำซ้อนกันต่ำเรียกวิธีนี้ว่า Max-Dependency คุณสมบัติที่ผ่านการเลือกโดยวิธี mRMR จะเป็นชุดของคุณลักษณะที่มีความสำคัญที่สามารถบ่งบอกประเภทของข้อมูลได้

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 Chen และ Zimbra [1] รีวิววิธีการทำเหมืองความคิดเห็นจากข้อมูลความคิดเห็นบนเว็บไซต์เช่น ข้อมูลความคิดเห็นบนบล็อกเว็บไซต์เครือข่ายสังคม และทวิตเตอร์ เป็นต้น เพื่อทำการวิเคราะห์ความคิดเห็นทำให้สามารถเข้าใจความรู้สึกของผู้ใช้งานได้ดียิ่งขึ้นโดยใช้วิธีหลักการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ในประมวลผลข้อความ หรือวิธีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อทำการประมวลผลข้อความ

2.5.2 Neethu และ Rajasree [7] นำเสนอวิธีการวิเคราะห์ความเชื่อมั่นบนทวิตเตอร์ซึ่งเป็นข้อมูลเกียวกับผลิตภัณฑ์อิเล็กทรอนิกส์โดยใช้วิธีการตัวจำแนกแบบเบย์อย่างง่าย และซัพพอร์ตเวกเตอร์แมชชีน มาทำการวิเคราะห์ความคิดเห็น ข้อมูลประกอบด้วยความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบที่ถูกโพสต์บนทวิตเตอร์

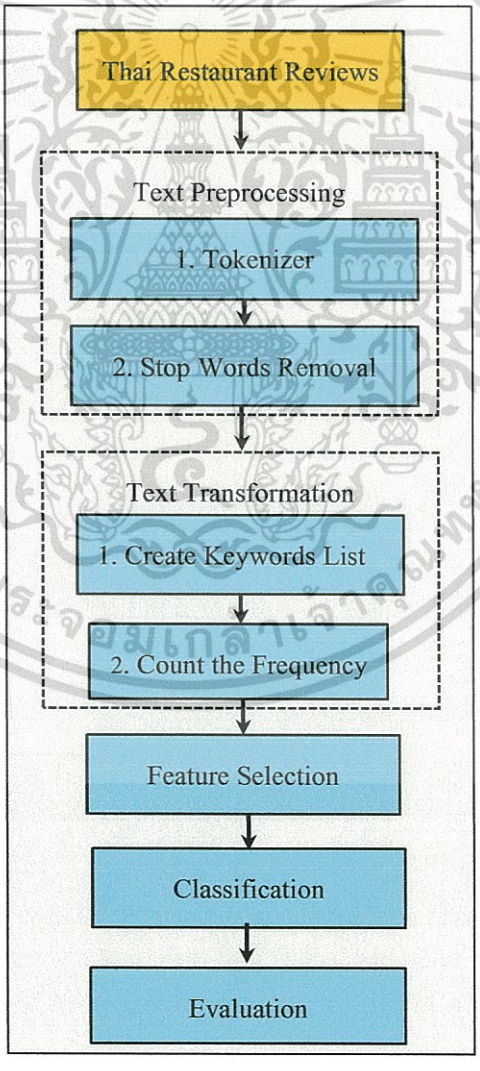
2.5.3 Singh, Piryani และ Uddin [8] นำเสนอวิธีการวิเคราะห์บทวิจารณ์ภาพยนตร์ โดยใช้แบบแผน SentiWordNet ทำการจัดลำดับแต่ละบทวิจารณ์ จากการทดลองพบว่าวิธีการที่นำเสนอวิเคราะห์ความคิดเห็นได้อย่างถูกต้อง

2.5.4 Raut และ Londhe [9] เสนอวิธีในการทำเหมืองความคิดเห็นจากข้อมูลความคิดเห็นของผู้ใช้บริการโรงแรมโดยใช้วิธีการเรียนรู้ของเครื่องประกอบไปด้วยตัวจำแนกแบบเบย์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้ตัดสินใจ จากผลการทดลองพบว่าวิธีการเรียนรู้ของเครื่องให้ค่าความถูกต้องในการจำแนกความคิดเห็น 87% จากการจำแนกประเภทความคิดเห็นเชิงบวกและเชิงลบ

บทที่ 3

การจำแนกประเภทความคิดเห็นร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียม และการเลือกคุณสมบัติแบบเอ็มอาร์เอ็มอาร์

งานวิจัยนี้นำเสนอวิธีการทำเหมืองข้อมูลความคิดเห็นเพื่อจำแนกประเภทความคิดเห็นของร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (MLP) และการเลือกคุณสมบัติแบบเอ็มอาร์เอ็มอาร์ (mRMR Feature Selection) ซึ่งขั้นตอนการจำแนกประเภทความคิดเห็นทั้งหมดแสดงดังรูปที่ 3.1



รูปที่ 3.1 ขั้นตอนการจำแนกประเภทความคิดเห็น

ข้อมูลความคิดเห็นของร้านอาหารไทยที่ใช้ในการทดลองนี้ทำการเก็บรวบรวมมาจาก www.tripadvisor.com ซึ่งเป็นเว็บไซต์แนะนำการท่องเที่ยวซึ่งเป็นที่นิยมของคนทั่วโลก จำนวนความคิดเห็นของร้านอาหารไทยได้มาจากการสุ่มเลือกความคิดเห็นจำนวน 1,060 ความคิดเห็น โดยที่ความคิดเห็นของลูกค้าของร้านอาหารที่เก็บรวบรวมมาประกอบด้วยความคิดเห็นเชิงบวก (Positive Reviews) และความคิดเห็นเชิงลบ (Negative Reviews) ความคิดเห็นที่เก็บรวบรวมมาจะเป็นภาษาอังกฤษ เพราะภาษาอังกฤษเป็นภาษากลางที่ใช้แสดงความคิดเห็นมากที่สุด

"The good time in DID"

I know DID from the blog of web page long time ago ,but I has no chance to try on ,so I can't an imagine how we could dinner in the dark? Until our team had the first time on the last December 2014 on Merry Christmas and Happy New Year celebration at Sheraton Grand Hotel, sukhumvit rd.,It was the good time to keep my experience we have got a good mind service from the blind people and I also deep understand the people live in the dark world. I exciting when I got the first menu was served by the blind people but we could be touch the good mind service from them tile I forgot the test of the food. The next menu I felled the food very delicious.

It is the good time in my memory.

I suggest you try on Dine in the dark It became the good experience for you.

รูปที่ 3.2 ตัวอย่างความคิดเห็นของลูกค้า

3.1 การเลือกความคิดเห็น

การเลือกความคิดเห็นของร้านอาหารไทยเพื่อใช้ในการจำแนกประเภทเป็นการเลือกความคิดเห็นของลูกค้าที่มีความรู้สึกต่อร้านอาหารทั้งทางด้านบวกและด้านลบ โดยที่ความคิดเห็นเชิงบวกนั้นจะเลือกความคิดเห็นที่มีคะแนนมากกว่าหรือเท่ากับ 4 หรือ 5 ดาวให้เป็นความคิดเห็นเชิงบวก และ ความคิดเห็นเชิงลบจะเลือกความคิดเห็นที่มีคะแนน 1 หรือ 2 ดาวเป็นความคิดเห็นเชิงลบ โดยความคิดเห็นทั้งสองประเภทจะทำการสุ่มเลือกจากเว็บไซต์ทั้งหมด 1,060 ควมคิดเห็น

ในแต่ละความคิดเห็นที่เลือกมานั้นประกอบไปด้วยข้อความที่แสดงความคิดเห็น ตัวเลข Stop Words และ อักขระที่ไม่ใช่ตัวอักษรจำนวนมากรวมอยู่ในความคิดเห็น ดังนั้นก่อนที่จะทำการจำแนกประเภทความคิดเห็นจึงต้องมีการเตรียมข้อมูลเพื่อให้ข้อมูลเหมาะสมกับวิธีการที่ใช้ในการจำแนกประเภท

ความคิดเห็น โดยทุกความคิดเห็นจะต้องผ่านขั้นตอนการเตรียมข้อมูล และขั้นตอนการแปลงข้อมูลเพื่อสร้างชุดข้อมูล (Datasets) เพื่อนำไปใช้ในการจำแนกประเภท

3.2 ขั้นตอนการเตรียมข้อมูล (Text Preprocessing)

ขั้นตอนการเตรียมข้อมูล เป็นขั้นตอนที่ใช้สำหรับการเตรียมข้อมูลให้เหมาะสมในการจำแนกประเภทความคิดเห็น และกำจัดข้อมูลที่ไม่จำเป็นออกไปเพื่อลดภาระในการประมวลผลข้อมูล ซึ่งขั้นตอนนี้เป็นขั้นตอนที่ใช้ในการเตรียมข้อมูลความคิดเห็นก่อนทำการแปลงข้อมูล ขั้นตอนการเตรียมข้อมูลแบ่งออกเป็นสองขั้นตอนดังนี้

3.2.1 ขั้นตอนการตัดคำ (Tokenization)

ในขั้นตอนนี้จะทำการตัดคำในแต่ละความคิดเห็นออกเป็นคำๆโดยใช้ช่องว่างระหว่างคำในความคิดเห็น เมื่อแต่ละคำถูกตัดออกจากกันแล้ว จากนั้นจะทำการลบอักขระที่ไม่ใช่ตัวอักษรออกไป เช่น ". + !" และทำการลบตัวเลข ออกจากความคิดเห็นทุกความคิดเห็น เพราะอักขระเหล่านี้ไม่มีความหมายในตัวเองจึงไม่จำเป็นต้องนำไปประมวลผล โดยแต่ละความคิดเห็นจะเหลือแต่คำที่มีความหมายเท่านั้น ในขั้นตอนนี้จะใช้ Regular Expression ในการตัดคำ และค้นหาอักขระที่ไม่ใช่ตัวอักษรในทุกความคิดเห็น Regular Expression ที่ใช้ในขั้นตอนนี้แสดงดังรูปที่ 3.3

```
W*$|d|^W*
```

รูปที่ 3.3 Regular Expression

3.2.2 ขั้นตอนการลบ (Delete Stop Words)

ในแต่ละความคิดเห็นประกอบด้วย Stop Words มากมาย เมื่อแต่ละความคิดเห็นทำการลบ Stop Words ออกไปแล้วจะไม่ทำให้ความหมายของความคิดเห็นเปลี่ยนไปเพราะ Stop Word เป็นคำที่ไม่มีความหมายในตัวเองเป็นคำที่มาขยายคำอื่นในประโยค จึงไม่จำเป็นต้องนำคำเหล่านี้ไปประมวลผล ตัวอย่าง Stop Words เช่น "and, the, of, I" เป็นต้น คำทุกคำในแต่ละความคิดเห็นจะถูกนำไปเทียบกับชุด Stop Words ถ้าพบว่าคำในความคิดเห็นเป็น Stop Words ก็จะถูกลบออกจากความคิดเห็น

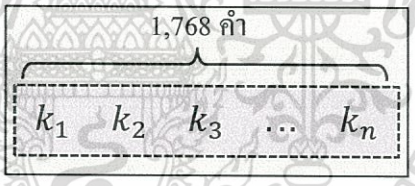
หลังจากที่ Stop Words ถูกลบออกจากทุกความคิดเห็นแล้วจะทำให้จำนวนของคำในแต่ละความคิดเห็นน้อยลงอย่างมากขนาดของชุดข้อมูลลดลงส่งผลให้ใช้เวลาในการประมวลผลข้อมูลน้อยลงด้วย

3.3 การแปลงข้อมูล (Text Transformation)

ขั้นตอนการแปลงข้อมูลเป็นขั้นตอนที่ใช้ในการเปลี่ยนความคิดเห็นที่อยู่ในรูปแบบของตัวอักษรให้อยู่ในรูปแบบของตัวเลขเพื่อสร้างชุดข้อมูลที่สามารถนำไปจำแนกประเภทได้โดยที่ข้อมูลนั้นจะต้องมีความเหมาะสมกับโครงข่ายประสาทเทียม ขั้นตอนการแปลงข้อมูลแบ่งออกเป็นสองขั้นตอนดังนี้

3.3.1 ขั้นตอนการสร้างชุดของคำสำคัญ

ในขั้นตอนนี้จะใช้คำทุกคำในความคิดเห็นของร้านอาหารที่ใช้เป็นชุดข้อมูลฝึกสอนทั้ง 510 ความคิดเห็น ทั้งความคิดเห็นเชิงบวก และ ความคิดเห็นเชิงลบที่ผ่านการลบ Stop Words แล้ว ชุดคำสำคัญเป็นชุดของคำที่มีความคำสำคัญเพื่อใช้ในการจำแนกประเภทของความคิดเห็น โดยใช้คำทั้งหมดในความคิดเห็นที่ใช้เป็นชุดข้อมูลฝึกสอนมาทำการลบคำที่ซ้ำกันออกไป ดังนั้นชุดคำสำคัญจึงเป็นชุดคำที่เป็นคำที่ไม่ซ้ำกันครอบคลุมทุกความคิดเห็นเพื่อใช้สำหรับสร้างชุดข้อมูล (Dataset)



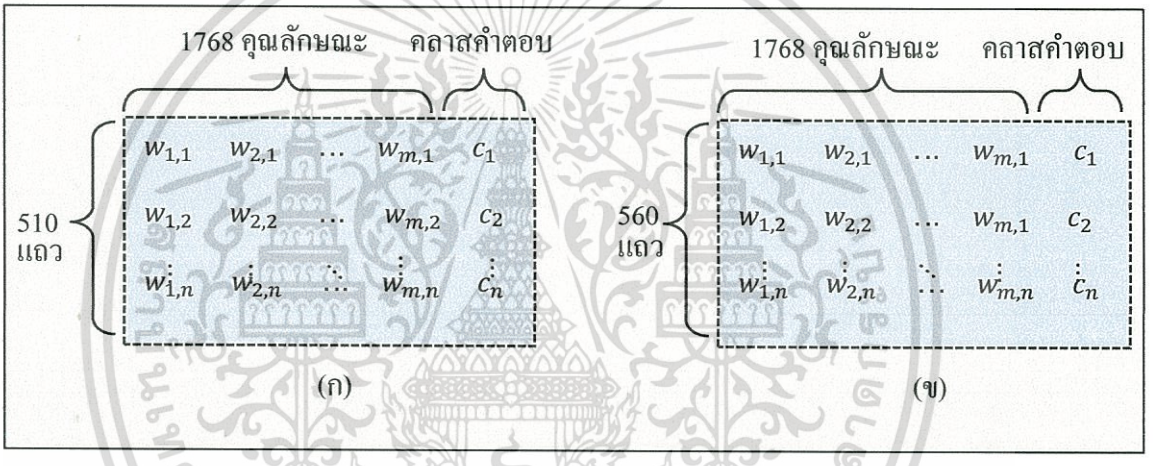
รูปที่ 3.4 ชุดของคำสำคัญ

หลังจากที่ได้ทำการสร้างชุดคำสำคัญจากคำที่ไม่ซ้ำกันแล้วทำให้ได้ชุดคำที่ไม่ซ้ำกันจำนวน 1,768 คำจากทั้ง 510 ความคิดเห็นของชุดข้อมูลฝึกสอน จากนั้นเราจะใช้คำสำคัญเป็นชุดของคำที่ใช้บ่งบอกประเภทของความคิดเห็น โดยคำสำคัญจะถูกนำไปใช้นับความถี่เพื่อสร้างชุดข้อมูลในขั้นตอนต่อไป

3.3.2. ขั้นตอนการแปลงข้อมูล

ในขั้นตอนนี้จะทำการแปลงข้อมูลจากทุกความคิดเห็นให้เป็นข้อมูลแบบตัวเลข เนื่องจากข้อมูลความคิดเห็นของลูกค้านั้นเป็นการแสดงความคิดเห็นแบบข้อความตัวอักษร ดังนั้นเพื่อให้ข้อมูลสามารถใช้ได้กับวิธีการที่ใช้ในการจำแนกประเภทความคิดเห็น ในงานวิจัยนี้ได้จึงต้องทำการแปลงข้อมูลจากข้อมูลตัวอักษรให้กลายเป็นตัวเลข โดยการนับความถี่ของแต่ละคำในชุดคำสำคัญที่มีปรากฏอยู่ในแต่ละความ

คิดเห็น วิธีการแปลงข้อมูลถูกทำโดยการนำคำในชุดของคำสำคัญมาเทียบกับคำในความคิดเห็นเพื่อหาความถี่ของคำนั้นว่าปรากฏในความคิดเห็นกี่คำ ทุกคำในชุดของคำสำคัญจะถูกนำมาเปรียบเทียบกับคำในความคิดเห็นทีละคำจนกระทั่งครบทุกคำในชุดของคำสำคัญ ความถี่ที่ได้จะถูกบันทึกลงไฟล์ข้อมูลทีละคำจนครบโดยที่ค่าสุดท้ายจะเป็นประเภทของความคิดเห็น ความถี่ของแต่ละความคิดเห็นจะอยู่ในบรรทัดเดียวกันและแต่ละความถี่จะเว้นช่องว่างหนึ่งช่อง เมื่อเริ่มทำการนับความถี่ในความคิดเห็นถัดไปก็จะทำการบันทึกลงไฟล์ในบรรทัดถัดไป วิธีการแปลงข้อมูลจะทำตามขั้นตอนนี้ไปเรื่อยๆจนครบทุกความคิดเห็นก็จะได้ชุดข้อมูลที่ใช้ในการจำแนกความคิดเห็น โดยที่ขนาดของคุณลักษณะของชุดข้อมูล (Features) จะมีขนาดเท่ากับจำนวนของคำในชุดของคำสำคัญ คือ1,768 คำ แสดงในรูปที่ 3.5



รูปที่ 3.5 ขนาดของชุดข้อมูลที่ใช้ในการทดลอง (ก) ขนาดของชุดข้อมูลฝึกสอน (ข) ขนาดของชุดข้อมูลทดสอบ

หลังจากที่ได้ทำการแปลงข้อมูลความคิดเห็นแล้วจะทำการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลฝึกสอน และ ชุดข้อมูลทดสอบ โดยแต่ละชุดข้อมูลมีขนาดของชุดข้อมูลดังนี้ ชุดข้อมูลฝึกสอนประกอบด้วย 1,769 คุณลักษณะ 510 อินพุตเวกเตอร์ และ ชุดข้อมูลทดสอบประกอบด้วย 1,769 คุณลักษณะ 550 อินพุตเวกเตอร์ โดยที่แต่ละชุดข้อมูลเราให้คุณลักษณะสุดท้ายเป็นคลาสคำตอบของแต่ละความคิดเห็น ความคิดเห็นในเชิงบวกแทนด้วยเลข 1 และความคิดเห็นในเชิงลบแทนด้วยเลข 0 เพื่อเป็นคำตอบของแต่ละชุดข้อมูล ข้อมูลที่ผ่านการแปลงจะถูกเขียนลงไฟล์ข้อมูลโดยให้แต่ละบรรทัดแทนข้อมูลแต่ละความคิดเห็น

1	1	1	1	1	2	1	1	2	1	1	1	1
0	0	0	1	0	0	0	0	0	0	0	2	0
0	0	3	2	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	2	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	1	0
0	0	0	5	0	0	0	0	0	0	0	1	0
0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	1	4	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	2	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	0	0	0	0	0	0	0
0	0	2	1	0	0	0	0	0	0	1	0	0
0	0	2	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0
0	0	0	4	0	0	0	0	0	0	0	1	0
0	0	1	1	0	0	0	0	0	0	0	0	0
0	0	1	2	0	0	0	0	0	0	0	0	0

รูปที่ 3.6 ตัวอย่างชุดข้อมูลที่ผ่านการแปลงข้อมูลแล้ว

3.4 การเลือกคุณลักษณะโดยใช้วิธีเอ็มอาร์เอ็มอาร์

วิธีการเลือกคุณสมบัติแบบเอ็มอาร์เอ็มอาร์ (mRMR Feature Selection) เป็นวิธีการที่ใช้ในการเลือกคุณลักษณะจากชุดข้อมูลโดยเลือกคุณลักษณะที่มีค่าความสัมพันธ์กันระหว่างคุณลักษณะกับคลาสคำตอบที่มีค่าสูงและมีความซ้ำซ้อนกันระหว่างคุณลักษณะที่ต่ำโดยใช้ค่า Mutual Information (MI) ในงานวิจัยนี้เราใช้วิธีการเลือกคุณลักษณะแบบ mRMR เพื่อหาชุดคำที่เป็นคำที่เหมาะสมในชุดของคำสำคัญจากการเลือกคุณลักษณะโดยใช้ชุดข้อมูลฝึกสอน วัตถุประสงค์ในการเลือกคุณลักษณะของงานวิจัยนี้มีดังนี้

1. ลดขนาดของข้อมูล
2. ลดเวลาในการเรียนรู้ของโครงข่ายประสาทเทียม
3. เพิ่มความถูกต้องในการจำแนกประเภทความคิดเห็น
4. ลดข้อมูลรบกวน
5. สามารถเลือกชุดคำสำคัญที่เหมาะสมที่บ่งบอกถึงประเภทของความคิดเห็นได้ดี
6. ช่วยลดภาระของผู้ใช้ในการเลือกคำสำคัญ

หลังจากที่ได้ทำการสร้างชุดข้อมูลของแต่ละความคิดเห็นผ่านขั้นตอนวิธีการเตรียมข้อมูล (Text Preprocessing) และ ขั้นตอนการแปลงข้อมูล (Text Transformations) ชุดข้อมูลที่ได้ทำการสร้างจากสองขั้นตอนนี้จะมีขนาดของคุณลักษณะเท่ากับ 1,768 คุณลักษณะซึ่งเท่ากับจำนวนของคำสำคัญทั้งหมด ทำให้ชุดข้อมูลมีขนาดใหญ่โครงข่ายประสาทเทียมใช้เวลานานในการเรียนรู้ประเภทของความคิดเห็น ดังนั้นวิธีการเลือกคุณลักษณะแบบ mRMR จะช่วยลดขนาดของข้อมูลลงโดยการเลือกคุณลักษณะที่มีความสัมพันธ์กับชุดคำตอบ เนื่องจากคำในชุดคำสำคัญนั้นเราใช้ทุกคำที่มีอยู่ในความคิดเห็นของชุดข้อมูลฝึกสอน ข้อมูลจึงประกอบไปด้วยคุณลักษณะที่ไม่เกี่ยวข้องกับความคิดเห็นหรือข้อมูลรบกวน (Noisy Data) อยู่เป็นจำนวนมาก การเลือกคุณลักษณะแบบ mRMR จะช่วยเลือกเฉพาะคุณลักษณะที่จำเป็นในการจำแนกประเภทความคิดเห็นออกมา

ขั้นตอนนี้เราใช้วิธี mRMR เพื่อเลือกคำสำคัญ ซึ่งเป็นชุดคำที่มีความสำคัญสามารถบอกประเภทของความคิดเห็น โดยการเลือกคำสำคัญนั้นจะใช้ชุดข้อมูลฝึกสอนเป็นข้อมูลที่ใช้ในการเลือกชุดของคุณลักษณะ โดยชุดของคุณลักษณะที่ผ่านการคัดเลือกก็คือชุดของคำสำคัญนั่นเอง ในขั้นตอนนี้กำหนดให้ mRMR เลือกชุดของคุณลักษณะจากชุดข้อมูลออกมาจำนวน 6 ชุดเพื่อให้ได้ชุดคำสำคัญๆที่ครอบคลุมทุกคำ แต่ละชุดคุณลักษณะที่เลือกออกมานั้นจะมีจำนวนคุณลักษณะดังต่อไปนี้ 50, 100, 150, 200, 250, 300 และ 400 คุณลักษณะ ซึ่งคุณลักษณะที่เลือกโดยวิธี mRMR คือ ชุดของคำสำคัญที่มีเฉพาะคำที่มีความสำคัญ และ ตัดคำที่ไม่จำเป็นออกไป โดยที่ผู้ใช้งานไม่จำเป็นต้องเลือกเองเพราะคำในชุดของคำสำคัญมีจำนวนมากและผู้ใช้ไม่สามารถทราบได้ว่าคำไหนจะเป็นคำที่สำคัญที่ใช้บอกประเภทความคิดเห็นได้ วิธี mRMR จึงถูกนำมาช่วยในการเลือกคำที่มีความสำคัญ การเลือกคุณลักษณะแบบ mRMR จึงเป็นวิธีที่ทำให้สามารถเลือกชุดคำได้อย่างถูกต้องตรงตามความต้องการ และยังใช้เวลาในการเลือกชุดคำน้อย ตัวอย่างชุดของคำสำคัญ ที่ผ่านการเลือกโดยวิธี mRMR แสดงดังรูปที่ 3.7

great, excellent, delicious, worst, disappointing, amazing, overpriced, average, bad, bar,

รูปที่ 3.7 ตัวอย่างชุดคำสำคัญที่เลือกโดยวิธีการ mRMR

จากตัวอย่างจะเห็นว่าวิธีการเลือกคุณลักษณะแบบ mRMR สามารถเลือกคำจากชุดคำสำคัญที่สามารถบอกถึงประเภทของความคิดเห็นได้เป็นอย่างดี แต่ละคำที่ผ่านการเลือกแล้วล้วนแต่เป็นคำที่มีทั้งคำทางด้านดีและด้านไม่ดีที่อยู่ในความคิดเห็นรวมอยู่ในชุดคำที่ผ่านการเลือกออกมา โดยที่แต่ละชุด คำที่

ผ่านการเลือกออกมาจะถูกนำไปใช้กับชุดข้อมูลทั้งขั้นตอนการฝึกสอนและขั้นตอนการทดสอบโครงข่ายประสาทเทียม ชุดข้อมูลทดสอบใช้ชุดของคุณลักษณะที่เลือกโดยวิธีการ mRMR ซึ่งเป็นชุดเดียวกันกับชุดข้อมูลฝึกสอน การสร้างชุดข้อมูลใหม่ที่ผ่านการเลือกคุณลักษณะแล้วจะทำให้คุณลักษณะที่ไม่เกี่ยวข้องถูกตัดออกไปเหลือแต่คุณลักษณะที่สำคัญทำให้ชุดข้อมูลมีขนาดเล็กลงช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทความคิดเห็นของโครงข่ายประสาทเทียมได้ดียิ่งขึ้น

3.5 การจำแนกประเภทความคิดเห็นโดยใช้โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น

ในขั้นตอนการจำแนกประเภทความคิดเห็นโดยใช้โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นดำเนินการทดลองโดยใช้โปรแกรม Matlab การจำแนกประเภทความคิดเห็นมีขั้นตอนดังนี้

3.5.1 การฝึกสอนโครงข่ายประสาทเทียม (Training)

การจำแนกประเภทข้อมูลความคิดเห็นร้านอาหารไทยในขั้นตอนนี้เริ่มจากการโหลดข้อมูลฝึกสอนที่ได้สร้างไว้ในรูปของ Text File เข้ามาเก็บไว้ในรูปแบบเมทริกซ์ข้อมูล จากนั้นทำการแยกคลาสคำตอบออกจากชุดข้อมูล โครงข่ายประสาทเทียมจะทำการเรียนรู้ข้อมูลความคิดเห็นจากชุดข้อมูลฝึกสอนเพื่อให้ได้ความถูกต้องในการจำแนกประเภทความคิดเห็นมากที่สุด การทดลองจะเลือกจำนวนโหนดโดยการปรับจำนวนของโหนดในชั้นซ่อนตั้งแต่ 2 โหนด ถึง 16 โหนดแล้วเลือกจำนวนโหนดที่มีความถูกต้องในการจำแนกประเภทความคิดเห็นมากที่สุด หลังจากการทดลองโครงข่ายประสาทเทียมที่มีโหนดในชั้นซ่อน 16 โหนดให้ความถูกต้องในการจำแนกประเภทความคิดเห็นมากที่สุด ในขั้นตอนการเรียนรู้เราแบ่งข้อมูลฝึกสอนออกเป็นสองส่วน ส่วนที่หนึ่งเป็นข้อมูลที่ใช้สำหรับสอนโครงข่ายประสาทเทียมกำหนดให้เท่ากับ 90% ของข้อมูลฝึกสอน ส่วนที่สองใช้สำหรับทดสอบโครงข่ายประสาทเทียมที่ได้เรียนรู้ประเภทของความคิดเห็นเพื่อป้องกันการเกิด Over Fitting ของการเรียนรู้โดยข้อมูลส่วนนี้กำหนดให้เท่ากับ 10% ของข้อมูลฝึกสอน โครงข่ายประสาทเทียมจะทำการเรียนรู้ประเภทของความคิดเห็นทุกชุดข้อมูลฝึกสอนทั้งที่ผ่านการเลือกคุณลักษณะ และชุดข้อมูลที่ไม่ผ่านการเลือกคุณลักษณะรวมทั้งหมด 9 ชุดข้อมูล การเรียนรู้แต่ละชุดข้อมูลจะถูกจับเวลาการเรียนรู้เอาไว้เพื่อใช้ในการเปรียบเทียบว่าจำนวนของคุณลักษณะมีผลทำให้เวลาในการเรียนรู้ของโครงข่ายประสาทเทียมลดลงหรือไม่ ในขั้นตอนการเรียนรู้โครงข่ายประสาทเทียมจะทำการปรับค่าน้ำหนักของทุกเส้นเชื่อมโยงจนกว่าโครงข่ายประสาทเทียมจะทำการจำแนกประเภทความคิดเห็นของชุดข้อมูลฝึกสอนได้อย่างถูกต้อง โครงข่ายประสาทเทียมที่ผ่านการเรียนรู้จะถูกนำไปใช้ในการจำแนกประเภทความคิดเห็นของร้านอาหารไทยในขั้นตอนการทดสอบ

3.5.2 การทดสอบประสิทธิภาพของโครงข่ายประสาทเทียม (Testing)

ในขั้นตอนนี้จะทำการจำแนกประเภทความคิดโดยใช้โครงข่ายประสาทเทียมที่ผ่านการเรียนรู้ข้อมูลความคิดเห็นของร้านอาหารไทยมาทำการจำแนกประเภทความคิดเห็นโดยใช้ชุดข้อมูลทดสอบ โครงข่ายประสาทเทียมจะทำการจำแนกประเภทความคิดเห็นของร้านอาหารโดย นำคำตอบที่ได้จากโครงข่ายประสาทเทียมมาเปรียบเทียบกับคลาสคำตอบของชุดข้อมูลทดสอบแล้วนำมาสร้าง Confusion Matrix เพื่อวัดประสิทธิภาพในการจำแนกประเภทข้อมูลของโครงข่ายประสาทเทียม [3] โดยที่ Confusion Matrix มีรายละเอียดดังรูป 3.8

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

รูปที่ 3.8 Confusion Matrix

ใน Confusion Matrix ประกอบด้วย

True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Positive

True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Negative

False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส Positive

False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส Negative

ในการทำการทดลองการจำแนกประเภทความคิดเห็นนี้จะใช้การวัดประสิทธิภาพในการจำแนกประเภทของโครงข่ายประสาทเทียมด้วยกันสามค่าคือ ค่าความความถูกต้อง (Accuracy) ค่าความแม่นยำของการทำนายคลาสบวก True Positive Rate (TPR) และ ค่าความแม่นยำของการทำนายคลาสลบ True Negative Rate (TNR) โดยมีรายละเอียดดังนี้

Accuracy เป็นค่าความถูกต้องทั้งหมดจากการจำแนกประเภทซึ่งคิดเป็นเปอร์เซ็น ค่า Accuracy เป็นค่าหนึ่งที่นิยมใช้วัดความแม่นยำในการจำแนกประเภทดังสมการที่ 3.1

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (3.1)$$

True Positive Rate (TPR) หรือ Sensitivity คือค่าที่ใช้วัดความเที่ยงตรงของการทำนายคลาสบวกจากจำนวนของคลาสบวกทั้งหมดแสดงดังสมการที่ 3.2

$$\text{TPR} = \frac{TP}{TP+FN} \quad (3.2)$$

True Negative Rate (TNR) หรือ Specificity คือค่าที่ใช้วัดความสามารถในการทำนายคลาสลบจากจำนวนของคลาสลบทั้งหมดแสดงดังสมการที่ 3.3

$$TNP = \frac{TN}{TN+FP} \quad (3.3)$$

ในการจำแนกข้อมูล (Classification) ค่า Accuracy TPR และ TNR เป็นค่าที่นิยมใช้ในการวัดประสิทธิภาพในการจำแนกประเภทข้อมูลของเครื่องมือที่ใช้ในการจำแนกประเภทข้อมูล

บทที่ 4

ผลการทดลอง

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองนี้เป็นข้อมูลความคิดเห็นของร้านอาหารไทยโดยเก็บรวบรวมมาจากความคิดเห็นบนเว็บไซต์ www.tripadvisor.com ความคิดเห็นที่เก็บรวบรวมมานั้นมีจำนวณ 1,060 ความคิดเห็นซึ่งประกอบด้วยความคิดเห็นเชิงบวก และความคิดเห็นเชิงลบ โดยที่ความคิดเห็นเชิงบวกคือความคิดเห็นที่มีคะแนนมากกว่าหรือเท่ากับ 4 ถึง 5 ดาว และความคิดเห็นเชิงลบคือความคิดเห็นที่มีคะแนน 1 ถึง 2 ดาว ความคิดเห็นทั้งสองประเภททำการสุ่มเลือกจากเว็บไซต์ทั้งหมด 1,060 ความคิดเห็น

ก่อนที่จะทำการจำแนกประเภทความคิดเห็นนั้น ต้องมีการเตรียมข้อมูลความคิดเห็นเพื่อให้ข้อมูลเหมาะสมกับการเรียนรู้ของโครงข่ายประสาทเทียมที่นำมาใช้ในการจำแนกประเภทความคิดเห็น โดยที่ทุกความคิดจะต้องผ่านขั้นตอนในการเตรียมข้อมูลดังนี้

4.1.1 การเตรียมข้อมูล (Text Preprocessing Method)

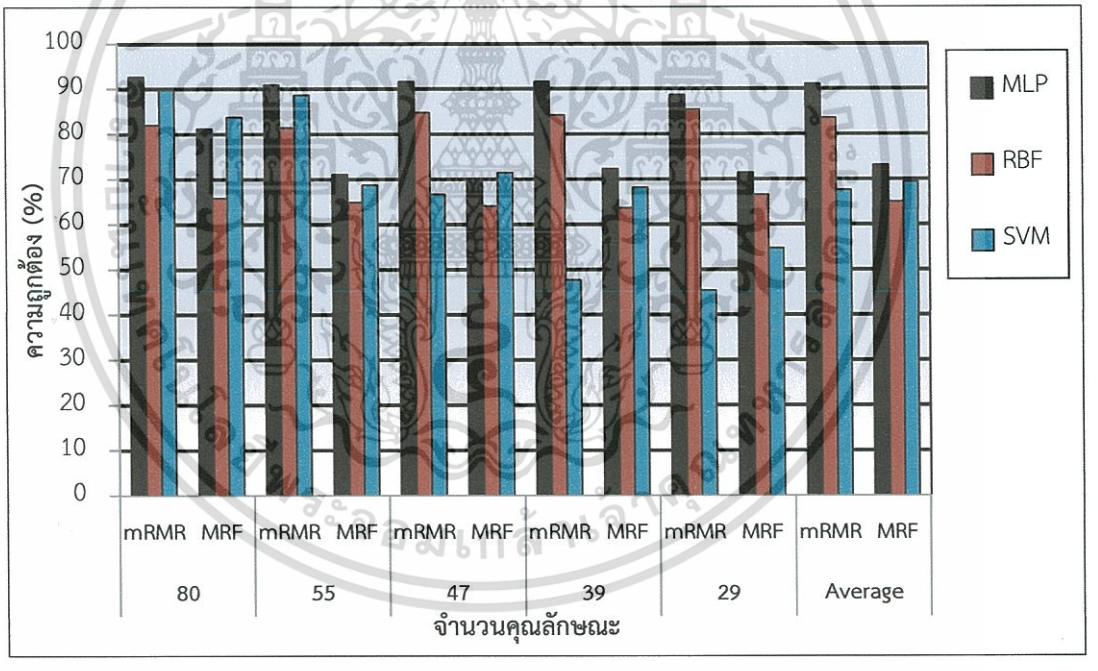
ข้อมูลแต่ละความคิดเห็นจากลูกค้าที่เก็บรวบรวมมานั้นต้องผ่านการลบอักขระที่ไม่ใช่ตัวอักษร ตัวเลข และ Stop Words เพื่อลดข้อมูลรบกวน และลดปริมาณคำที่ใช้ในการประมวลผลโดยการตัดคำที่ไม่จำเป็นเหล่านี้ออกไปจากความคิดเห็น

4.1.2 การแปลงข้อมูล (Text Transformation Method)

หลังจากทุกความคิดเห็นผ่านขั้นตอนการเตรียมข้อมูลเพื่อสร้างชุดข้อมูลแล้ว จากนั้นชุดข้อมูลนี้จะถูกแปลงจากตัวอักษรมาเป็นข้อมูลแบบตัวเลขโดยการนับความถี่ของคำใน Key Words ที่ปรากฏในแต่ละความคิดเห็น โดยชุดข้อมูลแบ่งออกเป็นสองชุดคือ ชุดข้อมูลฝึกสอน และ ชุดข้อมูลทดสอบ ชุดข้อมูลฝึกสอนประกอบด้วยข้อมูลจำนวน 510 แถว 1,769 คอลัมน์ และชุดข้อมูลทดสอบประกอบด้วยข้อมูลจำนวน 550 แถว 1,769 คอลัมน์ โดยที่คอลัมน์สุดท้ายของทั้งสองชุดเป็นคลาสคำตอบดังนี้ ตัวเลขเลข 1 แทนความคิดเห็นเชิงบวก และ ตัวเลข 0 แทนความคิดเห็นเชิงลบ

4.2 การเลือกคุณลักษณะด้วยวิธีเอ็มอาร์เอ็มอาร์

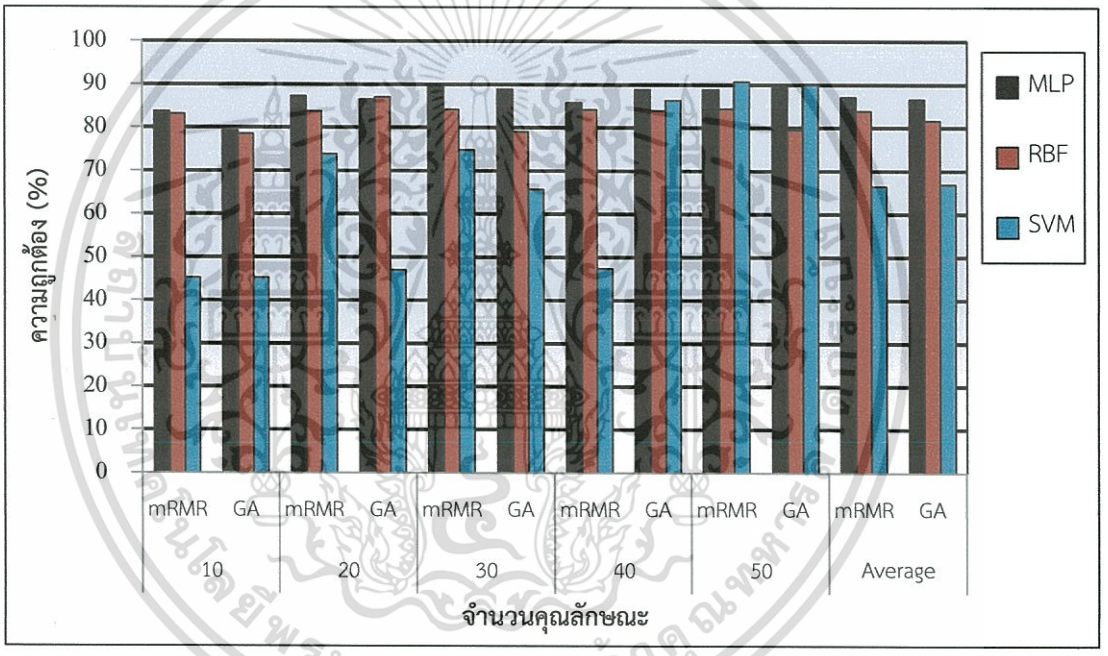
ในขั้นตอนนี้วิธีการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ (mRMR Feature Selection) ถูกนำมาใช้ในการเลือกชุดคำสำคัญที่เหมาะสมในการจำแนกประเภทความคิดเห็น จากชุดข้อมูลที่สร้างขึ้นมาในขั้นตอนการแปลงข้อมูลนั้นชุดข้อมูลประกอบด้วยคุณลักษณะจำนวณ 1,768 คุณลักษณะซึ่งจะเห็นว่าประกอบด้วยคำเป็นจำนวนมากซึ่งอาจรวมถึงคำที่ไม่จำเป็นต่อการประมวลผลรวมอยู่ด้วยจำนวนมาก ดังนั้น วิธีการเลือกคุณลักษณะแบบ mRMR จึงเป็นวิธีที่ใช้เลือกชุดของคำที่จำเป็นออกมา เนื่องจากต้องการตรวจสอบว่า การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์มีความเหมาะสมต่อการจำแนกประเภทความคิดเห็นมากที่สุดหรือไม่จึงได้ทำการเปรียบเทียบประสิทธิภาพของวิธีการเลือกคุณลักษณะแบบ mRMR กับ วิธีการเลือกคุณลักษณะอื่นๆ อีกสามวิธี คือ วิธีการเลือกคุณลักษณะแบบ MRF, GA และ Binary GA



รูปที่ 4.1 แผนภูมิเปรียบเทียบค่าความถูกต้องในการจำแนกประเภทความคิดเห็นจากการเลือกคุณลักษณะแบบ mRMR และ MRF

รูปที่ 4.1 เป็นการเปรียบเทียบประสิทธิภาพในการเลือกคุณลักษณะด้วยวิธีการ mRMR และ MRF [10] โดยกำหนดให้ทั้งสองวิธีเลือกจำนวนคุณลักษณะจากชุดข้อมูลฝึกสอนจำนวนห้าชุดดังนี้ 80,

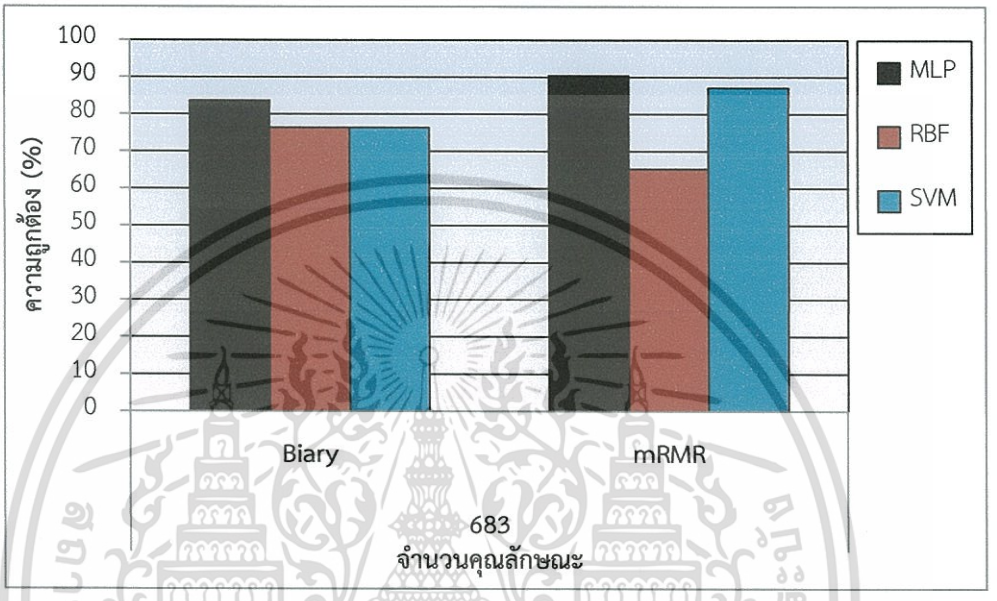
55, 47, 39, 29 คุณลักษณะ แต่ละชุดข้อมูลจะถูกนำไปใช้ในการจำแนกประเภทโดย MLP, RBF และ SVM จากแผนภูมิแสดงให้เห็นว่าชุดข้อมูลที่ผ่านการเลือกคุณลักษณะด้วยวิธี mRMR มีค่าความถูกต้องในการจำแนกประเภทความคิดเห็นด้วยวิธี MLP และ RBF มากกว่าการเลือกคุณลักษณะแบบ MRF ทุกชุดข้อมูล ค่าความถูกต้องของวิธีการจำแนกประเภทแบบ SVM การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ให้ค่าความถูกต้องมากกว่าการเลือกคุณลักษณะแบบ MRF ในชุดข้อมูลขนาด 80 และ 55 คุณลักษณะ แต่กลับให้ความถูกต้องที่น้อยกว่าการเลือกคุณลักษณะแบบ MRF ในชุดข้มูลขนาด 47, 39 และ 29 คุณลักษณะ ซึ่งแสดงให้เห็นว่า การเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์จะให้ประสิทธิภาพที่ดีกว่าเมื่อมีจำนวนคุณลักษณะที่มากพอ



รูปที่ 4.2 แผนภูมิเปรียบเทียบความถูกต้องในการจำแนกประเภทความคิดเห็นจากการเลือกคุณลักษณะแบบ mRMR และ GA

รูปที่ 4.2 เป็นการเปรียบเทียบประสิทธิภาพในการเลือกคุณลักษณะโดยวิธี mRMR และ GA [11] จากแผนภูมิวิธีการจำแนกประเภทแบบ MLP, RBF และ SVM มีค่าความถูกต้องโดยเฉลี่ยในการจำแนกประเภทความคิดเห็นที่ใกล้เคียงกับทั้งวิธีการเลือกคุณลักษณะแบบ mRMR และ GA แสดงให้เห็นว่าวิธีการเลือกคุณลักษณะแบบ mRMR และ GA มีประสิทธิภาพในการเลือกคุณลักษณะไม่แตกต่างกันมากนัก แต่วิธีการเลือกคุณลักษณะแบบ GA นั้นใช้เวลาในการเลือกคุณลักษณะมากกว่าการเลือกคุณลักษณะแบบ mRMR มาก โดยที่วิธีการเลือกคุณลักษณะแบบ GA นั้นใช้เวลาในการเลือกชุด

คุณลักษณะแต่ละชุด ประมาณ 30 นาที ในขณะที่วิธีการเลือกคุณลักษณะแบบ mRMR นั้นใช้เวลาในการเลือกคุณลักษณะแต่ละชุดประมาณ 0.1 วินาที แสดงให้เห็นว่าวิธีการเลือกคุณลักษณะแบบ mRMR นั้นให้ประสิทธิภาพที่ดีกว่า เนื่องจากให้ค่าความถูกต้องที่ดีกว่า GA เล็กน้อย อีกทั้งยังใช้เวลาในการเลือกที่น้อยกว่ามาก



รูปที่ 4.3 แผนภูมิเปรียบเทียบความถูกต้องในการจำแนกประเภทความคิดเห็นจากการเลือกคุณลักษณะแบบ mRMR และ Binary GA

จากรูปที่ 4.3 นั้นวิธีการเลือกคุณลักษณะแบบ mRMR นั้นมีค่าความถูกต้อง ในการจำแนกประเภทความคิดเห็นโดยวิธีการจำแนกแบบ MLP และ SVM มากกว่าการเลือกคุณลักษณะแบบ Binary GA [12] แต่วิธีการเลือกคุณลักษณะแบบ Binary GA ใช้เวลาในการเลือกชุดของคุณลักษณะมากกว่าการเลือกคุณลักษณะแบบ mRMR ประมาณ 15 นาที ดังนั้นวิธีการเลือกคุณลักษณะแบบ mRMR ให้ประสิทธิภาพในการเลือกคุณลักษณะมากกว่าเพราะมีความถูกต้องมากกว่าและใช้เวลาในการเลือกน้อยกว่ามาก

จากการเปรียบเทียบประสิทธิภาพการเลือกคุณลักษณะของทั้งสามวิธีนั้นแสดงให้เห็นว่าวิธีการเลือกคุณลักษณะแบบ mRMR นั้นมีประสิทธิภาพในการเลือกชุดของคุณลักษณะมากที่สุดทั้งในด้านความถูกต้องในการจำแนกที่สูงกว่าวิธีอื่นๆ และในด้านการใช้เวลาในการเลือกคุณลักษณะ ดังนั้นวิธีการเลือกคุณลักษณะแบบ mRMR จึงเป็นวิธีการที่เหมาะสมที่สุดในการเลือกชุดคำสำคัญจากความคิดเห็นที่ใช้ในการทดลองนี้

4.3 โครงสร้างของโครงข่ายประสาทเทียม

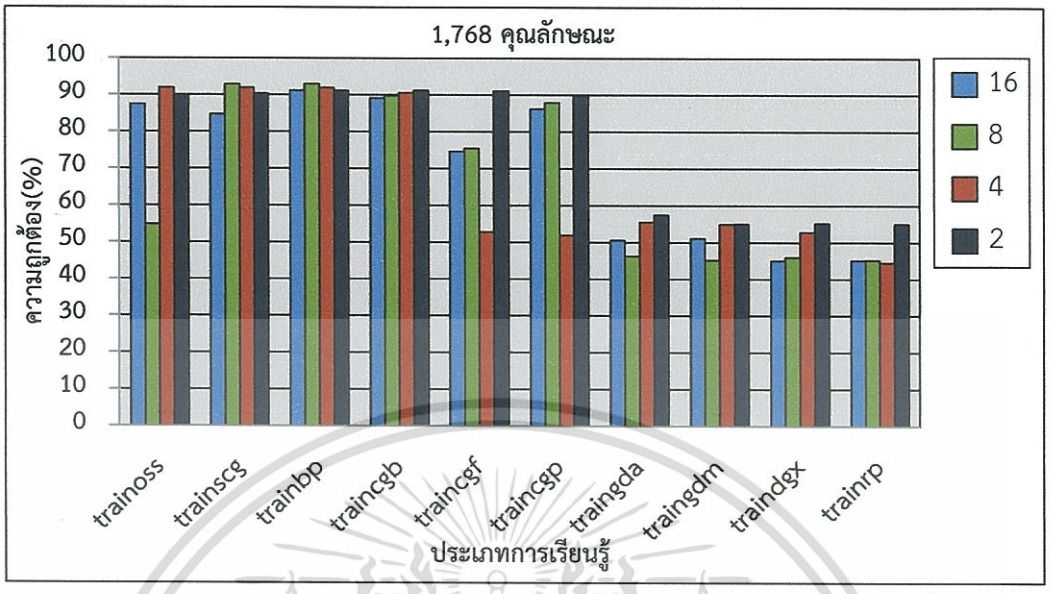
ในการทดลองนี้ใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับที่เป็นที่นิยมในการแก้ปัญหาการจำแนกประเภทข้อมูล โครงข่ายประสาทเทียมประกอบด้วยโหนดในชั้นซ่อนจำนวน 16 โหนด

4.3.1 การกำหนดค่าพารามิเตอร์ให้กับโครงข่ายประสาทเทียม

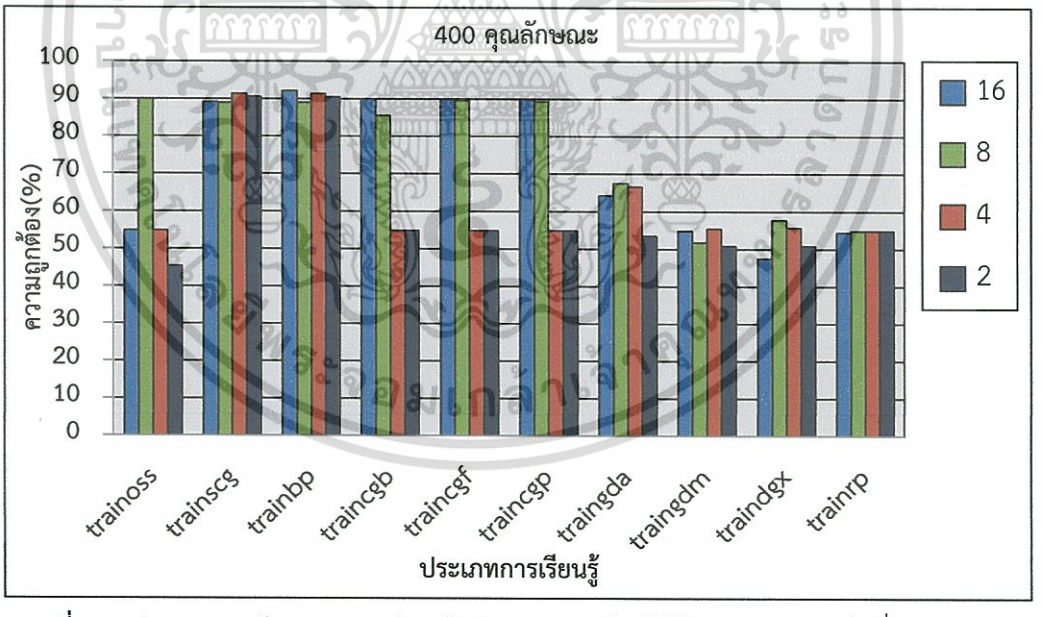
การกำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับโครงข่ายประสาทเทียมเป็นสิ่งสำคัญเพื่อให้ได้ค่าความถูกต้องในการจำแนกประเภทความคิดเห็นมากที่สุด ดังนั้นประเภทของการเรียนรู้ของโครงข่ายประสาทเทียมจึงเป็นสิ่งสำคัญที่จะช่วยให้โครงข่ายประสาทเทียมสามารถเรียนรู้ประเภทของความคิดเห็นได้อย่างถูกต้อง เนื่องจากประเภทของการเรียนรู้ของโครงข่ายประสาทเทียมนั้นมีหลายประเภทการทดลองนี้จึงทำการจำแนกประเภทความคิดเห็นของร้านอาหารไทยโดยใช้การเรียนรู้แบบต่างๆเพื่อหาประเภทการเรียนรู้ที่ให้ค่าความถูกต้องมากที่สุด ประเภทของการเรียนรู้ของโครงข่ายประสาทเทียมที่ใช้ทดสอบมีดังนี้

- One-Step Secant Backpropagation (oss) [13]
- Scaled Conjugate Gradient Backpropagation (scg) [14]
- Backpropagation (bp) [5]
- Conjugate Gradient Backpropagation with Powell-Beale restarts (cgb) [15]
- Conjugate Gradient Bachpropagation with Fletcher-Reeves updates (cgf) [16]
- Conjugate Gradient Bachpropagation with Polak-Ribiere updates (cgp) [16]
- Gradient Descent with Adaptive Learning rate Bakpropagation (gda) [17]
- Gradient Descent with Momentum Bakpropagation (gdm) [17]
- Gradient Descent with Momentum and Adaptive Learning rate Bakpropagation (gdx) [17]
- Resilient Backpropagation (rp) [18]

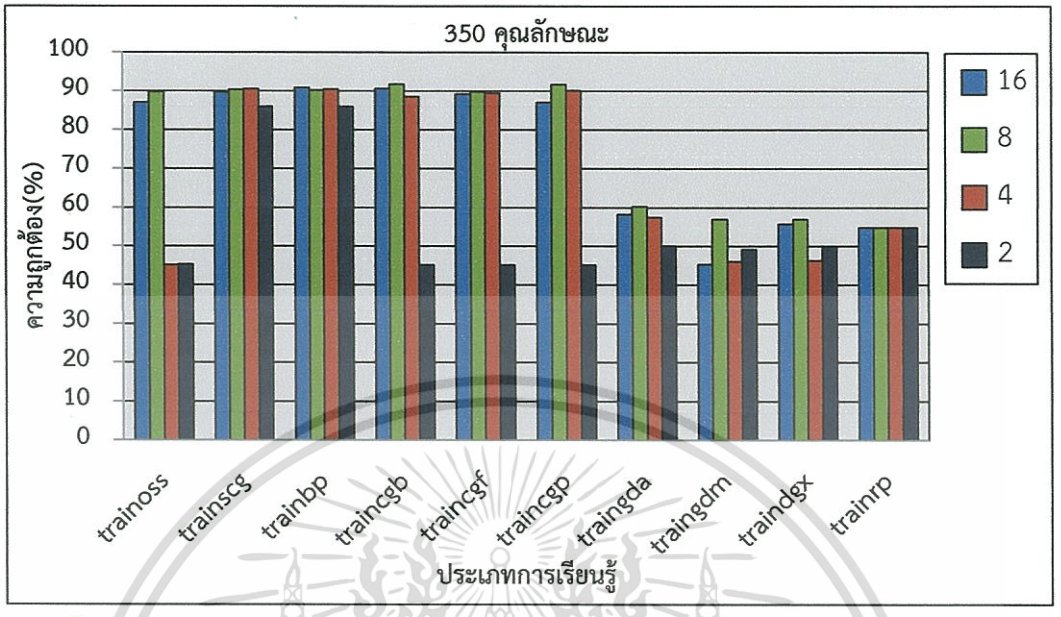
ค่าความถูกต้องของทั้ง 10 ประเภทการเรียนรู้ในแต่ละชุดของคุณลักษณะแสดงในรูปที่ 4.4 ถึง รูปที่ 4.10



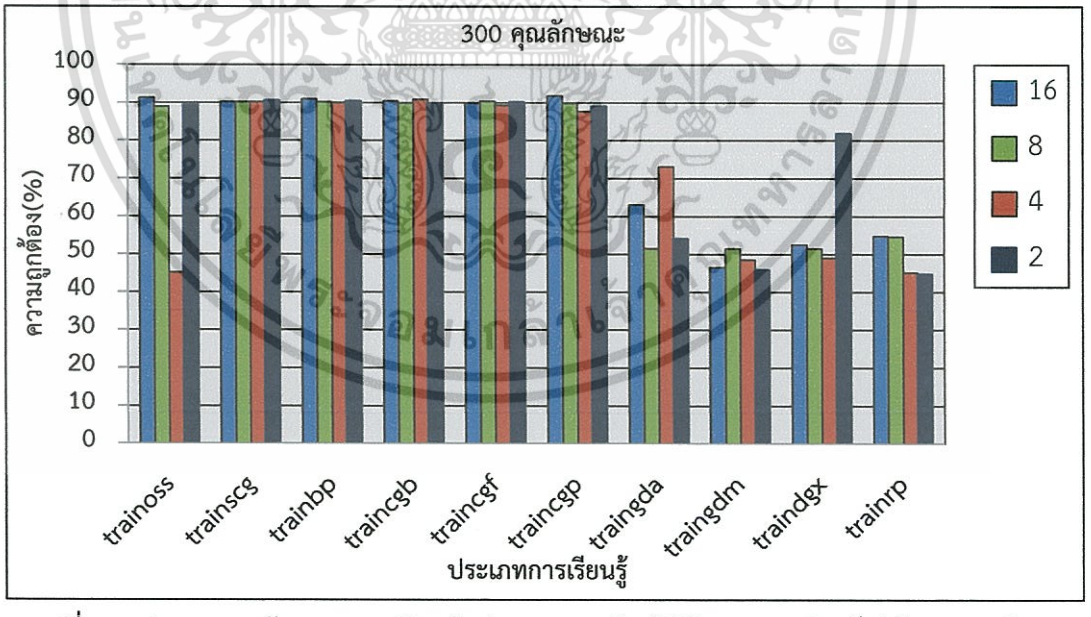
รูปที่ 4.4 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ชุดข้อมูลความคิดเห็นซึ่งมีจวน 1,768 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



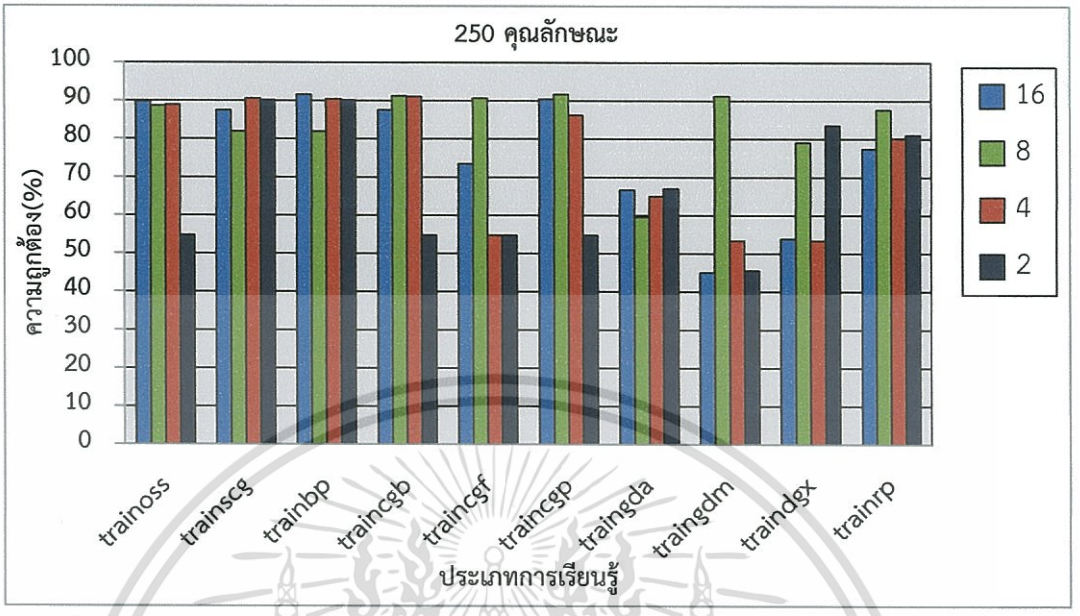
รูปที่ 4.5 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นที่ได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 400 คุณลักษณะที่ และใช้โหนดข้อมูลในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 ตามลำดับ



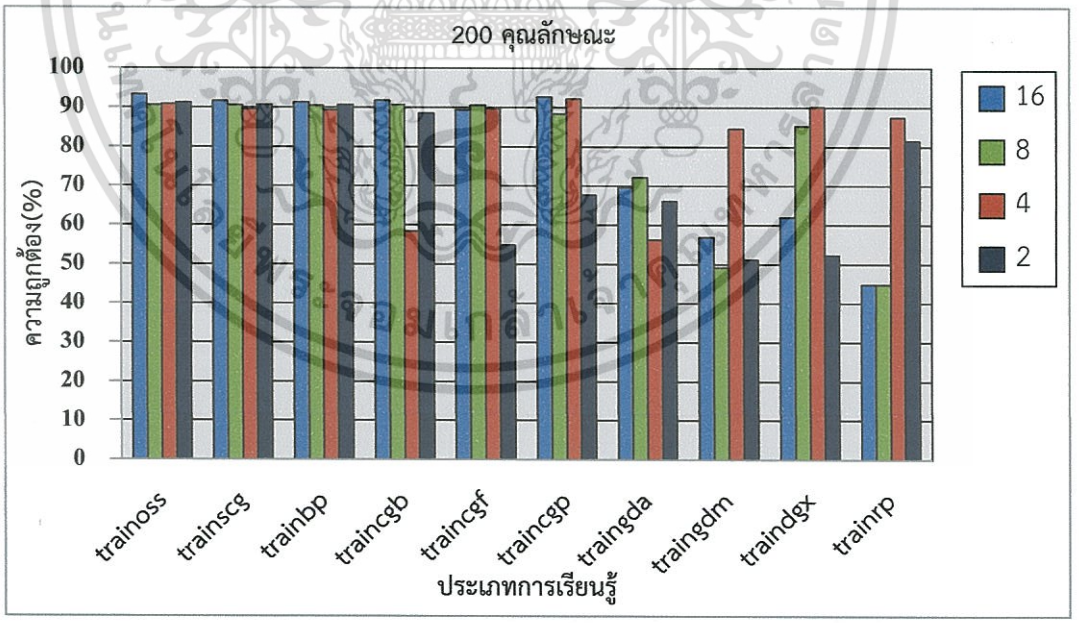
รูปที่ 4.6 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 350 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



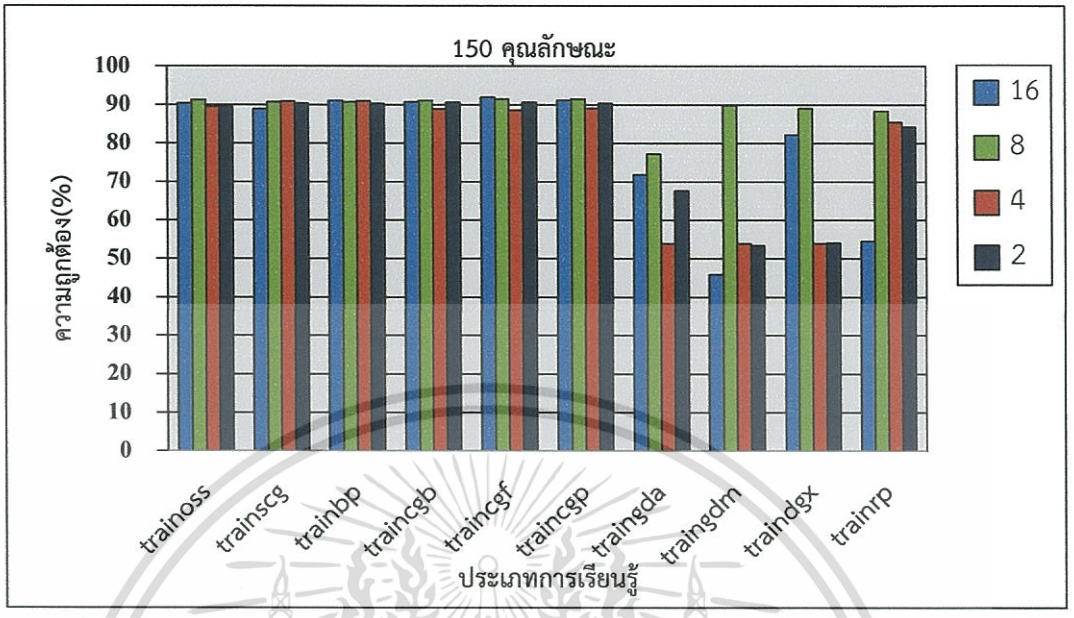
รูปที่ 4.7 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 300 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



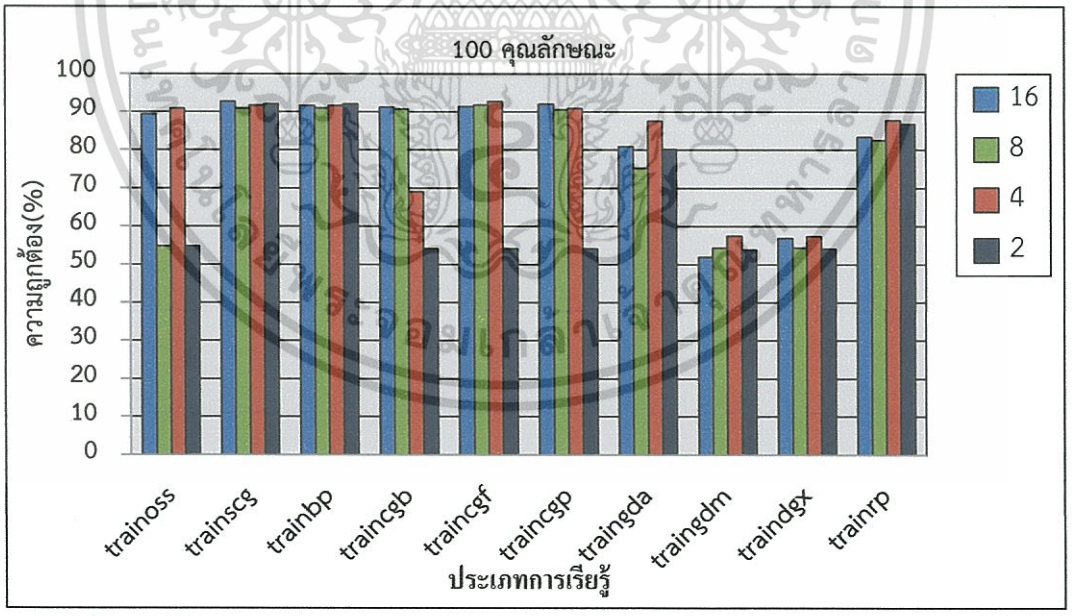
รูปที่ 4.8 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 250 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



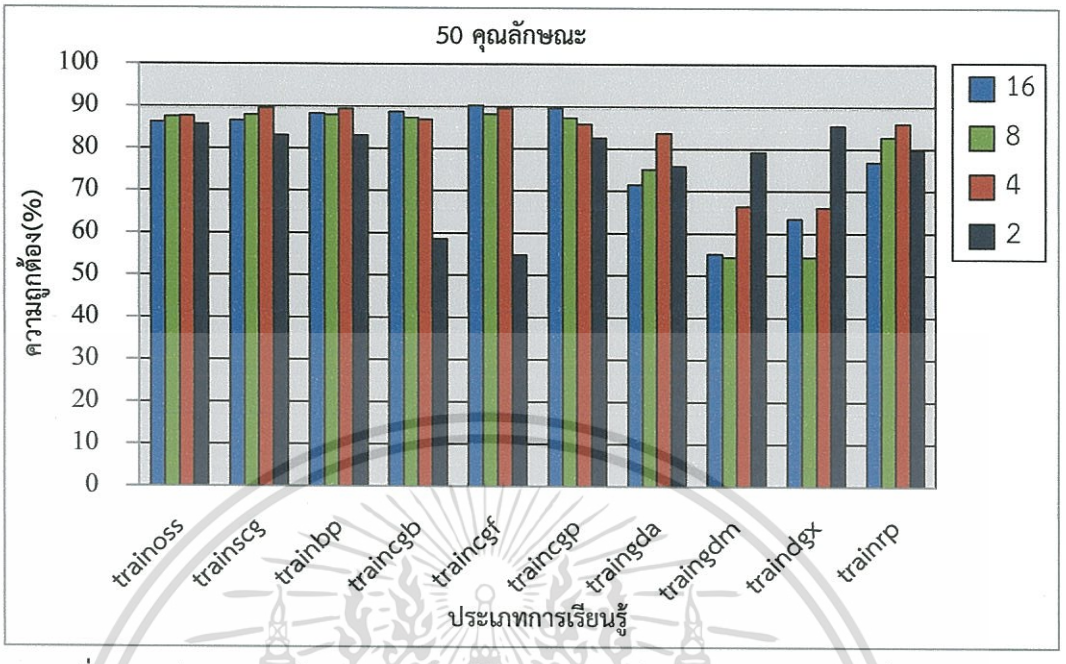
รูปที่ 4.9 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 200 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



รูปที่ 4.10 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 150 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



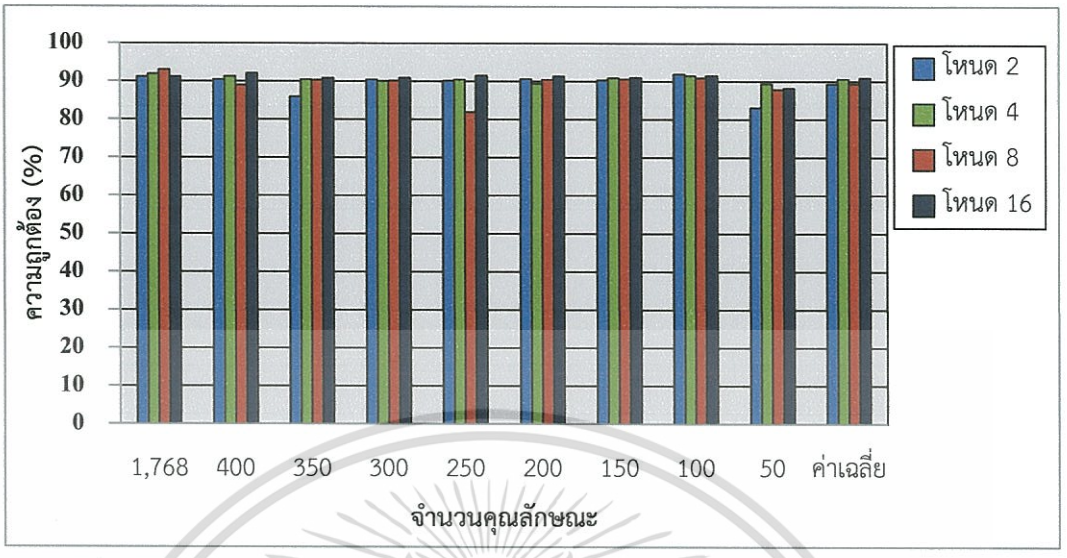
รูปที่ 4.11 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 100 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ



รูปที่ 4.12 ค่าความถูกต้องของการเรียนรู้แต่ละประเภทโดยใช้ข้อมูลความคิดเห็นได้จากการเลือกคุณลักษณะแบบเอ็มอาร์เอ็มอาร์ซึ่งมีจำนวน 50 คุณลักษณะ และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ

จากผลการทดลองของทุกประเภทการเรียนรู้ของโครงข่ายประสาทเทียมในการจำแนกประเภทความคิดเห็นแสดงให้เห็นว่าการเรียนรู้แบบแพร่กระจายย้อนกลับ (Back Propagation) เป็นการเรียนรู้ที่มีประสิทธิภาพมากที่สุด การเรียนรู้แบบแพร่กระจายย้อนกลับมีค่าความถูกต้องโดยเฉลี่ยในการจำแนกประเภทความคิดเห็นของร้านอาหารไทยได้ดีที่สุดในทุกชุดข้อมูลทั้งชุดข้อมูลที่ผ่านการเลือกคุณลักษณะและชุดข้อมูลที่ไม่เลือกคุณลักษณะ ทั้งหมด 9 ชุด โดยมีค่าความถูกต้องในการจำแนกประเภทความคิดเห็นไม่ต่ำกว่า 80% ในทั้งเก้าชุดข้อมูล ดังนั้นการเรียนรู้แบบแพร่กระจายย้อนกลับจึงเป็นประเภทการเรียนรู้ที่ดีที่สุดในการจำแนกประเภทความคิดเห็นในการทดลองนี้

การเลือกจำนวนโหนดที่เหมาะสมในชั้นซ่อนก็เป็นสิ่งสำคัญสำหรับโครงข่ายประสาทเทียมเช่นกัน ดังนั้น การทดลองนี้จึงได้นำชุดข้อมูลความคิดเห็นทั้ง 9 ชุด มาทำการจำแนกประเภทความคิดเห็นโดยใช้จำนวนโหนดในชั้นซ่อนที่แตกต่างกัน ดังนี้ คือจำนวน 2, 4, 8 และ 16 โหนด ดังแสดงในรูปที่ 4.13



รูปที่ 4.13 ค่าความถูกต้องของการเรียนรู้แบบแพร่กระจายย้อนกลับ โดยใช้ชุดข้อมูลความคิดเห็นที่มีจำนวนคุณลักษณะที่แตกต่างกัน และใช้โหนดในชั้นซ่อนจำนวน 16 ,8, 4 และ 2 โหนดตามลำดับ

จากรูป 4.13 ค่าความถูกต้องของการเรียนรู้แบบแพร่กระจายย้อนกลับ โดยใช้ชุดข้อมูลความคิดเห็นที่มีจำนวนคุณลักษณะที่แตกต่างกัน จากแผนภูมิจะเห็นได้ว่าค่าความถูกต้องของแต่ละจำนวนโหนดนั้นไม่แตกต่างกันมากนัก ดังนั้นในการทดลองนี้จึงเลือกจำนวนโหนดในชั้นซ่อนเท่ากับ 16 โหนด เพราะให้ความถูกต้องโดยเฉลี่ยในทุกชุดข้อมูลที่สูงที่สุดและใช้เวลาในการเรียนรู้น้อยไม่แตกต่างกับจำนวนโหนด อื่นๆด้วย ในการทดลองนี้จึงกำหนดให้โครงข่ายประสาทเทียมมีจำนวนโหนดในชั้นซ่อนเท่ากับ 16 โหนด

4.4 ขั้นตอนการทดลอง

1. ในการทดลองจะทำการทดลองนี้กับชุดข้อมูลความคิดเห็นที่ผ่านการเลือกคุณลักษณะโดยวิธีการเลือกคุณลักษณะแบบ mRMR
2. จำแนกประเภทความคิดเห็นโดยใช้โครงข่ายประสาทเทียม Radial Basis Function (RBF) และ Support Vector Machine (SVM)
3. MLP, RBF และ SVM ทำการจำแนกประเภทความคิดเห็นโดยแต่ละชุดข้อมูลจะทำซ้ำ 5 ครั้งเพื่อเลือกครั้งที่มีค่าความถูกต้องที่มากที่สุด
4. เปรียบเทียบประสิทธิภาพของการจำแนกประเภทความคิดเห็นในแต่ละวิธี

4.5 ผลการทดลอง

ในการทดลองนี้ได้ทำการกำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับ MLP, SVM และ RBF เพื่อให้สามารถจำแนกประเภทข้อมูลได้อย่างต้องมากที่สุด โดยวิธีการจำแนกแบบ MLP กำหนดให้มีจำนวนโหนดข้อมูลในชั้นซ่อนเท่ากับ 16 โหนด และใช้การเรียนรู้แบบแพร่กระจายย้อนกลับ(Backpropagation) วิธีการจำแนกแบบ RBF กำหนดให้ค่ารัศมีเท่ากับ 0.5 ค่าพารามิเตอร์ที่ถูกกำหนดเหล่านี้ ได้มาจากการทำการทดลองและปรับจนได้ค่าที่เหมาะสมซึ่งให้ความถูกต้องมากที่สุดในแต่ละวิธี

ตารางที่ 4.1 ผลการเปรียบเทียบการจำแนกประเภทความคิดเห็น

mRMR	**MLP**		**RBF**		**SVM**	
จำนวนคุณลักษณะ	**ความถูกต้อง (%)**	**เวลา (วินาที)**	**ความถูกต้อง (%)**	**เวลา (วินาที)**	**ความถูกต้อง (%)**	**เวลา (วินาที)**
50	88.3	0.15	84.4	0.02	**90**	0.02
100	**91.6**	0.22	78.8	0.03	88.4	0.05
150	**91.1**	0.14	75.9	0.03	86.9	0.05
200	**91.4**	0.16	74.4	0.03	90	0.05
250	**91.6**	0.21	73.7	0.03	88.2	0.06
300	**91**	0.17	74.4	0.03	87.1	0.09
350	**90.9**	0.22	70.8	0.03	87.8	0.07
400	**92.1**	0.34	70.2	0.03	87.7	0.03
1768	**91.2**	1.19	64.1	0.11	89.7	0.08
Average	**91**	**0.3**	**74.1**	**0.04**	**88.4**	**0.06**

ตารางที่ 4.1 เป็นการเปรียบเทียบค่าความถูกต้องในการจำแนกประเภทความคิดเห็น ด้วยวิธี MLP นั้นเปรียบเทียบกับ RBF และ SVM จากตารางผลการทดลองจะเห็นได้ว่า MLP สามารถจำแนกประเภทความคิดเห็นได้อย่างถูกต้องมากที่สุดเนื่องจากให้ค่าความถูกต้องในการจำแนกประเภทความคิดเห็นที่สูงที่สุดใน 8 ชุดข้อมูล ค่าความถูกต้องที่สุดในแต่ละชุดข้อมูลมีดังต่อไปนี้ ในชุดข้อมูลที่ไม่ผ่านการเลือกคุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91.2% และชุดข้อมูลที่ผ่านการเลือกคุณลักษณะโดยวิธี mRMR 400 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 92.1% ชุดข้อมูล 350 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 90.9% ที่ชุดข้อมูล 300 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91% ชุดข้อมูล 250 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91.6% ทุกชุดข้อมูล 200 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91.4% ชุดข้อมูล 150 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91.1% และ ชุดข้อมูล 100 คุณลักษณะ MLP มีค่าความถูกต้องสูงที่สุดเท่ากับ 91.6% และ ชุดข้อมูล 50 คุณลักษณะ SVM มีค่าความถูกต้องสูงที่สุดเท่ากับ 90% เมื่อคิดค่าเฉลี่ยความถูกต้องโดยรวมในทุกชุดข้อมูลแล้วพบว่า MLP มีค่าความถูกต้องเฉลี่ยสูงที่สุดที่ 91% ซึ่งสูงกว่า SVM และ RBF

ตารางที่ 4.2 ค่า True Positive Rate และ ค่า True Negative Rate ของ MLP, RBF และ SVM

จำนวนคุณลักษณะ	TPR			TNR		
	MLP	RBF	SVM	MLP	RBF	SVM
50	90.1	89.1	**92.2**	**91.2**	79.6	87.6
100	**93.9**	84.9	92.5	**90.9**	73.1	84.1
150	**92.4**	86.9	88.3	**91.5**	68	85.3
200	**94.3**	85.8	93.3	85.7	66.6	**86.5**
250	**93.2**	85.8	92.2	**88.8**	65.7	84.1
300	**93.7**	85.2	91.1	**86.5**	66.8	83
350	89.7	85.1	**91.5**	**93.5**	62.6	84
400	**95.9**	84.5	90.9	**84.4**	62.1	84.2
1768	**94.3**	91.3	91.8	86	56	**87.2**
เฉลี่ย	**93.1**	86.5	91.5	**88.7**	66.7	85.1

จากตารางที่ 4.2 เป็นการเปรียบเทียบค่า True Positive Rate และ ค่า True Negative Rate ของ MLP, RBF และ SVM จะเห็นได้ว่า MLP เป็นวิธีที่สามารถจำแนกประเภทความคิดเห็นได้อย่างมีประสิทธิภาพมากที่สุดเพราะ MLP ให้ค่า TPR มากที่สุดที่ชุดข้อมูล 100, 150, 200, 250, 300, 400 และ 1768 คุณลักษณะ และยังให้ค่าเฉลี่ยโดยรวมในทุกชุดข้อมูลมากที่สุดที่ 93.1% และให้ค่า TNR มากที่สุดที่ชุดข้อมูล 50, 100, 150, 250, 300, 350, 400 คุณลักษณะ และยังให้ค่าเฉลี่ยโดยรวมในทุกชุดข้อมูลมากที่สุดที่ 88.7% แสดงให้เห็นว่า MLP สามารถจำแนกประเภทความคิดเห็นทั้งเชิงบวกและเชิงลบได้ดีที่สุด

จากการทดลองพบว่า ชุดข้อมูลที่มีจำนวน 400 คุณลักษณะที่ได้จากวิธีการเลือกคุณลักษณะแบบ mRMR เป็นชุดคำสำคัญที่เหมาะสมที่สุดที่จะนำมาใช้ในการจำแนกประเภทความคิดเห็น เนื่องจากชุดคำสำคัญชุดนี้ประกอบด้วยคำที่สามารถบ่งบอกประเภทความคิดเห็นได้อย่างชัดเจนที่สุด ในตารางที่ 4.3 แสดงตัวอย่างชุดคำสำคัญจำนวน 100 คำจากทั้งหมด 400 คำ ที่ให้ค่าความถูกต้องในการจำแนกมากที่สุดจากการทดลอง

ตารางที่ 4.3 ลำดับคุณลักษณะและคำสำคัญ ที่เลือกมา 100คำ

คุณลักษณะ	คำสำคัญ	คุณลักษณะ	คำสำคัญ
280	great	1079	bar
297	excellent	471	reasonable
375	delicious	236	better
421	worst	594	maybe
504	poor	961	wonderful
613	disappointing	858	rude
543	amazing	1306	garden
447	overpriced	195	recommend
735	average	1020	lovely
688	bad	787	not good
1134	french	1367	maverick
1184	perfect	539	best
112	terrible	1070	poutine
1075	owner	189	bland
191	tasteless	33	bangkok
3	good	527	atmosphere
452	friendly	952	perfectly
242	worth	564	understand
579	nahm	450	ok
1093	relaxed	1383	team
402	sorry	334	fresh
1632	indigo	1210	homemade
455	not great	110	disappointment
484	fantastic	989	pizza
37	ordered	1081	short

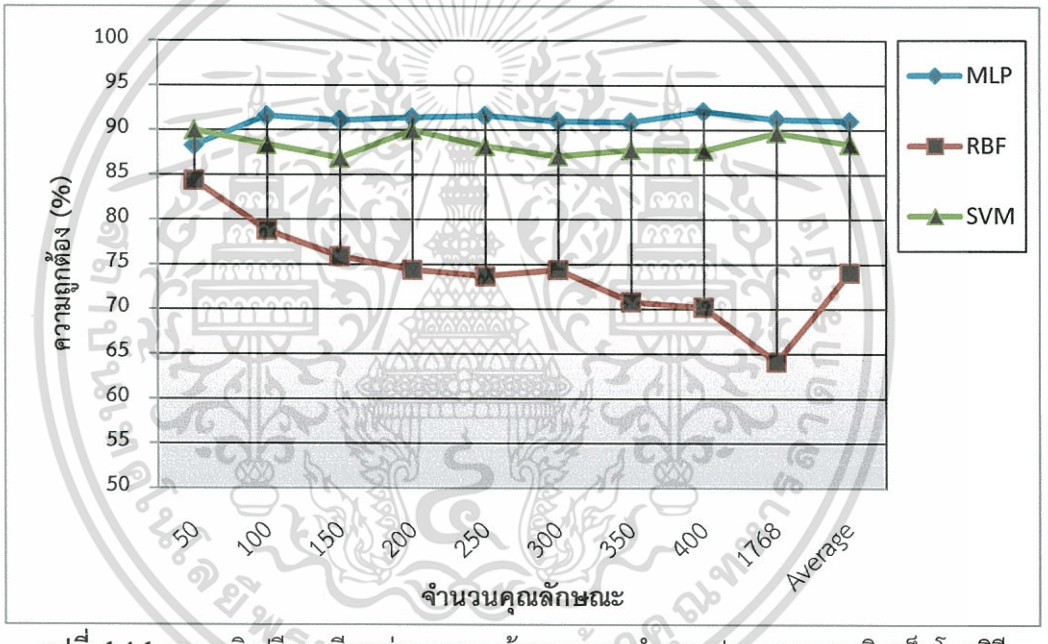
ตารางที่ 4.3 ลำดับคุณลักษณะและคำสำคัญ ที่เลือกมา 100 คำ (ต่อ)

คุณลักษณะ	คำสำคัญ	คุณลักษณะ	คำสำคัญ
17	left	892	loved
1384	work	144	manager
482	love	264	rock
696	worse	309	waitress
487	bill	606	not recommend
385	nothing	266	drink
558	ended	1076	welcome
475	definitely	233	back
781	slow	731	japanese
636	sure	391	since
170	old	1680	avoid
854	used	380	asked
488	unfortunately	651	ordinary
535	based	590	enjoyed
146	times	897	de
453	nice	123	word
164	shame	311	find
140	else	1157	not disappointed
261	much	582	not even
1011	selection	262	value
649	not worth	495	ate
1080	easy	1232	freshly
1635	silom	936	boyfriend
1185	fries	212	spicy
1082	khao	147	told

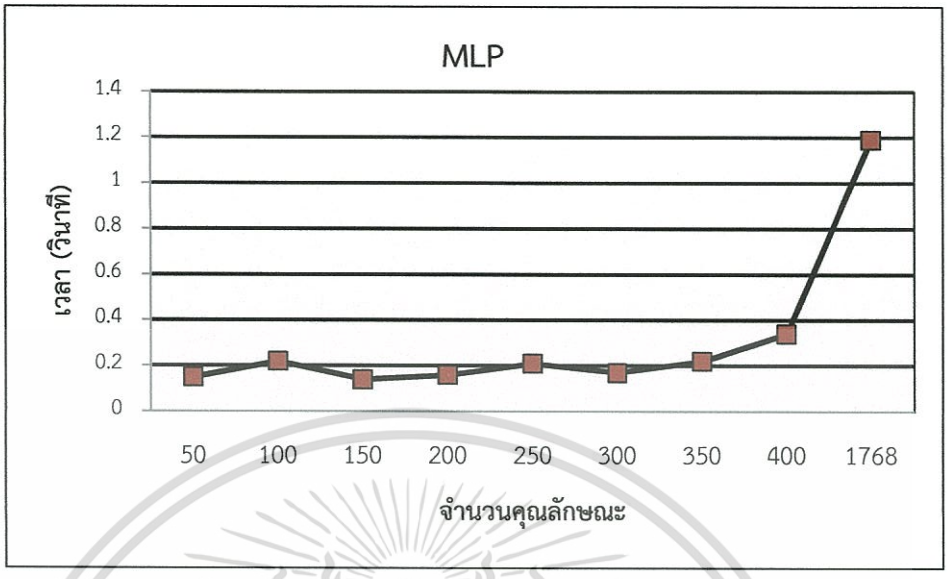
4.6 วิเคราะห์ผลการทดลอง

จากการทดลองจำแนกประเภทความคิดเห็นจะเห็นได้ว่าวิธีการที่นำเสนอนั้นสามารถนำมาใช้ในการจำแนกประเภทความคิดเห็นได้อย่างมีปประสิทธิภาพเพราะให้ความถูกต้องที่สูงที่สุด

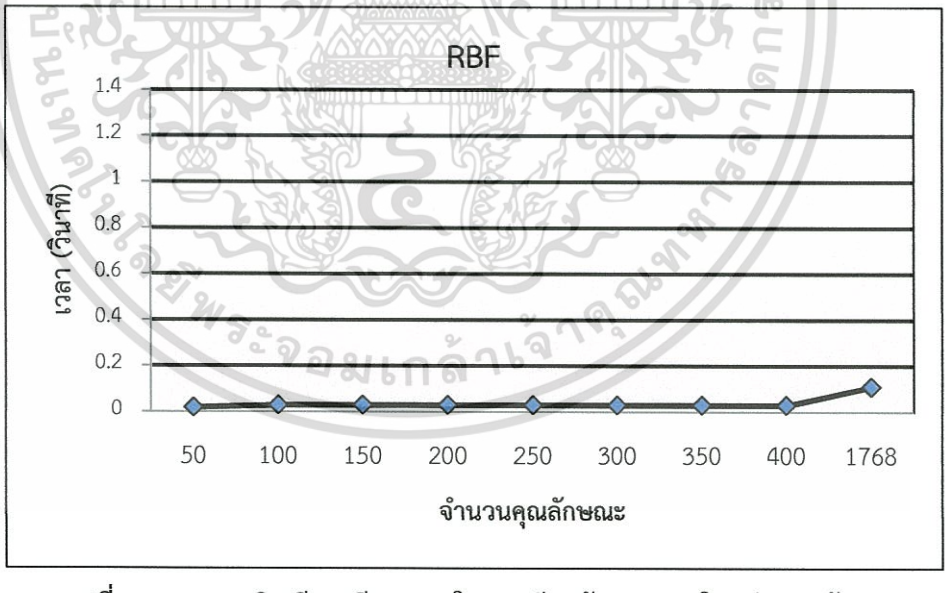
โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (MLP) เป็นวิธีการที่สามารถจำแนกประเภทความคิดเห็นของร้านอาหารไทยได้อย่างมีประสิทธิภาพมากที่สุดโดยมีค่าความถูกต้องโดยเฉลี่ยในการจำแนกโดยเฉลี่ยสูงที่สุดเมื่อเทียบกับ RBF และ SVM ในเกือบทุกชุดข้อมูลทั้งชุดที่ผ่านการเลือกคุณลักษณะ และชุดที่ไม่ผ่านการเลือกคุณลักษณะ ดังรูปที่ 4.14



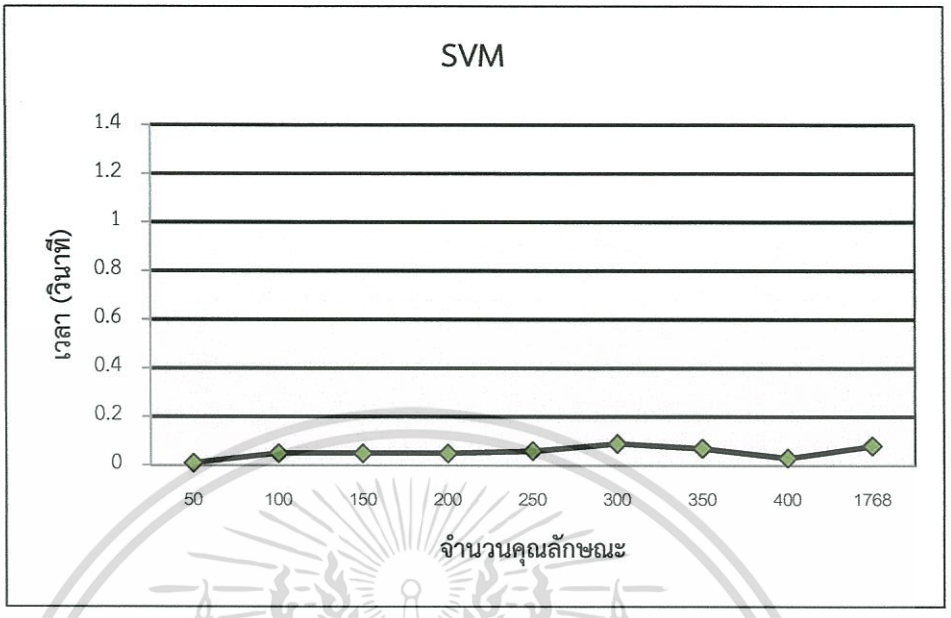
รูปที่ 4.14 แผนภูมิเปรียบเทียบค่าความถูกต้องของการจำแนกประเภทความคิดเห็นโดยวิธี MLP, RBF และ SVM



รูปที่ 4.15 แผนภูมิเปรียบเทียบเวลาในการเรียนรู้ของ MLP ในแต่ละชุดข้อมูล



รูปที่ 4.16 แผนภูมิเปรียบเทียบเวลาในการเรียนรู้ของ RBF ในแต่ละชุดข้อมูล



รูปที่ 4.17 แผนภูมิเปรียบเทียบเวลาในการเรียนรู้ของ SVM ในแต่ละชุดข้อมูล

จากรูปที่ 4.15 ถึง รูปที่ 4.17 แสดงการเปรียบเทียบเวลาในการเรียนรู้ของ MLP, RBF และ SVM จากชุดข้อมูลที่ผ่านการเลือกคุณลักษณะแบบ mRMR และชุดข้อมูลที่ไม่ผ่านการเลือกคุณลักษณะ โดยที่ MLP และ RBF นั้นชุดข้อมูลที่ผ่านการเลือกคุณลักษณะทุกชุดใช้เวลาในการเรียนรู้น้อยกว่าชุดข้อมูลที่ไม่ผ่านการเลือกคุณลักษณะและ SVM ที่ชุดข้อมูล 50, 100, 150, 200, 250 และ 400 คุณลักษณะนั้นใช้เวลาในการเรียนรู้น้อยกว่าชุดข้อมูลไม่ผ่านการเลือกคุณลักษณะ

จากการทดลองพบว่าวิธีการเลือกคุณลักษณะแบบ mRMR สามารถเลือกคุณลักษณะได้อย่างมีประสิทธิภาพ เพราะชุดข้อมูลที่ผ่านการเลือกคุณลักษณะนั้นทำให้ MLP สามารถจำแนกประเภทความคิดเห็นได้ถูกต้องมากยิ่งขึ้น โดยในชุดข้อมูลที่มีจำนวน 100, 200, 250 และ 400 คุณลักษณะ มากกว่าชุดข้อมูลเดิมที่ไม่ผ่านการเลือกคุณลักษณะ และเมื่อชุดข้อมูลผ่านการเลือกคุณลักษณะแล้วทำให้ชุดข้อมูลมีขนาดเล็กลง MLP จึงใช้เวลาในการเรียนรู้น้อยลงกว่าชุดข้อมูลเดิมที่ไม่ได้เลือกคุณลักษณะดังรูปที่ 4.15 ดังนั้นชุดข้อมูลที่เหมาะสมที่สุดที่ใช้ในการจำแนกประเภทความคิดเห็นคือชุดข้อมูลที่ผ่านการเลือกคุณลักษณะจำนวน 400 คุณลักษณะ เพราะให้ค่าความถูกต้องโดยเฉลี่ย ในการจำแนกที่ 92.1% ดังนั้นชุดคำสำคัญที่ได้จากการเลือกแบบเอ็มอาร์เอ็มอาร์ จำนวน 400 คำเหมาะสมที่สุดที่จะนำมาใช้ในการจำแนกประเภทความคิดเห็นของร้านอาหารไทย

บทที่ 5

สรุปและข้อเสนอแนะ

5.1 สรุป

งานวิจัยนี้ทำการพัฒนาวิธีการทำเหมืองข้อมูลจากข้อมูลความคิดเห็นของลูกค้าของร้านอาหารไทยโดยใช้โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (MLP) ในการจำแนกประเภทความคิดเห็น และวิธีการเลือกคุณลักษณะแบบ mRMR เพื่อหาชุดคำที่เหมาะสมของชุดคำสำคัญที่สามารถนำมาใช้ในการจำแนกประเภทได้อย่างถูกต้อง และช่วยลดเวลาในการเรียนรู้ประเภทความคิดเห็นของโครงข่ายประสาทเทียม จากการทดลองเพื่อปรับหาค่าพารามิเตอร์ที่เหมาะสมที่สุดในการสร้างโครงข่ายประสาทเทียมนั้นได้ทำการหาประเภทการเรียนรู้ที่สามารถเรียนรู้ประเภทความคิดเห็นได้ดี และจำนวนโหนดที่เหมาะสมในการสร้างโครงข่ายประสาทเทียมพบว่า โครงข่ายประสาทเทียมที่มีการเรียนรู้แบบแพร่กระจายย้อนกลับ (Backpropagation) และใช้จำนวนโหนดข้อมูลในชั้นซ่อน 16 โหนดเป็นโครงข่ายประสาทเทียมที่สามารถจำแนกประเภทความคิดเห็นได้อย่างถูกต้องโดยเฉลี่ยมากที่สุดและสามารถเรียนรู้ชุดข้อมูลได้อย่างรวดเร็ว

โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้นเมื่อทำการทดลองเปรียบเทียบกับ RBF และ SVM โดยทำการทดลองกับชุดข้อมูลทุกชุดคุณลักษณะพบว่าโครงข่ายประสาทเทียมสามารถจำประเภทความคิดเห็นได้ถูกต้องโดยเฉลี่ยมากที่สุด และชุดข้อมูลที่เหมาะสมที่สุดคือชุดข้อมูลที่มี 400 คุณลักษณะ ซึ่งชุดคำเหล่านี้เป็นคำที่สามารถบอกประเภทความคิดเห็นได้ดีที่สุดแสดงให้เห็นว่าการเลือกคุณลักษณะแบบ mRMR ช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทและช่วยลดเวลาในการเรียนรู้ได้เป็นอย่างดี ดังนั้นในการทำเหมืองข้อมูลความคิดเห็นของร้านอาหารไทยวิธีการที่นำเสนอเป็นวิธีที่มีประสิทธิภาพในการจำแนกประเภทความคิดเห็นตั้งแต่วิธีการเตรียมข้อมูล การแปลงข้อมูล วิธีการเลือกคุณลักษณะแบบ mRMR สามารถช่วยลดเวลาในการเรียนรู้และเพิ่มความถูกต้องให้กับโครงข่ายประสาทเทียม เพราะ mRMR สามารถเลือกเฉพาะคุณลักษณะที่สำคัญได้อย่างมีประสิทธิภาพ โครงข่ายประสาทเทียมที่นำเสนอยังสามารถจำแนกประเภทความคิดเห็นได้อย่างถูกต้อง 92.1% ซึ่งเป็นวิธีที่มีประสิทธิภาพมากที่สุดเมื่อเปรียบเทียบกับ RBF และ SVM

5.2 ข้อเสนอแนะ

จากผลการทดลองเพื่อศึกษาวิธีการจำแนกประเภทความคิดเห็นที่มีประสิทธิภาพนั้น วิธีการเลือกคุณลักษณะนั้นยังสามารถพัฒนาวิธีการเพื่อใช้ในการเลือกคุณลักษณะให้เหมาะสมมากกว่านี้เพราะบางชุดข้อมูลที่ผ่านการเลือกคุณลักษณะนั้นยังมีค่าความถูกต้องเมื่อผ่านการจำแนกประเภทแล้วน้อยกว่าชุดข้อมูลที่ไม่ผ่านการเลือกคุณลักษณะเล็กน้อย ปัญหาการเลือกคุณลักษณะนั้นยังคงต้องพัฒนาวิธีการต่อไปเพื่อให้สามารถเลือกคุณลักษณะได้อย่างมีประสิทธิภาพ

ปัญหาการเลือกความคิดเห็นเพื่อใช้ในการสร้างชุดข้อมูลฝึกสอนก็เป็นสิ่งสำคัญเพราะความคิดเห็นเหล่านี้จะถูกนำไปสร้างชุดข้อมูลเพื่อนำไปสอนให้กับโมเดลการเรียนรู้ของเครื่อง ซึ่งถ้าความคิดเห็นเหล่านี้ถูกเลือกมาโดยเป็นความคิดเห็นที่มีข้อมูลรบกวนอยู่มากหรือเป็นความคิดเห็นที่ไม่สามารถบอกประเภทความคิดเห็นได้เมื่อนำไปสร้างชุดข้อมูลฝึกสอนจะทำให้ชุดข้อมูลเต็มไปด้วยข้อมูลรบกวนส่งผลให้ชุดคำใน ชุดคำสำคัญ ที่ผ่านการเลือกออกมาเป็นคำที่ไม่สามารถบอกประเภทความคิดเห็นได้

ดังนั้นเพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ความคิดเห็น วิธีการที่นำเสนอนั้นเมื่อนำไปประยุกต์ใช้กับความคิดเห็นของสินค้าหรือบริการอื่นๆสามารถปรับเปลี่ยนพารามิเตอร์เพื่อให้เหมาะสมกับความคิดเห็นของสินค้าหรือบริการที่ต้องการจำแนกประเภทความคิดเห็นได้ เพื่อความถูกต้องสูงสุดในการจำแนกประเภทความคิดเห็น

เอกสารอ้างอิง

[1] H.Chen, D.Zimbra. 2010 "AI and Opinion Mining". IEEE Intelligent Systems". 25(3) : pp. 74 – 80.

[2] ธนาวุฒิ ประกอบผล. 2552. "โครงข่ายประสาทเทียม Artificial Neural Networks" วารสาร มฉก.วิชาการ 73. 12(24).

[3] J. Han, M. Kamber, J. Pei. 2012. "**Data Mining Concepts and Techniques**". Third Edition. Elsevier.

[4] H.Chen, D.Zimbra. 1990. "An algebraic proof for backpropagation in acyclic neural networks", 17 – 21. **IJCNN International Joint Conference**. San Diego : IEEE.

[5] S. Haykin. 2005. "**NEURAL NETWORKS A Comprehensive Foundation**" Second Edition. Ontarrio:Pearson Education.

[6] H. Peng, F. Long and C. Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and MinRedundancy". IEEE Pattern Analysis and Machine Intelligence. 27(8) : 1226 – 1238.

[7] M.S. Neethu, R. Rajasree. 2013. "Sentiment analysis in twitter using machine learning techniques". 1-5. **Computing, Communications and Networking Technologies.** Tiruchengode : IEEE.

[8] V.K. Singh, R. Piyani, A. Uddin, P. waila. 2031. "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification". 712 – 717. **Automation, Computing, Communication, Control and Compressed Sensing.** Kottayam : IEEE.

[9] V.B. Raut, D.D. Londhe. 2014. "Opinion Mining and Summarization of Hotel Reviews". 556 – 559. **Computational Intelligence and Communication Networks.** Bhopal : IEEE.

เอกสารอ้างอิง (ต่อ)

[10] Q. Cheng, H. Zhou, and J. Cheng. 2011 "The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multiclass Classification with Applications to High-Dimensional Data". IEEE Pattern Analysis and Machine Intelligence. 27(8) : 1217– 1233.

[11] O. Ludwig, U. Nunes. 2010. "Novel Maximum-Margin Training Algorithms for Supervised Neural Networks;". IEEE Transactions on Neural Networks. 21(6) : 972 - 984.

[12] B. Oluleye, A. Leisa J, L. Jinsong and D. Dean. 2014. "Zernike Moments and Genetic Algorithm: Tutorial and Application", British Journal of Mathematics & Computer Science, 4(15) : 2217- 2236.

[13] Battiti, R. 1992. "First and second order methods for learning: Between steepest descent and Newton's method," Neural Computation. 2(2) : 141–166.

[14] Scales, L.E. 1985. Introduction to Non-Linear Optimization. New York. Springer-Verlag.

[15] Powell, M.J.D. 1977. "Restart procedures for the conjugate gradient method," Mathematical Programming. 241–254.

[16] Cales, L.E. 1985. "Introduction to Non-Linear Optimization". New York. Springer-Verlag.

[17] Hagan, M.T., H.B. Demuth, M.H. Beale. 1996. "Neural Network Design". Boston. MA: PWS Publishing.

[18] Riedmiller, M., and H. Braun. 1993. "A direct adaptive method for faster backpropagation learning: The RPROP algorithm,". **IEEE International Conference on Neural Networks**. 586–591. IEEE.

ภาคผนวก

ผลงานที่ได้รับการตีพิมพ์

Computer Science and Engineering Conference (ICSEC),

2014 International Conference

Opinion Mining for Thai Restaurant Reviews using Neural Networks and mRMR Feature Selection

Niphat Claypo and Saichon Jaiyen

Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand

s5650805@kmitl.ac.th, kjsaicho@kmitl.ac.th

Abstract— **Currently, Thai restaurants are popular around the world. There are tons of reviews related to foods and services in social networking websites. These tons of customer reviews make it difficult to analyze the opinions of customer toward foods and services. To help the businesses, the model of opinion mining is proposed for classifying the reviews and to analyze the attitude of customers for improving their products and services. In this research, the artificial neural network is applied to classify the positive and negative reviews. In addition, the mRMR feature selection is used to select the features of data in order to reduce the number of features in the data set. Consequently, the computational times of learning algorithms for neural networks are reduced. The experimental results show that the neural network is an effective model for classifying the Thai restaurant reviews.**

Keywords— Classification; Feature selection; minimal-redundancy-maximal-relevance (mRMR); multilayer perceptron (MLP); radial basis function (RBF); Support Vector Machine (SVM).

I. INTRODUCTION

The opinion mining is an approach for analyzing the user comments for products and services. The user opinions can be collected from the social media, e-commerce websites, product review sites, blogs etc. These collected comments are divided into two classes which are positive and negative comments. There are many techniques are used to classify the data in the opinion mining. The classification methods based on machine learning including Support Vector Machine (SVM), Naïve Bayesian (NB), K-nearest neighbor (KNN) are most widely used [3], [4], [5], [13]. Classifying customer reviews for online products using Sentence Weight algorithm is proposed [6]. The Sentiment fuzzy classification algorithm is proposed to predict the positive and negative comments for the movie review [7]. The Back-Propagation Neural Network (BPANN) is used for classifying the movie and hotel reviews [8]. The self-organizing map (SOM) is used to extract the feature of data vectors [9]. The opinions on Twitter micro blog data are used for classifying the positive and negative opinions [10]. The text pre-processing techniques including Tokenization, Stop-word removal (STR), Lemmatization (LM), Number replacement (NMR), Synonym recognition (SYR) and Word generalization (WG) are proposed to optimize the experiments [12]. The natural language processing techniques based on sentiment analysis are applied to online reviews to find the most influential part-of-speech [14]. The mRMR feature selection is used to select the features of omics data based on three relevance evaluation measures including MI, CC, and MIC [15]. A two-stage feature selection algorithm by combining ReliefF and mRMR are used for finding a set of genes [16].

On social networking websites, users can post comments or reviews about products and services. The opinions of products, evens, news, and so on can be found in Blogs, social media, forums and others. For online business, the user opinions are very important because they can help the business to understand the user needs and feelings. Therefore, the opinion mining is the popular tool for analyzing their customers. Furthermore, it can be applied to business intelligence, ecommerce, etc.

In this paper, we propose opinion mining method for classifying the reviews of Thai restaurant using multilayer perceptron neural network (MLP) and mRMR feature selection. The data used in the experiments are Thai restaurant reviews consisting of positive and negative reviews. The text preprocessing techniques are used for preparing the dataset. The neural network model is used to classify the positive and negative reviews and the mRMR feature selection technique is used for selecting the features of data set for reducing the number of the features of data in the data set. The experimental results are compared with Support Vector Machines (SVM) and Radial Basis Function Neural Network (RBF). The experimental results show that the multilayer perceptron neural network gives the better performance.

II. OPINION MINING

Web 2.0 technology and social media cause many opinions on the websites. There are a lot of opinions on websites toward social events, political movements, company strategies, marketing campaigns, and product preferences, and so on. The comments and reviews can be found in various blogs, forums, social media and social networking sites, virtual worlds, and tweets. The opinion mining is a discipline of web mining and computational intelligence which is used to gather the opinions in various online sources, social media comments, and other user-generated contents for extracting, classifying, and understanding. The opinion mining can be applied for sentiment analysis [1]. The opinion mining collects opinions from websites for classifying through the mining process such as SVM, MLP, decision tree, and so on. The opinion mining can help business to know the positive or negative attitude of their customers about their products and services. Moreover, the opinion mining can help them to understand the advantages and disadvantages of their products and services in order to improve their further products and services.

978-1-4799-4963-2/14/$31.00 ©2014 IEEE

III. THE PROPOSED METHOD

A. Text Preprocessing

In this research, the text preprocessing technique is used for data preparation which is divided into two steps. The first step is tokenization used to split the reviews into tokens or words. The second step is the removal of the stop words which are words that have no meaning or do not make the meaning of the sentence changes when remove these words such as "a", "an", "the", "of", "I", etc. Each document consists of many stop words. When these stop words are deleted form the document, it can make the number of words in documents lower. Consequently, the size of data is reduced and the computational time is also decreased.

B. Text Transformation

The data used in the experiments must be appropriate for the classification methods. Because the methods used in the experiments are the computational method, they must use the data in numeric format as their input. Therefore, the documents which are in the text format must be transformed to the numeric format. In this paper, the words after removing all stop words of all positive and negative comments in the training set and testing set, the words in the training set are used as a keyword for creating the input vectors. Then, each document is transformed into the input vector by calculating the frequency of keywords appearing in that document.

C. Minimum Redundancy and Maximum Relevance (mRMR) Feature Selection

The mRMR is a popular method applied to select the features by using the relationship between the feature and the target class in order to reduce the size of the dataset. The basic idea of minimum redundancy is to choose the features such that they are mutually maximally dissimilar to other features. Let S denote the subset of features and |S| is the number of features in S. The minimal redundancy (Min-Redundancy) condition is

$$\min R(S),\ R(S) = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i; x_j) \quad (1)$$

where c is the target class and $I(x_i; x_j)$ is the mutual information between the individual feature x_i and x_j.

The mutual information between individual feature x_i and the target class c is the measure of relevance of that feature. Thus, maximal relevance criterion (Max-Relevance) is to maximize the average relevance of all features in S:

$$\max D(S, c),\ D(S, c) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) \quad (2)$$

where $I(x_i; c)$ is the mutual information between individual feature x_i and the target class c. The mRMR feature set is obtained by optimizing the conditions in Equation (1) and (2) simultaneously.

D. Artificial Neural Network

Artificial neural network is one of disciplines of artificial intelligence (AI). Artificial neural network is the computational model that simulates the function of neurons in the human brain. In data mining neural network is another popular to do classification that can classify the data correctly and efficiently. In this paper, the opinion mining model using multilayer perceptron (MLP) neural network are proposed in order to classify the reviews of Thai restaurants. The MLP consists of three layers including input layer, hidden layer, and output layer as shown in Fig. 1. The input layer is the first layer to get data. The hidden layer is the computational layer that maps the input data in the input space into a feature space where it becomes linearly separable. The last layer is output layer used to identify the class of data.

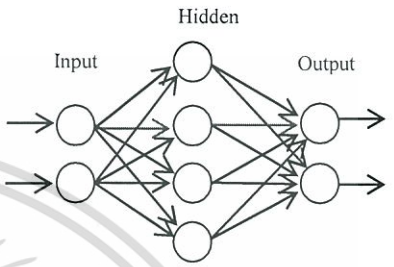


Fig. 1. Example of a simple neural network.

The MLP use the supervised learning algorithm to learn the data. The data set is divided into training and testing set. The training set consists of learning sample data and the answers called "target". The training data is fed into the network to calculate the output. Then, the output is compared to the answers to find the error for adjusting weights. These weights will be applied to other unseen data to predict the class of input data. The learning algorithm for MLP is called Back-propagation algorithm. The processes of the learning algorithm are shown in the following algorithm.

Back-propagation Algorithm

1. Input the training data to the MLP neural network and compute the network outputs.
2. For each output neuron j, calculate the local gradient δ_j defined by

$$\delta_j = e_j \varphi'(\mathrm{v_j}) \quad (3)$$

where $\varphi(\cdot)$ is the activation function, $v_j = \sum_i w_{ji} y_i$, and $y_j = \varphi(v_j)$.

3. For each hidden neuron j, calculate the local gradient δ_j defined by

$$\delta_j = \varphi'(v_j) \sum_{\mathrm{k}} \delta_{\mathrm{k}}(\mathrm{w_{kj}}) \quad (4)$$

4. Update each network weight w_{ji} by

$$w_{ji}^{new} = w_{ji}^{old} + \Delta w_{ji} \quad (5)$$

where $\Delta w_{ji} = \eta \delta_j x_i$ and η is the learning rate.

IV. EXPERIMENT

A. Data Sets

The reviews of Thai Restaurant used in this paper are collected from th.tripadvisor.com. There are 1,060 reviews randomly collected from this website. Each reviewed document include vocabularies, stop words, numbers and non-alphabet characters. Before training, the reviews documents are preprocessed by using text-preprocessing method and text transformation method as mention above and yield the results as the input vectors. These input vectors are divided into two classes including class 1 representing the positive review and class 0 representing the negative review. The dimension of the input vector is equal to the number of keywords in all comments after using text-preprocessing method. All data is divided into two set which are the training set and the testing set. In the training set, there are 510 input vectors with 1768 features and there are 550 input vectors with 1768 features in the testing set.

B. Experimental Results

In this experiment, three types of classifiers including the proposed MLP, RBF, and SVM are applied to solve this classification problem. To evaluate the performance, the experimental results of the proposed MLP method are compared with the results of RBF and SVM. This experiment is conducted using the neural network toolbox on Matlab. For the proposed MLP, we conduct the experiment with 2, 4, 8, 16 hidden neurons and the hidden neurons that give the best accuracy are chosen. After trial and error, the 8 hidden neurons that give the best accuracy are chosen. For RBF, we try to change the spread of the basis function and select the spread that give the best accuracy. After trial and error, the spread is set as 0.5 for RBF. The mRMR feather selection method is used to select the features of data. From the data set, we generate new six data sets. Each set has a number of features as 50, 100, 150, 200, 250 and 300 features.
All data sets are divided into training and testing sets. The simulations are run on all pairs of training and testing sets. The experimental results of the proposed MLP are then compared with RBF and SVM as shown in Table I. In the experiment, the comparative results from the classification by the proposed MLP, RBF, and SVM are illustrated in Table I. For all methods, reviews are classified into either positive or negative classes. The experimental results include accuracy and time obtained from six pairs of training and testing sets along three types of classification method. All three methods are simulated three times for all training and testing sets and the best accuracy are chosen. According to Table I, it has shown that the mRMR feature selection can reduce a number of features that make the data smaller and less training time in all classification methods, while it does not reduce the accuracy of all methods as well. From the comparative results, the proposed MLP method can achieve the best performance at 93.5% of accuracy, while SVM gains 90% of accuracy and RBF gains 78.9% of accuracy.

Fig. 2 shows the accuracy of MLP trained by all features and the accuracies of MLP trained by selected features using mRMR method. The accuracies of six data sets with selected features are higher than the accuracy of the data set with all features. Fig. 3 shows the accuracy of RBF trained by all features and the accuracies of RBF trained by selected features using mRMR method. The accuracies of six data sets with selected features are higher than the accuracy of the data set with all features. Fig. 4 shows the accuracy of SVM trained by all features and the accuracies of SVM trained by selected features using mRMR method. The accuracies of five data sets with selected features are higher than the accuracy of the data set with all features. However, the accuracy of SVM for the data set with 150 features is slightly less than the data set with all features. Fig. 5 illustrates the time comparison of the three methods using all features and selected features. The training times of MLP and RBF method with all features are very higher than the data sets with selected features using mRMR method. However, the training time of SVM method is almost no difference between all features and selected features because the number of all features is not sufficiently large. From the experimental results, the input vectors with 200 features give the best accuracy. Therefore, the mRMR feature selection can reduce the number of features. Enhance the accuracy and use less times for classifying reviews of Thai restaurants. From the experimental results, all three classification methods can efficiently classify the review data. However, the experimental results show that the proposed MLP is most effective for classifying reviews of Thai restaurants.

Table I. Comparison result of MLP, RBF, and SVM

Method	Non feature select		mRMR Feature Selection											
	1768		50		100		150		200		250		300	
	Time (second)	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*
MLP	24.31	90	0.58	91.7	0.65	92.2	0.697	92	0.8	93.5	0.58	92.2	0.47	90.9
RBF	35.3	72.5	0.24	85.3	0.24	78.9	0.24	77.5	0.24	75.7	0.43	74.4	0.65	75.9
SVM	0.54	89.7	0.2	90	0.5	88.9	0.39	86.9	0.6	90	0.33	88.2	0.3	87.1

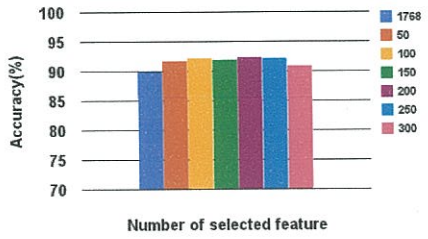


Fig. 2. Comparison of feature selection using MLP.

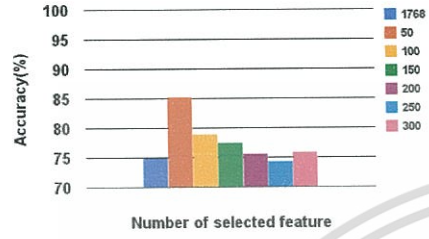


Fig. 3. Comparison of feature selection using RBF.

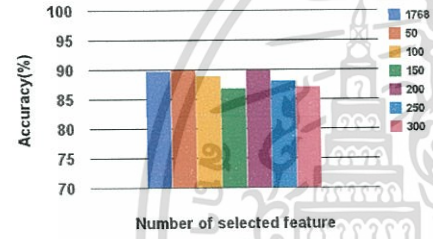


Fig. 4. Comparison of feature selection using SVM.

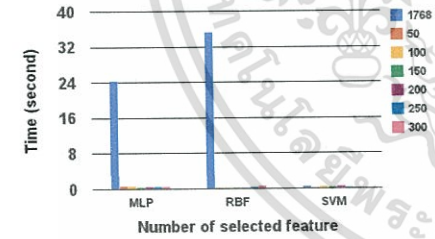


Fig. 5. Comparison time of the three methods using all features

V. CONCLUSION

In this paper, we propose the model to classify the opinions of customers for Thailand restaurants using opinion mining. For the proposed method, the mRMR feature selection is adopted to optimize the classification of Thailand restaurant reviews and the neural networks are used to classify them into positive and negative reviews. The experimental results have shown that high accuracy can be gained in all classification methods. When mRMR method is used to reduce the size of data set, it enhances the performance of the neural networks, spends less training time, and preserve high accuracy. Furthermore, an advantage of the neural networks is that they can adjust themselves to recognize patterns of the reviews and thus, they result in high accuracy in the classification. From the experimental results, they have shown that the proposed method, which utilizes the benefit of mRMR feature selection and MLP neural network, can achieve the best performance for opinion mining for Thai restaurant reviews.

VI. REFERENCES

[1] H.Chen and D.Zimbra, "AI and Opinion Mining", Intelligent Systems, Vol. 25, pp. 74 – 80, IEEE. 2010.

[2] H. Peng, F. Long and C. Ding. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", Pattern Analysis and Machine Intelligence, Vol. 27, pp. 1226 - 1238, IEEE, 2005.

[3] C. Zhang, W. Zuo, T. Peng and F. He, "Sentiment Classification Reviews Using Machine Learning Methods Based on String Kernel", Convergence and Hybrid Information Technology, Vol. 2, pp. 909 – 914, IEEE, 2008.

[4] A.Khan, B.Baharudin, K.khan, "Sentence Based Sentiment Classification from Online Customer Reviews", Frontiers of Information Technology, ACM, 2010.

[5] N.Aleebrahim, M.Fathian and M.Reza Gholamian, "Sentiment Classification of Online Product Reviews Using Product Features", Data Mining and Intelligent Information Technology Applications, pp. 242 – 245, IEEE, 2010.

[6] X.Hu and B.Wu, "Classification and Summarization of Pros and Cons for Customer Reviews", Web Intelligence and Intelligent Agent Technologies, Vol. 3, pp. 73 – 76, IEEE, 2009.

[7] K.Mouthami, K.Nirmala Devi and V.Murali Bhaskaran, "Sentiment Analysis and Classification Based On Textual Reviews", Information Communication and Embedded Systems, pp. 271 – 276, IEEE, 2011.

[8] A.Sharma and S.Dey, "A Document-Level Sentiment Analysis Approach Using Artificial Neural Network and Sentiment Lexicons", ACM SIGAPP Applied Computing Review, VOL. 12, pp. 67-75, ACM, 2012.

[9] S.Nirkhi, "Potential use of Artificial Neural Network in Data Mining", Computer and Automation Engineering, Vol. 2, pp. 339 – 343, IEEE, 2010.

[10] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N.Prasath, A. Perera, "Opinion Mining and Sentiment Analysis on a Twitter Data Stream", Advances in ICT for Emerging Regions, pp. 182 – 188, IEEE, 2012

[11] G. R. Brindha and B. Santhi, "Application of Opinion Mining Technique in Talent Management", Management Issues in Emerging Economies , pp. 127 – 132, IEEE, 2012.

[12] Z. Ceska and C. Fox, "The Influence of Text Pre-processing on Plagiarism Detection", Industrial and Information Systems, pp. 376 – 380, IEEE, 2009.

[13] K. Gayathri, A. Marimuthu, "Text Document Pre-Processing with the KNN for Classification Using the SVM", Intelligent Systems and Control, pp. 453 – 457, IEEE, 2012.

[14] S. Thanangthanakij, E. Pacharawongsakda, N. Tongtep, P. Aimmanee, T. Theeramunkong, "An Empirical Study on Multi-Dimensional Sentiment Analysis from User Service Reviews", Knowledge, Information and Creativity Support Systems, pp. 58 – 65, IEEE, 2012.

[15] J. Yang, Z. Zhu, S. He and Z. Ji, "Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification", Computational Intelligence in Bioinformatics and Computational Biology, pp. 246 – 251, IEEE, 2013.

[16] Y. Zhang, C. Ding and T.Li, "A Two-Stage Gene Selection Algorithm by Combining ReliefF and mRMR", Bioinformatics and Bioengineering, pp. 164 – 171, IEEE, 2007.

[17] M. Wisniewski and T. P. Zielinski, "MRMR-based feature selection for automatic asthma wheezes recognition", Signals and Electronic Systems, pp. 1 – 5, IEEE, 2010.

Knowledge and Smart Technology (KST), 2015 7th International Conference

Opinion Mining for Thai Restaurant Reviews using K-Means Clustering and MRF Feature Selection

Niphat Claypo and Saichon Jaiyen

Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand

s5650805@kmitl.ac.th, kjsaicho@kmitl.ac.th

***Abstract*—Opinion mining on millions of Thai restaurant reviews in an unsupervised manner is a challenging task to survey feedbacks of the customers on their products and services. This is extremely helpful for owners to improve their business. In this paper, we propose an opinion mining on Thai restaurant reviews using K-Means clustering and MRF feature selection. The proposed method begins with text preprocessing for breaking reviews into words and removing stop words, followed by text transformation for creating keywords and generating input vectors. MRF feature selection is subsequently adopted for selecting relevant features from a large number of features extracted. Then, K-Means is employed for clustering into positive and negative reviews. From the experimental results, MRF feature selection can efficiently reduce the number of features in the data set so the computational time is significantly decreased. In addition, K-means can achieve the best clustering performance, when compared with Self-Organizing Map, Fuzzy C-Means, and Hierarchical Clustering. Thus, the cooperation of K-means with MRF feature selection is an effective model for clustering Thai restaurant reviews.**

Keywords-Opinion Mining; K-Means; Self-Organizing Map neural network (SOM); Fuzzy C-Means (FCM); MRF feature selection; Hierarchical.

I. INTRODUCTION

Opinion mining is a web mining technique for analyzing underlying sentiments from user-generated content, such as online product reviews, blogs, discussion forums, and so on. P. Jiang and et al. [3] presented opinion mining of online product reviews using tree kernels for sentiment expression extraction and sentiment classification. Their experimental results illustrated that tree kernels can achieve a relatively high performance when trained by small data set. Then, X. Yu and et al. [4] proposed Sentiment PLSA (S-PLSA) in the movie domain for sales performance prediction. It can help us move from simple "negative or positive" classification toward a deeper comprehension of the sentiments in blogs. Amir A. Sheibani [5] conducted his research on finding post spam reviews of products by review spam recognition. In addition, K. Jędrzejewski and M. Morzy [6] discussed the role and importance of social networks followed by presenting P-proportional method for classification real world data sets. Likewise, R. Xu and C. Kit [7] also presented a Corse-fine opinion mining framework for classifying NTCIR-6 and NTCIR-7 data sets.

In unsupervised classification, k-means is widely used method for data clustering. T. Hitendra Sarma and et al. [8] proposed a two stage hybrid approach to fast kernel k-means clustering method and the experimental result showed that, with a small loss of quality, their method can significantly reduce the time taken than the conventional kernel k-means clustering method. Furthermore, S. N. Sulaiman and N. A. Mat Isa [9] presented a new clustering algorithm called Adaptive Fuzzy-K-means (AFKM) clustering for image segmentation, and it illustrated efficient segmentation results. Additionally, K-means was combined with manifold learning algorithms into a coherent framework in order to improve k-means algorithm [10, 11]. The results showed that the method can effectively improve the speed and accuracy of clustering, while reduce the computational complexity of k-means simultaneously. Particularly, the framework KCM (K-means clustering with manifold) has shown the good clustering results on UCI data sets.

Fuzzy C-Means (FCM) [2] is a typical clustering algorithm used for clustering large data. It is one of unsupervised learning algorithms in machine learning. Timothy [12] applied fuzzy c-means (FCM) method for clustering very large (VL) data and compared it with three methods. The experimental results show that it is a good choice for approximating the VL data, B. Mohamed [13] proposed New Allied Fuzzy C-Means algorithm for Takagi-Sugeno Fuzzy model Identification. The Particle Swarm Optimization method (PSO) combined with the NAFCM algorithm (NAFCM-PSO) is applied. The experimental results demonstrated that the NAFCM algorithm combined with the PSO algorithm give the best result of identification.

On websites and social network, users can post messages to comment their products and services of restaurants. There are a large number of these opinions distributed over the networks. From the viewpoint of business, these opinions and reviews can help them know the users' needs and feelings because each review report expresses what the user thinks about their products and services. Therefore, opinion mining methods are used to understand the opinions, and machine learning algorithms are adopted as the popular tool to analyze them. This can help the business to understand various kinds of the opinion groups and patterns.

Since clustering of the user opinions is beneficial to various viewpoints of business, this paper proposes opinion mining method for clustering the reviews of Thai restaurant using K-Means clustering and MRF feature selection. The data sets used in this experiment are the reviews of Thai restaurants collected from travel guide website. The data sets are transformed to the input vectors of the clustering model by using text-preprocessing and text transformation methods. In addition, MRF feature selection technique is also used for selecting relevant features in the data sets. Then, the size of data is reduced so that they contained only relevant features.

978-1-4799-6049-1/15/$31.00 ©2015 IEEE

Consequently, K-Means clustering is used for clustering reviews of Thai restaurants into positive and negative groups. To evaluate the performance, the experimental results were compared with Fuzzy C-Means algorithm (FCM), Self-Organizing Map neural network (SOM) [14, 15] and Hierarchical. For this experiment, it has shown that K-Means clustering can achieve the best performance of clustering Thai restaurant reviews.

The remaining of this paper is organized as follows. Firstly, Section II provides a fundamental knowledge of opinion mining. Secondly, Section III explains the proposed methods including procedures of text preprocessing, text transformation, MRF feature selection, and K-Means Clustering. Thirdly, Section IV describes the data sets and parameter setting used in this experiment followed by discussing the experimental results. Lastly, Section V concludes all of this paper.

II. OPINION MINING

Opinion mining is also known as Sentiment analysis that can analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. Opinion mining consists of various techniques in information retrieval, natural language processing and machine learning. The techniques are used to analyze opinions, comments and reviews of people to find what people think. Currently, social network can cause many opinions on the websites. The opinions of people on website are collected for classifying by the mining process. The comments and reviews can be found in various blogs, forums, social media and social networking sites, virtual worlds, and tweets [16]. The machine learning process can help us know patterns of the opinions. From the viewpoints of business, opinion mining can help the owners understand their customers' needs and feelings through reviews of products and services. This helps agencies and organizations find information feedback to the customer relationship management. Consequently, the opinion mining can help the owners improve their products and services that will be launched in the future.

III. THE PROPOSED METHOD

A. Text preprocessing

In this experiment, the text preprocessing is divided into two steps. The first step is tokenization used to split the review documents into tokens or words. The second step is the removal of the stop words, which are words that have no meaning or do not make the meaning of the sentence changes when remove them such as "is", "to", "and", "of", "in", etc. Each review document consists of many stop words. When these stop words are deleted from the document, it make the number of words in documents lower. Therefore, the size of data is reduced and the computational time is also decreased.

B. Text Transformation

For text transformation, it is divided into two steps. The first step is creating a list of keywords. After removing all stop words in both of positive and negative comments of the dataset set, the words obtained from 510 reviews are used as keywords. Moreover, duplicate words in 510 reviews are also removed such that the obtained keywords are the unique words. These keywords are subsequently used for creating the input vectors. Since feeding an appropriate format of inputs to the clustering methods is necessary, the second step will transform the documents, which are in the text format, to the numeric format. Then, each document is transformed into the input vector by calculating the frequency of keywords appearing in it. The obtained input vectors are subsequently fed to the clustering methods.

C. MRF feature selection

MRF feature selection is a method applied to select the features using Markov random field optimization techniques [1].The MRF feature selection uses the value of γ and the value of β for determining the number of chosen features. These values are the global threshold. The MRF algorithms are as follows.

Fisher-Markov Selector with Linear Polynomial Kernel (LFS) algorithm with $d = 1$.

1. Input a data matrix of data sets example $[x_1, \ldots, x_n] \in R^{p \times n}$ for g groups. For the vector of group labels of the data sets $y = [x_1, \ldots, x_n]$ where $y_k \in \{w_1, \ldots, w_n\}, k = 1, \ldots n.$
2. Compute the Markov coefficients θ_j by

$$\theta_j = \frac{1}{n}\sum_{i=1}^{g}\frac{1}{n_i}\sum_{u,v=1}^{n_i} x_{vj}^{(i)}x_{vj}^{(i)} - \frac{\gamma}{n}\sum_{i=1}^{n} x_{ij}^2 + \frac{\gamma-1}{n^2}\sum_{u,v=1}^{n} x_{uj}x_{vj} \quad (1)$$

3. Solve LFS problem by

$$\theta_j > \beta \Leftrightarrow \alpha_j^* = 1 \quad (2)$$

4. The output is the estimated feature selector of α^*.

Fisher-Markov Selector with Quadratic Polynomial Kernel (QFS) algorithm with $d = 2$.

1. Input a data matrix of data sets example $[x_1, \ldots, x_n] \in R^{p \times n}$ for g groups. For the vector of group labels of the data sets $y = [x_1, \ldots, x_n]$ where $y_k \in \{w_1, \ldots, w_n\}, \ k = 1, \ldots n.$
2. Compute the Markov coefficients θ_j by (2) and θ_{jl} by

$$\theta_{jl} = \frac{1}{n}\sum_{i=1}^{g}\frac{1}{n_i}\sum_{u,v=1}^{n_i} x_{vj}^{(i)}x_{vj}^{(i)}x_{vj}^{(i)}x_{vj}^{(i)}\frac{\gamma}{n}\sum_{i=1}^{n} x_{ij}^2 x_{ij}^2$$

$$+ \frac{\gamma-1}{n^2}\sum_{u,v=1}^{n} x_{uj}x_{vj}x_{uj}x_{vj},\ 1 \le \mathrm{j,l} \le \mathrm{p}, \quad (3)$$

3. Solve MRF maximization problem of (QFS) by MRF solver.
4. The output is the estimated feature selector of α^*.

D. K-Means Clustering

K-Means is a typical clustering algorithm which is one of unsupervised learning algorithms in machine learning. It is widely used for solving clustering problems in various kinds of application. In this experiment we apply K-Means clustering to classify Thai restaurant reviews. K-Means clustering assign all data into kth groups by minimizing the objective function which is defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} ||p - \mu_i||^2 \quad (4)$$

where k is number of clusters, μ_i is the centroid of clusters C_i, and p is the data point in each clusters. In K-Means clustering, firstly, the K points are placed in the space. These points represent initial cluster centroids. Secondly, each data is assigned to the cluster that has the closest centroid based on the Euclidean distance. The Euclidean distance can be defined as

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_n - y_n)^2}. \quad (5)$$

Thirdly, after all data have been assigned, the positions of the K centroids are recalculated. Finally, these steps excepting the initial step are repeated until the centroids no longer move [10].

IV. Experiment

A. Data Sets

The data used in this experiment are reviews of Thai restaurants collected from th.tripadvisor.com. The 1060 reviews are randomly picked from the website. Each review includes vocabulary, stop words, numbers and non-alphabet characters. For this experiment, all reviews in the data set are preprocessed by text-preprocessing and text transformation methods for creating input victors. From 1060 reviews, 510 reviews are used as the training set which are used for creating the keywords and 550 reviews are used as the testing set. The input vectors are computed from the frequency of the obtained keywords. So, there are 550 input vectors, which each of them contains 1768 features. In this experiment, the data set is divided in to two groups, the first group contains the positive reviews and the second group contains the negative reviews. In addition, the MRF feature selection technique is used for selecting relevant features of the data sets. In order to find an optimal performance of this technique, the β values are defined differently. Accordingly, the beta values are set as 0.05, 0.04, 0.03, 0.02 and 0.01, while γ value is set as -0.5 for defining the number of selected features. In this experiment, we adjust the value of β for choosing features until this value optimized for each data set. The MRF feature selection can reduce the number of features and make the data smaller. Thus, features in the input vectors of the data set are selected by MRF feature selection. After selection, new six data sets are generated. Each set had 550 input vectors with the number of features as 103, 135, 183, 261 and 467, respectively.

B. Expeimental Results

In this experiment, an unsupervised learning algorithm is used for analyzing the reviews of Thai restaurants. Four types of clustering algorithms, which are K-Means, Hierarchical, Fuzzy C-Means (FCM) and Self-organizing map (SOM), are selected. Then, the experimental results of the proposed K-Means clustering are compared with these methods to evaluate the performance. In this experiment, Matlab toolbox is used for conducting the experiment. For the configuration of the clustering algorithms, a type of distance measurements in K-means and Hierarchical clustering is defined as Euclidean distance, and numbers of cluster of SOM are defined as 2. All clustering methods are simulated on six data sets, and each method is run five iterations for each data set to choose the best accuracy. All methods clustered the reviews into two groups each group include positive or negative reviews. The experimental results of K-Means clustering are compared with the results of FCM, Hierarchical and SOM, as illustrated in Table I. For the data set with 103 features, K-Means can achieve the best accuracy at 71.7%. The accuracy of K-Means is higher than the accuracy of other methods. Besides, K-Means can also achieve the best accuracy at 70.3% for the data set with 135 features, while K-Means can obtain the best accuracy at 74.6% and 75.5% for the data sets with 183 features and 261 features, respectively. Then, K-Means can achieve the best accuracy at 74.6% for the data sets with 467 features, while Hierarchical can obtain the best accuracy at 77.9% for the data set with 1768 features. Figure 1 illustrates the average of accuracy values obtained from all six data sets. The experimental results have shown that K-Means can achieve the best average accuracy at 71.73% for clustering Thai restaurant reviews. Figure 2 illustrates the running time of each clustering algorithm applied to six data sets. From Figure 2, it can be seen that the running time of all algorithms are reduced when the number of features are reduced.

Table I. Comparative results of K-Means, Hierarchical, FCM and SOM

MRF Feature Selection	Unsupervised Learning Method							
	K-Means		Hierarchical		FCM		SOM	
Number of features	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*
103	71.7	0.116	63.0	0.048	58.3	0.006	59.9	0.7
135	70.3	0.116	30.7	0.056	59.5	0.009	59.2	0.695
183	74.6	0.139	46.6	0.059	62.3	0.023	58.4	0.719
261	75.5	0.139	66.6	0.073	57.2	0.012	59.2	1.236
467	74.6	0.153	33.2	0.09	61.1	0.012	58.6	1.866
1768	69.3	0.889	77.9	0.31	59.2	0.039	59.0	5.55
Average	**71.73**	**0.2586**	**53**	**0.106**	**59.6**	**0.0168**	**59.22**	**1.794**

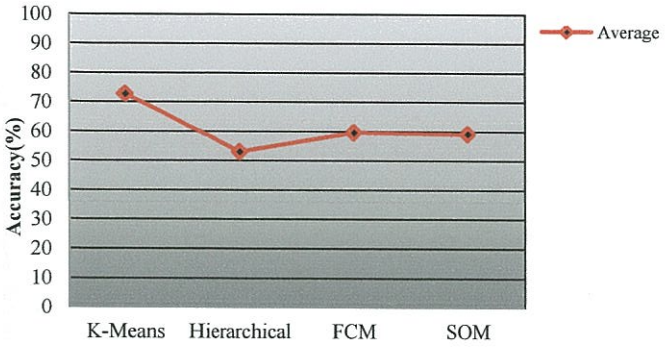


Fig. 1. Comparative averaged accuracies of all clustering methods.

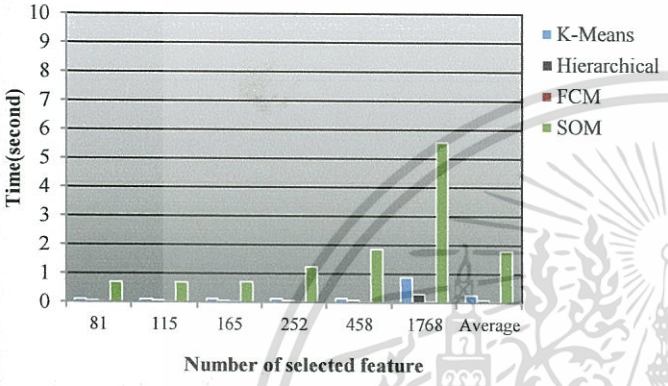


Fig. 2. Comparative running times of all clustering methods after applying MRF feature selection.

V. CONCLUSION

In this paper, we propose the clustering of customer opinions for Thai restaurants by using unsupervised learning algorithm. In addition, the proposed method apply MRF feature selection technique for selecting relevant features. This can effectively reduce the number of features and computational times. Then, K-Means is adopted for clustering the reviews of Thai restaurants into positive and negative groups. The experimental results showed that K-Means clustering is compatible with MRF feature selection since it can achieve the best performance in the clustering. Furthermore, the proposed opinion mining method can be further improved and deployed to other clustering problems in business, especially tourism industry. This proposed method can be applied to determine business strategy and to improve products and services according to requirements of customers.

VI. REFERENCES

[1] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multiclass Classification with Applications to High-Dimensional Data," Pattern Analysis and Machine Intelligence, Vol. 33, pp. 1217–1233, IEEE, June . 2011.

[2] Chiu, S., "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, Spet. 1994.

[3] P. Jiang, C. Zhang, H. Fu, Z. Niu1 and Q. Yang, "An Approach Based on Tree Kernels for Opinion Mining of Online Product Reviews", Data Mining, pp. 256–265, IEEE, Dec. 2010.

[4] X. Yu, Y. Liu, J. X. Huang and A. An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", Knowledge and Data Engineering, Vol. 24, pp. 720–734, IEEE, April. 2012.

[5] Amir A. Sheibani, "Opinion Mining and Opinion Spam", Telecommunications, pp. 1109–1113, IEEE, Nov. 2012.

[6] K. Jędrzejewski and M. Morzy, "Opinion Mining and Social Networks: a Promising Match", Advances in Social Networks Analysis and Mining, pp. 599 -604, IEEE, July. 2011.

[7] R. Xu, C. Kit, "CORSE-FINE OPINION MINING", Machine Learning and Cybernetics, pp. 3469–3474, IEEE, July 2009.

[8] T. Hitendra Sarma, P. Viswanath and B. Eswara Reddy, "A Fast Approximate Kernel k-means Clustering Method For Large Data sets", Recent Advances in Intelligent Computational Systems, pp. 545-550, IEEE, Sept. 2011.

[9] S. N. Sulaiman and N. A. Mat Isa, "Adaptive Fuzzy-K-means Clustering Algorithm for Image Segmentation", Consumer Electronics, Vol. 56, pp. 2661-2668, IEEE, Nov. 2010.

[10] L. Wei, W. Zeng and H. Wang, "K-means Clustering with Manifold", Fuzzy Systems and Knowledge Discovery, Vol. 5, pp. 2095-2099, IEEE, Aug. 2010.

[11] S. Na and L. Xumin, "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm", Intelligent Information Technology and Security Informatics, pp. 63-67, IEEE, April 2010.

[12] Timothy C. Havens, James C. Bezdek, C. Leckie, Lawrence O. Hall, and M. Palaniswami, "Fuzzy c-Means for Very Large Data", Fuzzy Systems, Vol. 20, pp. 1130 - 1146, IEEE, May 2012.

[13] B. Mohamed, T. Ahmed, H. Lassad and C. Abdelkader, "New Allied Fuzzy C-Means algorithm for Takagi-Sugeno Fuzzy model Identification", Electrical Engineering and Software Applications, pp. 1-7, IEEE, March 2013.

[14] T. Kohonen, "The Self-organizing Map", Proceedings of the IEEE. Vol. 78, pp. 1464-1480, IEEE, Sep. 1990.

[15] J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map", Neural Networks, IEEE, Vol. 110, pp. 586 - 600, May 2000.

[16] H. Chen and D. Zimbra, "AI and Opinion Mining", Intelligent Systems, IEEE , Vol 25 , pp. 74 – 80, June 2010.

ประวัติผู้เขียน

ชื่อ – สกุล	นายนิพัทธ์ คล้ายโพธิ์
วัน เดือน ปีเกิด	12 กันยายน 2532
ที่อยู่	37 หมู่ 6 ตำบลบ้านกร่าง อำเภอศรีประจันต์
	สุพรรณบุรี 72140
ประวัติการศึกษา	2554 จบการศึกษาปริญญาบริหารธุรกิจบัณฑิต
	สาขาวิชาระบบสารสนเทศทางคอมพิวเตอร์-คอมพิวเตอร์ธุรกิจ
	เกรดเฉลี่ย 3.31 คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี
ผลงานทางวิชาการ	1. Opinion Mining for Thai Restaurant Reviews using
	Neural Networks and mRMR Feature Selection
	2. Opinion Mining for Thai Restaurant Reviews using
	K-Means Clustering and MRF Feature Selection