การเปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล

วิทยานิพนธ์

ของ

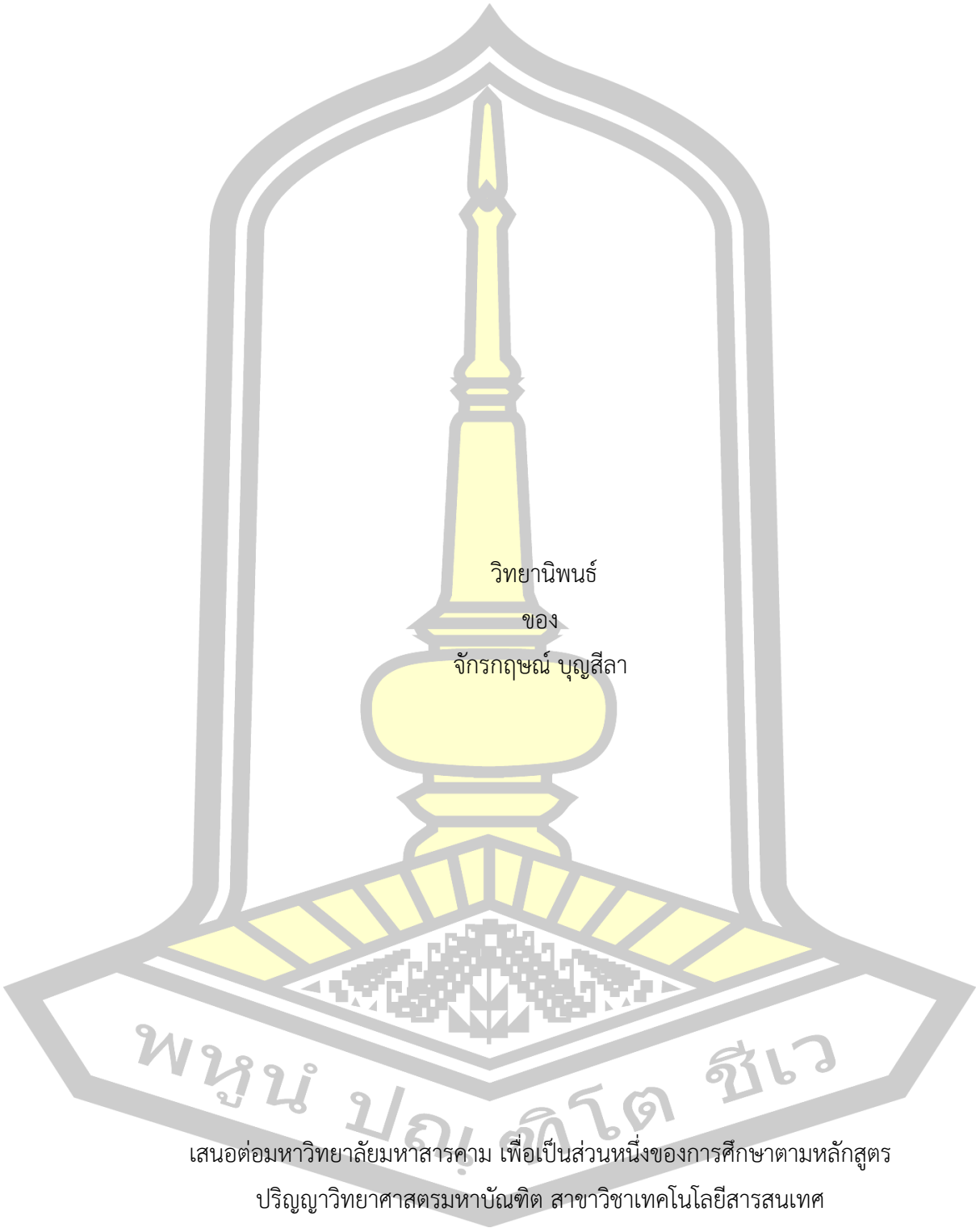
จักรกฤษณ์ บุญสีลา

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

สิงหาคม 2562

การเปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล



วิทยานิพนธ์

ของ

จักรกฤษณ์ บุญสีลา

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

สิงหาคม 2562

A Comparison of Classification of Subject Web board using Data Mining Techniques

Jakkrid Boonseela

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

August 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายจักรกฤษณ์ บุญสีลา แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรปภา อารีราษฎร์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จิรัฏฐา ภูบุญอบ)

.....กรรมการ
(ผศ. ดร. แกมกาญจน์ สมประเสริฐศรี)

.....กรรมการ
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(ผศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล
ผู้วิจัย จักรกฤษณ์ บุญสีลา
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จิรัฏฐา ภูบุญอบ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอผลการเปรียบเทียบแบบจำลอง 3 แบบจำลอง คือ 1) นาอีฟเบย์ (Naive Bayes) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และ 3) โครงข่ายประสาทเทียม (Neural Network) ซึ่งแบบจำลองที่มีประสิทธิภาพสูงที่สุดจะถูกนำมาใช้เพื่อพัฒนาระบบจำแนกหัวข้อกระดานข่าวบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคามอัตโนมัติ ข้อมูลที่ใช้ในการวิจัยนี้ เป็นข้อมูลที่เกี่ยวกับกระดานข่าวและ Facebook บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม ตั้งแต่ปี 2552-2561 จำนวน 1,102 Record การวิจัยได้แบ่งหมวดหมู่คำถามออกเป็น 4 กลุ่ม ได้แก่ การรับเข้า (Admission) งานระบบสารสนเทศ (Information) มาตรฐานบทนิพนธ์ (Thesis Standard) และงานมาตรการสอบ (Exam Standard) จากการทดลองพบว่าแบบจำลองที่ให้ประสิทธิภาพดีที่สุด คือซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ร้อยละ 83.33 อัลกอริทึม อัลกอริทึม นาอีฟเบย์ (Naive Bayes) ร้อยละ 80.33 และโครงข่ายประสาทเทียม (Neural Network) ร้อยละ 73.53 ตามลำดับ

คำสำคัญ : กระดานข่าว, อีฟเบย์, ซัพพอร์ตเวกเตอร์แมชชีน, โครงข่ายประสาทเทียม

TITLE A Comparison of Classification of Subject Web board using Data Mining Techniques

AUTHOR Jakkrid Boonseela

ADVISORS Assistant Professor Jiratta Phuboon-ob , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

This research presents the comparison of three models: 1) naive bayes 2) support vector machine 3) neural network We compares the most efficient models to be used to automatic classification of subject web board of graduate school Mahasarakham University. The data used in this research were the bulletin board and facebook page graduate school Mahasarakham University. from 2009 to 2017 with 1,102 records. The research was divided into 4 groups: admission, information system, thesis standard and exam standard. From the experiment, the best simulation model was support vector machine with 83.33%, naive bayes 80.33% and neural network 73.53%, respectively.

Keyword : Webboard, Naive Bayes, Support Vector Machine, Neural Network

กิตติกรรมประกาศ

งานวิจัยนี้ สำเร็จลุล่วงได้ด้วยความอนุเคราะห์อย่างสูงจาก ผศ.ดร.จิรัฏฐา ภูบุญอบ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ ปรึกษา รวมทั้งเสนอแนะแนวทางในการดำเนินงานวิจัยด้วยความเอาใจใส่อย่างดียิ่ง ขอขอบคุณบุคลากรบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคามที่ได้ให้ความช่วยเหลือในการจัดหาข้อมูลที่เกี่ยวข้อง ตลอดจนคำแนะนำต่าง ๆ ขอขอบคุณมหาวิทยาลัยมหาสารคามที่เอื้อเฟื้อสถานที่ในการจัดทำวิจัย ตลอดจนให้โอกาสทางการศึกษา ซึ่งเป็นประโยชน์ต่อผู้วิจัยอย่างยิ่ง

จักรกฤษณ์ บุญสีลา

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพประกอบ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

หน่วยงานไม่ว่าจะภาครัฐ หรือเอกชน การพัฒนาเว็บไซต์ พร้อมทั้งกระดานข่าว ถึงเป็นช่องทางหนึ่งในการแจ้งข้อมูลข่าวสารตามความต้องการของผู้ใช้ พร้อมทั้งเป็นแหล่งช่วยแก้ปัญหาและชี้แจงรายละเอียดต่างๆ ได้เป็นอย่างดี และในปัจจุบันความนิยมของการใช้สมาร์ทโฟนคงปฏิเสธไม่ว่าเข้ามามีบทบาทในทุกด้าน และสูงขนาดไหน การพัฒนาระบบส่วนของเว็บไซต์ ด้านกระดานข่าว (Webboard) ควบคู่กับช่องทางการสื่อสารโทรศัพท์มือถือสมาร์ทโฟนอีกด้านหนึ่ง ซึ่งถือเป็นสื่อสารธารณะอินเตอร์เน็ตได้อย่างอิสระ เสรีผ่านเครือข่ายโทรศัพท์มือถือต่าง ๆ ด้วยระบบอินเตอร์เน็ตความเร็วสูงอย่าง 3G หรือเทคโนโลยีใหม่อย่าง 4G

และปัจจุบันปัญหาการติดต่อสื่อสารระหว่างบัณฑิตวิทยาลัยกับนิสิตหรือคณาจารย์บุคลากร ในมหาวิทยาลัย ตลอดจนให้บริการแก่ผู้ปกครองนิสิต และหน่วยงานต่าง ๆ ทั้งภายในและภายนอกมหาวิทยาลัย ได้มีช่องทางการติดต่อสื่อสารหลายช่องทาง เช่น โทรศัพท์ อีเมล เบอร์โทร เฟสบุ๊ค และกระดานข่าว ณ ปัจจุบันได้มีบุคคลหรือนิสิตมีการสอบถามผ่านช่องทางกระดานข่าวบัณฑิตวิทยาลัย ผู้วิจัยได้ศึกษาเกี่ยวกับกระดานข่าวของบัณฑิตวิทยาลัย ซึ่งเปิดให้ผู้ที่ตั้งคำถามสามารถเลือกหมวดหมู่ของคำถามเองและในการเลือกนั้นอาจทำให้เลือกผิดหมวดหมู่ ก่อให้เกิดปัญหาบุคลากรอ่านข้อคำถามแล้วตอบประเด็นคำถามไม่ตรงประเด็น ก่อให้เกิดผลเสียต่อหน่วยงาน

การทำเหมืองข้อความ (Data Mining) [1] เป็นอีกหนึ่งวิธีในการวิเคราะห์ข้อมูลเพื่อค้นหาความรู้ และค้นหารูปแบบความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ในชุดข้อมูล เพื่ออยู่ในรูปฐานความรู้ (Knowledge Base) เพื่อนำไปใช้ประโยชน์ต่อไป เทคนิคการทำเหมืองข้อมูลมีหลายเทคนิค ได้แก่ การวิเคราะห์ทางด้านสถิติ (Statistical analysis) การจัดกลุ่ม (Clustering) การพยากรณ์ (Prediction) และการจำแนกประเภท (Classification) ในปัจจุบันการทำเหมืองข้อมูล ไปประยุกต์ใช้กับงานด้านต่าง ๆ ผู้วิจัยจริงมีแนวความคิดที่จะพัฒนาระบบการจำแนกหัวข้อกระดานข่าวอัตโนมัติของบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม โดยวิธีเหมืองข้อมูลเพื่อวิเคราะห์หัวข้อกระดานข่าวกระดานข่าวของบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคามเพื่อมาแก้ไขปัญหาดังกล่าว

งานวิจัยนี้ได้นำเสนอผลเปรียบเทียบประสิทธิภาพข้อมูลระหว่างเทคนิคจำแนกข้อมูล เทคนิคนาอีฟเบย์ (Naive Bayes) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และโครงข่ายประสาทเทียม (Neural Network) ซึ่งได้นำแบบจำลองที่จำแนกข้อมูลได้ดีที่สุดมาพัฒนาระบบการจำแนกหัวข้อกระดานข่าวกระดานข่าวของบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคามอัตโนมัติ

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล

1.3 ความสำคัญของการวิจัย

เนื่องจากบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม เป็นหน่วยงานกลางที่ให้การสนับสนุนการจัดการเรียนการสอนในระดับบัณฑิตศึกษาในระดับปริญญาโทและปริญญาเอก โดยมีบทบาทหน้าที่วิทยาลัย เป็นหน่วยงานระดับคณะ มีภารกิจในการบริหารและจัดการเรียนการสอนระดับบัณฑิตศึกษา ในบทบาทของการเป็นผู้ประสานงานและสนับสนุนการจัดการเรียนการสอน กำกับและควบคุมมาตรฐานการศึกษาระดับบัณฑิตศึกษาและบริหารการจัดการข้อมูลหลักสูตรของคณะ และหลักสูตรสาขาวิชาร่วมให้มีประสิทธิภาพและประสิทธิผล ดั้งนั้นการติดต่อสื่อสารระหว่างบัณฑิตวิทยาลัยกับนิสิตหรือคณาจารย์ บุคลากร ในมหาวิทยาลัย ตลอดจนให้บริการแก่ผู้ปกครองนิสิต และหน่วยงานต่าง ๆ ทั้งภายในและภายนอกมหาวิทยาลัย เพื่ออำนวยความสะดวกบัณฑิตวิทยาลัยจึงเพิ่มช่องทางการติดต่อผ่านทางเว็บไซต์ในลักษณะของกระดานข่าวและFacebook ซึ่งผู้ใช้งานหรือผู้ติดต่อสามารถตั้งคำถามเพื่อที่จะสอบถามข้อมูลเกี่ยวกับ การรับเข้า (Admission) งานระบบสารสนเทศ (Information) มาตรฐานบทนิพนธ์ (Thesis Standard) และงานมาตรการสอบ (Exam Standard) เกี่ยวกับมหาวิทยาลัยมหาสารคามได้

งานวิจัยนี้จึงได้ทำการศึกษาเกี่ยวกับการจำแนกกลุ่มคำถามบนกระดานข่าวแบบอัตโนมัติเพื่อที่ผู้ตอบคำถามสามารถที่จะตอบคำถามได้ตรงหมวดหมู่คำถามที่เกี่ยวข้องกับการให้บริการของหน่วยงานตนเอง ซึ่งจะเป็นการลดปัญหาการโทรมาถามเจ้าหน้าที่ตลอดจนแบ่งเบาการตอบปัญหาต่างๆในกระดานข่าวและ Facebook และงานวิจัยนี้ยังสามารถพัฒนาเพื่อที่จะศึกษาการใช้งานกระดานข่าวในด้านอื่น ๆได้อีกด้วย

1.4 ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในการวิจัยนี้ เป็นข้อมูลที่เกี่ยวกับกระดานข่าวและFacebookบัณฑิตวิทยาลัยมหาวิทยาลัยมหาสารคาม ตั้งแต่ปี 2552-เมษายน 2561 จำนวน 1,102 Record การวิจัยได้แบ่งหมวดหมู่คำถามออกเป็น 4 กลุ่ม ได้แก่ การรับเข้า (Admission) งานระบบสารสนเทศ (Information) มาตรฐานบทนิพนธ์ (Thesis Standard) และงานมาตรการสอบ (Exam Standard)

2. เปรียบเทียบประสิทธิภาพการจำแนกกลุ่มคำถามจาก 3 เทคนิควิธี ได้แก่ เทคนิคอีฟเบย์ (Naive Bayes)เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และโครงข่ายประสาทเทียม (Neural Network)

1.5 ผลที่คาดว่าจะได้รับจากงานวิจัยครั้งนี้

ได้ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ข้อมูลระหว่างเทคนิคอีฟเบย์ (Naive Bayes)เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และโครงข่ายประสาทเทียม (Neural Network)

1.6 นิยามศัพท์เฉพาะ

กระดานข่าว หมายถึง ข้อความที่ได้มาจากผู้ตั้งคำถามบนกระดานข่าว ลักษณะของการถาม-ตอบประเด็นปัญหาหรือข้อสงสัยต่างๆ ในรูปแบบข้อความ ผ่านเว็บไซต์ โดยมีเจ้าหน้าที่เป็นผู้ให้คำตอบข้อคำถามตามประเด็นต่างๆ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีของการทำเหมืองข้อความและการจำแนกข้อความ เพื่อจำแนกคำถามบนกระดานข่าว อัตโนมัติ ซึ่งผู้วิจัยได้ศึกษาเกี่ยวกับทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้องสามารถนำมาประยุกต์ใช้งานกับงานวิจัยได้ โดยมีรายละเอียด ดังนี้

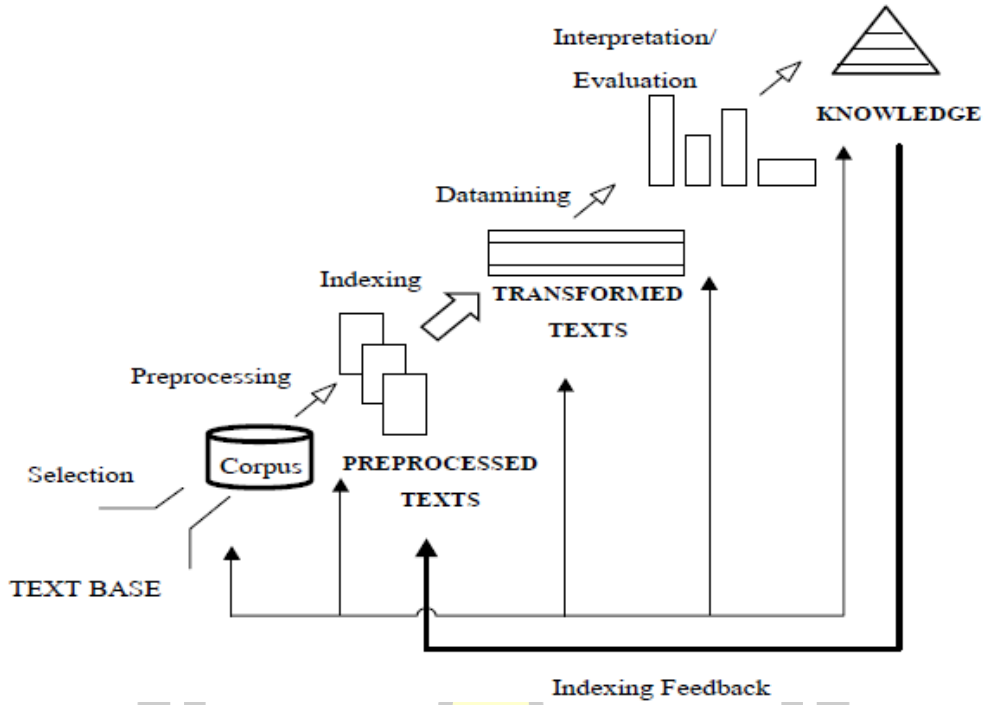
2.1 เหมืองข้อความ

การทำเหมืองข้อความ (Text Mining) เป็นกระบวนการในการค้นหาความรู้ใหม่หรือข้อเท็จจริงที่แฝงอยู่ในชุดข้อความ หรืออาจกล่าวได้ว่าเป็นกระบวนการในการวิเคราะห์หาความหมายที่อยู่ในข้อความ การทำเหมืองข้อความเน้นไปที่ข้อมูลประเภทไม่มีโครงสร้าง (Unstructured Data) เช่น ข้อความในอีเมล์ ข้อความบนเว็บบอร์ด หรือข้อมูลประเภทกึ่งโครงสร้าง (Semi-Structured Data) เช่น ข้อความในรูปแบบ XML หรือข้อความในรูปแบบ HTML [1]

การทำเหมืองข้อความเป็นเทคนิคในการทำเหมืองข้อมูล (Data Mining) สาขาหนึ่ง ซึ่งเป็นการค้นพบองค์ความรู้จากฐานข้อมูล (Knowledge Discovery in Database - KDD) และเป็นการค้นพบข้อมูลซึ่งข้อมูลนั้นยังไม่ทราบมาก่อน [2] โดยการสกัดแบบอัตโนมัติจากแหล่งที่มาของข้อมูลที่แตกต่างกัน สาเหตุที่เหมืองข้อความได้รับความนิยมอย่างแพร่หลาย เนื่องจากปริมาณข้อมูลที่อยู่ในลักษณะไม่มีโครงสร้าง (Unstructured) หรือกึ่งโครงสร้าง (Semi-Structured) ที่อยู่ในองค์กรหรือหน่วยงานต่าง ๆ มีปริมาณมาก เทคนิคเหมืองข้อความจึงเป็นเครื่องมือที่สามารถจัดการข้อมูลเหล่านั้นได้เป็นอย่างดี [3]

2.1.1 กระบวนการทำเหมืองข้อความ

กระบวนการทำเหมืองข้อความ (Text Mining Process) มีกระบวนการคล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล ทั้งนี้หากผลจากการวิเคราะห์ในแต่ละขั้นตอนมีความถูกต้องหรือความน่าเชื่อถือต่ำเกินไป จะต้องกลับไปขั้นตอนที่ต่ำกว่า [4] หรือทำการเลือกข้อมูลมาใหม่เพื่อทำให้กระบวนการทำเหมืองข้อความมีคุณภาพ โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ ดังภาพประกอบ 1 กระบวนการทำเหมืองข้อความ มีกระบวนการทำงานดังนี้



ภาพประกอบ 1 กระบวนการทำเหมืองข้อความ [5]

2.1.1.1 การเลือกข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนามาใช้ในการทำเหมืองข้อความ รวมถึงการนาข้อมูลที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการทำการศึกษา

2.1.1.2 การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนามาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนาข้อมูลที่ไม่ถูกต้องออก หรือเป็นขั้นตอนที่อาจต้องแก้ไขข้อมูลก่อนนาไปใช้งาน

2.1.1.3 การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป เช่น การตัดบางคอลัมน์ที่ไม่จาเป็นออก

2.1.1.4 การทำเหมืองข้อมูล (Data mining) เป็นขั้นตอนประมวลผล โดยใช้อัลกอริทึมต่าง ๆ เพื่อแก้ไขปัญหาหรือหารูปแบบของข้อมูล (Patten Search) การจัดหมวดหมู่ (Classification) การค้นหากฎความสัมพันธ์ (Association Rule Discovery) การแบ่งกลุ่มข้อมูล (Data Clustering) และการสร้างจินตทัศน์ (Visualization) โดยขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูลนั้น ๆ

2.1.1.5 การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความ และการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนาเสนอผลการวิเคราะห์ในรูปแบบที่สามารถเข้าใจได้ง่าย

2.1.2 การทำเหมืองข้อความเพื่อให้ได้มาซึ่งองค์ความรู้

การทำเหมืองข้อความจะคล้ายกับการทำเหมืองข้อมูลยกเว้นเครื่องมือในการสร้างเหมืองข้อมูล ซึ่งได้รับการออกแบบสำหรับโครงสร้างข้อมูล แต่การทำเหมืองข้อความสามารถทางานร่วมกับข้อมูลที่ไม่มีโครงสร้างหรือข้อมูลกึ่งโครงสร้างได้ เช่น Email, Full Text Document, HTML file เป็นต้น รูปแบบขององค์ความรู้ที่ได้มาจากการทำเหมืองข้อความสามารถทำได้หลายรูปแบบ [6, 7] ดังนี้

2.1.2.1 การสกัดสารสนเทศ (Information Extraction) เป็นกระบวนการสกัดนิพจน์(Entity) ที่อยู่ในความสนใจและเกี่ยวข้องกับหัวข้อที่กาลังพิจารณา ซึ่งในการประมวลผลภาษาธรรมชาตินั้น ได้มีการศึกษาและวิจัยเทคนิคการจดจานิพจน์ระบุนาม (Named Entity Recognition – NER)นิพจน์ระบุนาม ได้แก่ ชื่อบุคคล ชื่อองค์กร ชื่อสถานที่ ปริมาณ/จำนวน วัน/เวลา ซึ่งเหมาะกับการวิเคราะห์เหตุการณ์จากข่าว อย่างไรก็ตามการสกัดสารสนเทศสามารถประยุกต์ให้เข้ากับโดเมนหรือหัวข้อเฉพาะที่สนใจได้ เช่น การสกัดชื่อยีน (Gene expression) จากบทความวิชาการทางชีวการแพทย์เป็นต้น

2.1.2.2 การตรวจจับและติดตามหัวข้อ (Topic Detection and Tracking) เป็นกระบวนการสำหรับตรวจจับหัวข้อที่สนใจจากสารสนเทศที่เกิดขึ้นและไหลเข้ามาอย่างต่อเนื่อง เช่น บทความข่าว บทความตีพิมพ์ และ ข้อความสั้นบนเว็บไซต์ทวิตเตอร์ เป็นต้น เทคนิคนี้เหมาะกับการตรวจจับและติดตามหัวข้ออย่างอัตโนมัติให้กับผู้ใช้ตามหัวข้อที่สนใจ ทำให้ผู้ใช้ไม่ต้องเสียเวลาเข้าไปในเว็บไซต์เพื่อตรวจสอบเอง

2.1.2.3 การสรุปเอกสารข้อความ (Text Summarization) เป็นการลดความซับซ้อนและลดขนาดของเอกสารข้อความ โดยการเลือกเฉพาะสาระที่สำคัญของเอกสาร และตัดเนื้อหาที่ไม่สำคัญออก โดยความหมายของเอกสารข้อความยังคงเดิม

2.1.2.4 การจำแนกประเภทเอกสารข้อความ (Text Classification) เป็นการจำแนกกลุ่มเอกสารข้อความตามประเภทที่ได้กำหนดไว้ โดยการใช้ชุดข้อมูลตัวอย่างของเอกสารข้อความเพื่อใช้ในการฝึกฝน (Training Set) และเพื่อใช้ในการทดสอบ (Test Set) อัลกอริทึมในการแบ่งประเภทเอกสาร เช่น Supervised Learning Neural Networks และ C4.5 Decision Tree

2.1.2.5 การแบ่งกลุ่มเอกสารข้อความ (Text Clustering) เป็นการจัดแบ่งเอกสารข้อความออกเป็นกลุ่ม โดยใช้การวัดความคล้ายคลึงและความแตกต่างของคุณลักษณะของเอกสารข้อความ ซึ่งจะไม่มีข้อมูลกลุ่มตัวอย่างในการฝึกฝนหรือการทดสอบ ข้อมูลเอกสารจะถูกแปลงให้เป็นชุด

ข้อมูลตัวเลขโดยวิธีการ DFxIDF (Vector Space Model) อัลกอริทึมในการแบ่งกลุ่มเอกสาร เช่น KMean Unsupervised Learning Neural Networks และ Hierarchical Clustering

2.1.2.6 การถามตอบ (Question and Answering) เป็นเทคนิคการค้นคืน

สารสนเทศรูปแบบหนึ่ง โดยผู้ใช้สามารถป้อนคำถามเป็นคิวรีให้กับระบบ และผลลัพธ์ที่ได้คือ คำตอบที่ตรงกับคำถามที่ระบุ เช่น คำถาม "เมืองหลวงของประเทศไทยคืออะไร" คำตอบคือ "กรุงเทพฯ" เป็นต้นเทคนิคนี้สามารถหาไปพัฒนาระบบถามตอบ (Question Answering System) โดยมีความแตกต่างจากระบบค้นคืนสารสนเทศ (Information Retrieval System) และเสิร์ชเอ็นจิน (Search Engine) ซึ่งผู้ใช้ระบุคิวรีเป็นคำสำคัญและผลลัพธ์เป็นรายการเอกสารหรือหน้าเว็บที่มีคำสำคัญที่ผู้ใช้ระบุปรากฏอยู่

2.1.3 ประโยชน์ของการทำเหมืองข้อความ

ปัจจุบันเหมืองข้อความได้ถูกนามาใช้ประโยชน์ทางด้านการค้นหาข้อมูลที่ซ่อนอยู่ในข้อความเอกสารจำนวนมหาศาล ทำให้ค้นพบข้อมูลที่มีประโยชน์ช่วยในการสนับสนุนด้านต่าง ๆ ตัวอย่างการใช้ประโยชน์ในการทำเหมืองข้อความ [8, 9] เช่น

2.1.3.1 ด้านธุรกิจ มีการทำเหมืองข้อความเพื่อการวิเคราะห์ข้อมูลการใช้บริการทางธุรกิจของลูกค้าเพื่อนาผลการวิเคราะห์มาสนับสนุนการวางแผนการตลาดหรือการลงทุน

2.1.3.2 ด้านการแพทย์ เพื่อทำการวิเคราะห์ข้อมูลผู้ป่วย ข้อมูลโภชนาการ ข้อมูลยา หรือการค้นหาความสัมพันธ์ของโรคต่อการดูแลรักษา หรือยารักษาโรค เป็นต้น

2.1.3.3 ด้านการพยากรณ์ โดยการวิเคราะห์ข้อมูลเพื่อศึกษาแนวโน้มของเหตุการณ์หรือการค้นหาความสัมพันธ์ของข้อมูลด้านต่าง ๆ

การประมวลผลข้อความ (Text Preprocessing) เป็นการจำแนกเอกสารภาษาไทยอัตโนมัติมีขั้นตอนหลักในการทำงาน คือ ขั้นตอนแรกจะทำการสกัดคุณลักษณะด้วยการตัดคำเพื่อให้ได้คุณลักษณะจากเอกสารออกมา จากนั้นทำการกาจัดคำหยุดและทำรากศัพท์จากฐานข้อมูลภาษาไทยที่กำหนดขึ้น หลังจากนั้นทำการให้ค่าน้ำหนักดัชนีของคำในเอกสาร (Term Weighting) แล้วทำการลดขนาดคุณลักษณะเพื่อมิติของเอกสารลดลง จากนั้นให้ทำการเรียนรู้แบบมีผู้สอน (Supervised Learning) [9]

2.3.1 การตัดคำ (Word Segmentation)

การประมวลผลการจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพ พบว่าเกิดปัญหาในการตัดคำภาษาไทย ซึ่งลักษณะภาษาไทยมีการเขียนติดต่อกันเป็นสายอักขระโดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำหรือสัญลักษณ์ที่สามารถบ่งบอกถึงขอบเขตของคำที่เรียงต่อเนื่องกันทั้งประโยคเหมือนกับภาษาอังกฤษที่ใช้ช่องว่าง (Space) คั่นระหว่างขอบเขตของคำ [10, 11] ดังนี้

2.3.1.1 หลักการตัดคำโดยใช้กฎ (Rule-Based Approach) การตัดคำโดยใช้กฎเป็นการตัดคำโดยการตรวจสอบจากฎเกณฑ์ทางอักขระวิธีที่สร้างขึ้นมาโดยอาศัยหลักไวยากรณ์

ภาษาไทยซึ่งเริ่มจากการพัฒนาการตัดพยางค์ขึ้นมาก่อน เนื่องจากพยางค์มีรูปแบบและกฎเกณฑ์ที่แน่นอนมากกว่าคำ จากนั้นจึงนามาเป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้มีข้อจากัดในการทำงานคือ ผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดคำออกไปได้อีก ความถูกต้องของการตัดคำที่ได้จะมีค่าต่อเนื่องจากภาษาไทยมีกฎเกณฑ์ที่ไม่แน่นอนจำนวนมากแต่มีข้อดี คือ การทำงานของระบบมีความรวดเร็วและใช้ทรัพยากรในการประมวลผลน้อย

2.3.1.2 หลักการตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) การตัดคำโดยใช้พจนานุกรม เป็นการตัดคำโดยใช้การเปรียบเทียบคำกับคำที่จัดเก็บอยู่ในพจนานุกรมร่วมกับการใช้กฎในการตัดคำด้วย ซึ่งเป็นการแก้ปัญหาการตัดคำของการใช้กฎ เนื่องจากการใช้กฎเพียงอย่างเดียวไม่สามารถหาขอบเขตของคำได้ถูกต้องทั้งหมด โดยวิธีการนี้จะต้องมีการจัดเก็บคำทั้งหมดไว้ในพจนานุกรม แล้วนาข้อความที่ป้อนเข้ามาไปค้นหาและเปรียบเทียบสายอักขระกับคำในพจนานุกรมปัญหาที่พบในวิธีการตัดคำโดยใช้พจนานุกรม คือ ไม่สามารถจัดเก็บคำทั้งหมดลงในพจนานุกรมได้เนื่องจากมีคำใหม่ ๆ เกิดขึ้นอยู่ตลอดเวลา ดังนั้นความถูกต้องของวิธีนี้จึงขึ้นอยู่กับปริมาณของคำในพจนานุกรม และต้องสูญเสียพื้นที่ในการจัดเก็บคำตามปริมาณคำศัพท์ในพจนานุกรมด้วย วิธีการตัดคำโดยพจนานุกรม สามารถแบ่งออกเป็น 2 วิธี คือ

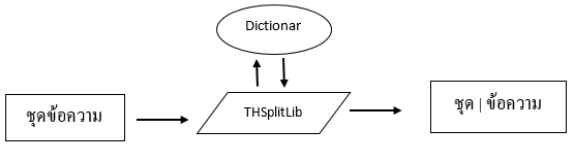
1) วิธีการเทียบคำที่ยาวที่สุด (Longest Matching) โดยวิธีการเทียบคำที่ยาวที่สุดนั้นจะทำการเปรียบเทียบคำที่ป้อนเข้ามากับคำที่อยู่ในพจนานุกรม ถ้าไม่พบคำที่สามารถเทียบคำที่อยู่ในพจนานุกรมได้ ระบบจะทำการลดความยาวของข้อความลงทีละตัวอักษรตามหลักทางอักขระวิธีจนสามารถที่จะเทียบคำกับคำในพจนานุกรมได้ ซึ่งวิธีการตัดคำนี้มีความถูกต้องมาก แต่มีข้อด้อย คือ การเทียบคำที่มีความยาวมากเกินไปตั้งแต่แรกทำให้การตัดคำที่มีตามมา มีความผิดพลาดได้เนื่องจากคำในภาษาไทยบางคำเป็นคำประสมที่เกิดจากหลายคำรวมกัน

2) การตัดคำให้ได้จำนวนคำและคำที่ไม่พบในพจนานุกรมมากที่สุด (Maximal Matching) เป็นวิธีที่ใช้ในการแก้ปัญหาจากการตัดคำด้วยวิธีการเทียบคำที่ยาวที่สุด ที่เกิดความผิดพลาดจากการเทียบกับคำที่มีความยาวมากเกินไปตั้งแต่ต้น วิธีการนี้จะใช้การตัดคำโดยวิธีการเทียบคำที่ยาวที่สุดก่อนจากนั้นจะทำการย้อนกลับ (Backtracking) เพื่อทำการตัดคำที่สามารถเป็นไปได้อีกครั้ง โดยใช้วิธีพิจารณาทางเลือกของการตัดคำที่เป็นไปได้ทั้งหมดก่อนที่จะเลือกการตัดคำและผลลัพธ์ของการตัดคำ
วิธีนี้จะเลือกจากการตัดคำที่ค่าต้นทุนน้อยที่สุดและได้คำที่ไม่พบในพจนานุกรมมากที่สุด

2.3.1.3 หลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) เป็นวิธีการตัดคำที่นาหลักทางสถิติหรือวิธีการกลไกการเรียนรู้เข้ามาใช้ในกระบวนการประมวลผลทางภาษา โดยใช้คลังข้อมูลทางภาษาที่ใช้แรงงานคนในการเตรียมข้อมูลมาเป็นฐานความรู้ที่จะใช้สำหรับการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลสามารถแบ่งได้ 2 วิธี คือ วิธีการตัดคำที่อาศัยค่าความน่าจะเป็น และ

วิธีการตัดคำที่อาศัยคุณลักษณะของคำโดยวิธีการตัดคำที่อาศัยค่าความน่าจะเป็นเป็นการตัดคำโดยใช้แบบจำลองไตรแกรมกากับหน้าที่ของคำ(Part-Of-Speech Trigram Model) ในการหารูปแบบของการตัดคำ และลาดับหมวดคำ (TagSequence) ที่เป็นไปได้มากที่สุด ซึ่งวิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและการกำกับหน้าที่ของคำเตรียมไว้ก่อนแล้ว วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ เป็นวิธีการที่แก้ไขข้อผิดพลาดของ
การตัดคำที่อาศัยค่าความน่าจะเป็นของการจัดประเภทของคำที่จะนามาเป็นแบบจาลองในการตัดคำ ซึ่งวิธีการนี้เป็นวิธีการแบบผสม (Hybrid Approach) โดยนาคุณลักษณะของคำที่มีความกากวมในการตัดคำมาช่วยเลือกการตัดคำที่ถูกต้อง หลังจากที่มีการตัดคำจากวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด โดยการสร้างแบบจาลองอาจจะใช้กลไกการเรียนรู้ เช่น ต้นไม้ตัดสินใจ เบย์เซียนเน็ตเวิร์ค วินโดว์ เป็นต้น

งานวิจัยนี้ใช้ โปรแกรม THSplitlib [12] ซึ่งเป็นโปรแกรม Open Source ในการตัดคำภาษาไทย ใช้หลักการตัดคำซึ่งดาเนินการโดยใช้พื้นฐานจากพจนานุกรม (Dictionary-Based Approach) ในการเปรียบเทียบการตัดคำกับคำที่จัดเก็บในพจนานุกรม ดัง ภาพประกอบ 2



ภาพประกอบ 2 การตัดคำ

จากภาพประกอบ 2 โดยอธิบายการทำงาน 3 ขั้นตอนดังนี้

ขั้นตอนที่ 1 เพิ่มคำถามใหม่ผ่านฟอร์มเพิ่มคำถาม

ขั้นตอนที่ 2 คำถามที่ป้อนเข้ามาจะถูกนำมาตัดคำโดยเทียบกับพจนานุกรม คำศัพท์ด้วยวิธีเทียบคำที่ยาวที่สุดที่พบในพจนานุกรม THSplitlib

ขั้นตอนที่ 3 เป็นการเปรียบเทียบคำที่ตัดแล้วกับคำสำคัญในฐานข้อมูล ว่าพบคำเหล่านั้นปรากฏในประโยคหรือไม่

2.1 เทคนิคอีฟเบย์ (Naive Bayes)

อีฟเบย์ (Naive Bayes)เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ อีฟเบย์ (Naive Bayes) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดีเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน การจำแนกประเภทเบย์อย่างง่ายสามารถนำไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดี ทำให้ผู้วิจัยเลือกวิธีการนี้มาใช้ในงานวิจัย เนื่องจากเป็นวิธีการจำแนกข้อมูลที่มีประสิทธิภาพและมีอัลกอรึทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่น ๆ [13, 14]

สมมติฐานของการจำแนกประเภทเบย์อย่างง่าย คือ เรากำหนดให้คุณสมบัติแต่ละตัวไม่ขึ้นต่อกันกับคุณสมบัติอื่น ๆ ซึ่งสามารถเขียนแทน $P(a_1, a_2, ...a_n \mid v_j)$ ด้วยผลคูณความน่าจะเป็น ดังสมการ

$$P(a_1, a_2, ...a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j)$$

โดยที่

$\prod$ คือ ผลคูณของค่า P(ai | vj) ทั้งหมด i = 1,2,3,...,n และ j = 1,2,3,...,n

การนาเทคนิควิธีการเรียนรู้เบย์อย่างง่ายไปใช้งาน มีวิธีการดังนี้

1) หาความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนาค่า 1 2 n j P(a , a , ..., a | v) จากสมการที่ 2-11 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้น ๆ คือ j P(v) ได้เท่ากับ *NB V*

2) นาค่าที่ได้มาเปรียบเทียบกับกลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้วิธีการจำแนกประเภทอย่างง่าย ดังสมการ

$$V_{NB} = \arg\max_{vieV} P(v_j) x \prod_{i=1}^{n} P(a_i \mid v_j)$$

ข้อดีของวิธีการเรียนรู้เบย์อย่างง่าย สามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior knowledge) เข้ามาช่วยในการเรียนรู้ได้ ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดี สามารถใช้ได้ดีกับข้อมูลที่มีจำนวนไม่มาก และข้อมูลที่ไม่ขึ้นต่อกัน [15]

2.3 เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวคเตอร์แมชชีนเป็นอัลกอลิทึมที่ใช้ในการแบ่งประเภทซึ่งได้รับการเสนอโดย Vapnik ในปี ค.ศ.1992 เป็นระบบการเรียนรู้ที่อาศัยหลักทฤษฎีการเรียนรู้ทางสถิติ (Statistical LearningTheory) พร้อมกับทฤษฎีการหาค่าที่เหมาะสม (Optimization Theory) ซัพพอร์ตเวกเตอร์แมชชีนถือเป็นโมเดล เชิงเส้น แต่สามารถนามาใช้แก้ไขปัญหาที่มีความซับซ้อนไม่เป็นเชิงเส้นได้โดยอาศัยการแปลงข้อมูลไปยังอีกปริภูมิหนึ่งที่มีจานวนมิติมากกว่าเดิมโดยอาศัยเคอร์เนลฟังก์ชัน (Kernel Function) ทาให้แก้ไขปัญหาได้ง่ายยิ่งขึ้น ซึ่งในการใช้ซัพพอร์ตเวคเตอร์แมชชีนอย่างมีประสิทธิภาพนั้นจาเป็นจะต้องเข้าใจถึงการทางานของซัพพอร์ตเวคเตอร์แมชชีน เช่น การตระเตรียมข้อมูลการเลือกใช้เคอร์เนล (Kernel) และสุดท้ายการเลือกค่าพารามิเตอร์ของซัพพอร์ตเวคเตอร์แมชชีนและเคอร์เนล หากไม่เข้าใจถึงกระบวนการเหล่านี้จะทาให้ประสิทธิภาพในการทางานของซัพพอร์ตเวคเตอร์แมชชีนลดลง

Support Vector Machines (SVM) ตัวแบบของ SVM มีความคล้ายคลึงกับเอร์เซฟตรอน ซึ่งเป็นข่ายงานประสาทเทียมแบบง่ายมีหน่วยเดียวที่จำลองลักษณะของเซลล์ประสาท ด้วยการใช้ Kernal Function ในสื่อตีพิมพ์เกี่ยวกับ SVM จะเรียกตัวแปรในการตัดสินใจว่าคุณสมบัติและตัวแปรที่เปลี่ยนแปลงใช้การกำหนดระนาบหลายมิติ เรียกว่าคุณลักษณะ (feature) ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า การคัดเลือกคุณลักษณะ (feature selection) จำนวนเซตของลักษณะที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่า เวกเตอร์ (vector) ดังนั้น จุดมุ่งหมายของตัวแบบ SVM คือการประโยชน์สูงสุดจากระนาบหลายมิติที่แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรเป้าหมายที่อยู่ข้างหนึ่งของระนาบและกรณีกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดนี้เรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vector)

SVM เป็นวิธีการที่สามารถนำมาใช้การจำแนกรูปแบบหรือกลุ่มของข้อมูลได้ โดยจะอาศัยระนาบมาใช้ในการแบ่งเขตของข้อมูลออกเป็นสองฝั่ง และ support vector machines นี้จะมีคุณลักษณะแบบ inner-product ระหว่างตัว support vector และ inpout vector

$$\varnothing(w,\xi) = \frac{1}{2} w^T w + C\sum_{i=1}^{n} \xi_i$$

จากสมการที่ (1) เป็นการแสดงเวกเตอร์ค่าน้ำหนักของ w โดยจะพยายามลดค่าในเทอมแรกของสมการที่ (1) ให้มีค่าน้อยที่สุด และค่า C เป็นค่าคงที่ที่ใช้ สำหรับกำหนดค่าความผิดพลาดในการแยกกลุ่มข้อมูลและค่า ξ_i หรือ slack variable ซึ่งจะเป็นการวัดค่า ความผิดพลาดที่คลาดเคลื่อนไปจากตำแหน่งที่เหมาะสม

$$\sum_{i-1}^{n} a_i d_i K(x, x_i) = 0$$

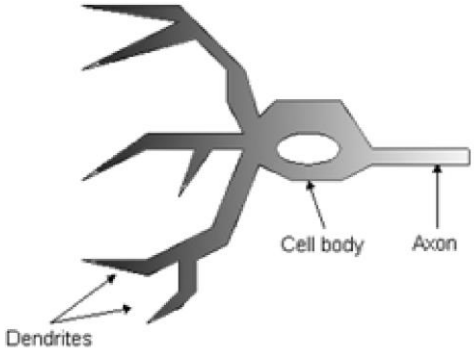
จากสมการที่ (2) แสดงค่า decision surface โดยที่ K(x,xi) เป็น Inner-Product Kernel และ a_i คือค่า Lagrange multipliers และ d_i คือค่า target output สำหรับ kernel ของ SVM ที่นิยมใช้กันคือแบบ polynomial เป็นการคำนวณหาเส้นแบ่งโดยใช้สมการเชิงเส้นที่มี degree มากว่าสองและแบบ RBF ซึ่งเป็นการคำนวณหาขอบเขตข้อมูลโดยอาศัยวิธีการแบบ Radial Basis เข้ามาช่วยในการคำนวณดังแสดงวในสมการที่ (3) และ (4) ตามลำดับ

$$K(x, x_i) = (x^T x_i + 1)^P$$

$$K(x, x_i) = \exp(-y \| x - x_i \|^2)$$

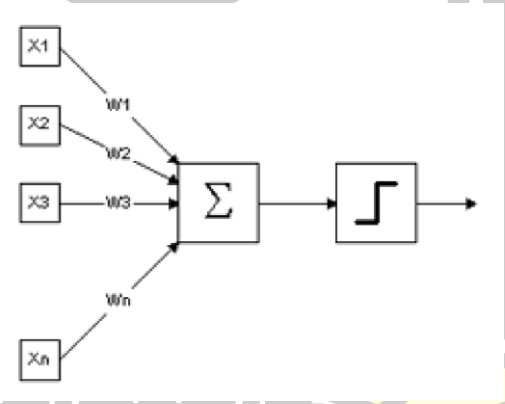
2.4 เทคนิคโครงข่ายประสาทเทียม (Neural Network)

โครงข่ายประสาทเทียม (Artificial neural network) หรือที่มักจะเรียกสั้น ๆ ว่า ข่ายงานประสาท (neural network หรือ neural net) คือโมเดลทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศด้วยการคำนวณ แบบคอนเนคชันนิสต์ (connectionist) เพื่อจำลองการทำงานของเครือขายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำแบบรูป(Pattern Recognition) และการอุปมาน ความรู้ เช่นเดียวกับความสามารถที่มีในสมองมนุษย์แนวคิดเริ่มต้นของเทคนิคนี้ ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท หรือ “นิวรอน” (neurons)และ จุดประสานประสาท (synapses)แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับ กระแสประสาท เรียกว่า "เดนไดรท์" (Dendrite) ซึ่งเป็น input และปลายในการส่งกระแสประสาทเรียกว่า "แอ คซอน" (Axon) ซึ่งเป็นเหมือน output ของเซลล์เซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมีเมื่อมีการกระตุ้นด้วย สิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่าน ทางแอคซอนของมัน



ภาพประกอบ 3 แสดง Model ของ Neuron ในสมองมนุษย์ [16]

โครงสร้าง นักวิจัยส่วนใหญ่ในปัจจุบันเห็นตรงกันว่าข่ายงานประสาทเทียมมีโครงสร้างแตกต่างจากข่ายงานใน สมอง แต่ก็ยังเหมือนสมอง ในแง่ที่ว่าข่ายงานประสาทเทียม คือการรวมกลุ่มแบบขนานของหน่วยประมวลผล ย่อยๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญาของข่ายงาน เมื่อพิจารณาขนาดแล้วสมองมี ขนาดใหญ่กว่าข่ายงานประสาทเทียมอย่างมาก รวมทั้งเซลล์ประสาทยังมีความซับซ้อนกว่าหน่วยย่อยของ 2 ข่ายงาน อย่างไรก็ดีหน้าที่สำคัญของสมอง เช่น การเรียนรู้ยังคงสามารถถูกจำลองขึ้นอย่างง่ายด้วยโครงข่าย ประสารทนี้



ภาพประกอบ 4 แสดง Model ของ Neuron ในคอมพิวเตอร์ [16]

ประเภทของโครงข่ายประสาทเทียม

Dave Anderson and George McNeill ได้เสนอแนวคิดเรื่องประเภทของโครงข่ายประสาทเทียมว่า แนวคิดของเซลล์ประสาทเทียมที่ต้องมีการเชื่อมต่อและมีการคำนวณค่าแอคติเวชั่นฟังก์ชั่น โดยมากจะเป็นตัวบ่งบอกโครงสร้างทางสถาปัตยกรรม ซึ่งจะมีกฎกำหนดวิธีการของโครงข่ายโดยแบ่งออกเป็นประเภทได้ 1 ประเภท ดังนี้

ตาราง 1 แสดงความแตกต่างระหว่างประเภทของโครงข่ายประสาทเทียม

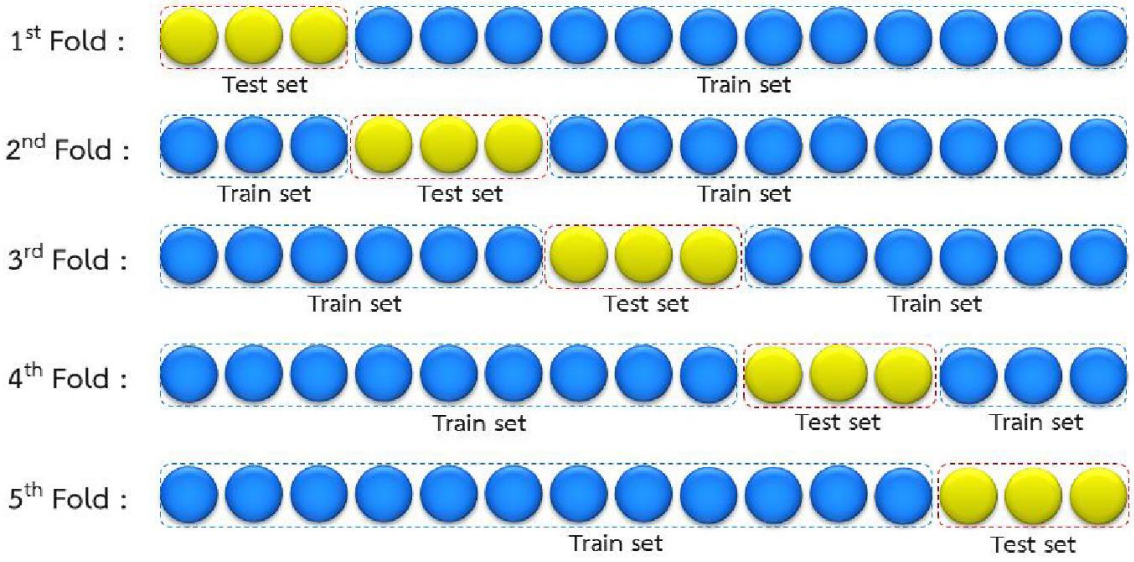
ประเภท	ชื่อโครงข่าย	ลักษณะการใช้งาน
การคาดเดา Prediction	- Back-propagation - Delta Bar Delta - Extended Delta Bar Delta - Directed Random Search - HigherOrder Neural Networks - Self-organizing map into Back-propagation	ใช้ค่าอินพุตเพื่อคาดเดาเอาต์พุต
การจัดหมวดหมู่ Classification	- Learning Vector Quantization - Counter-propagation - Probabilistic Neural Networks	ใช้ค่าอินพุตเพื่อกำหนดการจัดหมวดหมู่
การเชื่อมโยงข้อมูล Data Association	- Hopfield - Boltzmann Machine - Hamming Network - Bidirectional associative Memory	เหมือนกับ Classificationแต่มันจะจดจำข้อมูลที่มีerror ด้วย
กระบวนการสร้างความคิดData Conceptualization	- Adaptive Resonance Network - Self-Organizing Map	วิเคราะห์อินพุตเพื่อการจัดกลุ่ม
การกลั่นกรองข้อมูล Data Filtering	- Recirculation	ทำให้สัญญาณอินพุตเรียบสม่ำเสมอ

โดยโครงข่ายที่จะกล่าวถึงในงานวิจัยนี้ ขอกล่าวถึงเพียงแค่โครงข่ายที่ใช้สำหรับในงานวิจัยนี้เพียงเท่านั้น เทคนิคโครงข่ายประสาทเทียมแบบชั้นเดียว (Single Layer) โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer) ดังที่จะกล่าวต่อไป

2.5 การทดสอบประสิทธิภาพการทำงาน

2.5.1 การประเมินผลวิธี K-fold Cross Validation

ขั้นตอนการสร้างแบบจำลองเพื่อจำแนกกลุ่มข้อความด้วยเทคนิคเหมืองข้อความจำเป็นต้องมีการแบ่งชุดข้อมูลทดสอบ (Testing Data) เพื่อให้ได้ผลการวิจัยที่มีความถูกต้องแม่นยำ โดยใช้วิธีการประเมินผลวิธี K-fold Cross Validation ซึ่งเป็นวิธีที่ได้รับความนิยมอย่างแพร่หลายและเหมาะสมกับข้อมูลทดสอบ โดยแบ่งข้อมูลออกเป็นกลุ่มจำนวน k กลุ่ม (k-Fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไป classifier จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นๆ ที่เหลือเป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อยๆ จนครบ k กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยของค่าความถูกต้องในแต่ละกลุ่ม ว่า k กลุ่มไหนให้ผลลัพธ์ในการจำแนกหมวดหมู่ดีที่สุด วิธีการนี้ข้อมูลทุกตัวอย่างจะได้เป็นทั้งชุดทดสอบและชุดสอนดังภาพประกอบ 5 การแบ่งกลุ่มประเมินผลด้วยวิธี K-fold Cross Validation



ภาพประกอบ 5 การแบ่งกลุ่มประเมินผลด้วยวิธี K-fold Cross Validation

2.5.2 การประเมินประสิทธิภาพ

งานวิจัยนี้ ใช้การประเมินประสิทธิภาพการทำงานด้วยค่าความเที่ยง (Precision) ค่าความระลึก (Recall) และค่าความถูกต้อง (Accuracy) ผลการจำแนกหมวดหมู่ที่สามารถเกิดขึ้นได้สามารถเขียนเป็นตารางการณ์ (Contingency Table) ดังตาราง 2

ตาราง 2 ตารางการณ์ของการจำแนกหมวดหมู่

การจำแนกหมวดหมู่	อยู่ในหมวดหมู่ cj	ตัดสินโดยผู้เชี่ยวชาญ	
		ใช่	ไม่ใช่
ตัดสินโดยตัวจำแนกอัตโนมัติ	ใช่	TP (True Positive)	FP (False Positive)
	ไม่ใช่	FN (False Negative)	TN (True Negative)

TP คือ จำนวนเอกสารที่อยู่ในหมวดหมู่ C_j และตัวจำแนกอัตโนมัติทำนายว่าอยู่ในหมวดหมู่ C_j

FP คือ จำนวนเอกสารที่ไม่อยู่ในหมวดหมู่ C_j แต่ตัวจำแนกอัตโนมัติทำนายว่าอยู่ในหมวดหมู่ C_j

FN คือ จำนวนเอกสารที่อยู่ในหมวดหมู่ C_j แต่ตัวจำแนกอัตโนมัติทำนายว่าไม่อยู่ในหมวดหมู่

2.6 งานวิจัยที่เกี่ยวข้อง

Huan Huang และคณะ [17] ศึกษาเกี่ยวกับการจัดหมวดหมู่ข้อความสั้นแบบอัตโนมัติ โดยการทำให้คอมพิวเตอร์จัดการการเชื่อมโยงเนื้อหาของข้อความให้ตรงกับประเภทหมวดหมู่ข้อความโดยอัตโนมัติ ขั้นตอนการจัดหมวดหมู่ทั่วไป เช่น Support Vector Machine (SVM)หรือ k-Nearest Neighbor (k-NN) จะใช้คำในการจัดหมวดหมู่ คุณสมบัติจึงขึ้นอยู่กับความถี่ของคำที่เพิ่มสูงขึ้น คำหรือข้อความทั่วไปไม่สามารถเป็นตัวแทนของข้อความที่ดีได้ นอกจากนี้ขั้นตอนวิธีการจัดหมวดหมู่แบบเดิมไม่สามารถสกัดคุณลักษณะที่มีเงื่อนไขมากเกินไป ดังนั้นประสิทธิภาพการจัดหมวดหมู่จึงค่อนข้างต่ำ ยังเป็นปัญหาสำคัญในการจัดหมวดหมู่ข้อความขนาดใหญ่ คำบางคำในหัวข้อสามารถสะท้อนให้เห็นถึงรูปแบบและเป็นตัวแทนของงานวิจัยได้ การสกัดคำในหัวข้อจึงไม่เพียงแต่จะปรับปรุงความถูกต้องของการจัดหมวดหมู่ แต่ยังสามารถเพิ่มประสิทธิภาพ

ของการจัดหมวดหมู่ข้อความได้งานวิจัยนี้ยังเป็นแนวทางในการแยกคุณลักษณะของคำอ่านและคำพูดทั่วไปได้ จึงใช้ขั้นตอนวิธีBayesian Classification ในการจำแนกและจัดหมวดหมู่ข้อความ

สิงห์ทัย สุขสว่างโรจน์ [18] เพื่อสร้างระบบถาม-ตอบภาษาไทยเพื่อใช้ในการตัดสินใจของนักศึกษามหาวิทยาลัยรามคำแหงในกิจกรรมที่เกี่ยวข้องกับมหาวิทยาลัย โดยการวิเคราะห์คำถาม (Question Analysis) และสร้างฐานข้อมูลคำตอบ องค์ประกอบของการสร้างระบบใช้เทคนิคการทำเหมืองข้อความ ซึ่งประกอบด้วย การตัดคำการแจงประโยคด้วยกฎ และการแบ่งกลุ่มด้วยวิธี K Mean เพื่อหารูปแบบคำถามและความรู้ ผลการทดสอบระบบได้ค่าความถูกต้อง 0 88 ค่าความแม่นยำ 0.88 และค่าF (F Measure) 0.936

ผลจากการศึกษาระบบถาม-ตอบอัตโนมัติแบบทันที (Real Time Q-A) ผู้ศึกษาสามารถสรุปผลการศึกษาได้ ดังนี้ 1) สามารถลดค่าใช้จ่ายของหน่วยบริการ รวมทั้งผู้ถามจะได้คำตอบเพื่อดำเนินการในกิจกรรมการศึกษาได้อย่างรวดเร็วทันความต้องการสามารถพัฒนาตัวแบบเพื่อแก้ไขปัญหาความล่าช้าของกระบวนการระบบถาม-ตอบได้ 2) สามารถนำไปปรับใช้เพื่อเพิ่มประสิทธิภาพของระบบถาม-ตอบ และระบบอื่นที่มีลักษณะการทำงานในแบบเดียวกันได้ 3) ได้ตัวแบบระบบถาม-ตอบภาษาไทยนำร่องเพื่อพัฒนาให้มีประสิทธิภาพมากยิ่งขึ้น

ราชวิทย์ ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์ สมประเสริฐศรี [19] กระดานสนทนาเป็นสื่อในการแลกเปลี่ยนความคิดเห็นหรือซักถามข้อสงสัยที่ได้รับความนิยมและแพร่หลายในปัจจุบัน การนำกระดานสนทนามาใช้ในหน่วยงานหรือองค์กร เป็นการเพิ่มช่องทางการติดต่อสื่อสารอีกช่องทางหนึ่งที่ก่อให้เกิดประโยชน์และประสิทธิภาพในการให้บริการของหน่วยงาน แต่คำถามบนกระดานสนทนาที่ยังไม่มีการจัดหมวดหมู่อาจส่งผลกระทบต่อการตอบคำถาม ทำให้ผู้ตอบตอบคำถามไม่ตรงประเด็นหรือตอบคำถามไม่ได้เนื่องจากบางคำถามนั้นไม่ตรงกับหน่วยงานของตนเอง งานวิจัยนี้จึงนำเสนอการจัดกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ ซึ่งได้ทำการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกข้อความของ 3 เทคนิควิธี คือ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจและเทคนิคการเรียนรู้เบย์อย่างง่าย ผลการเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่า เทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยค่าความถูกต้องเท่ากับ 0.89 ค่าความเที่ยง เท่ากับ 0.9 ค่าความระลึก เท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892

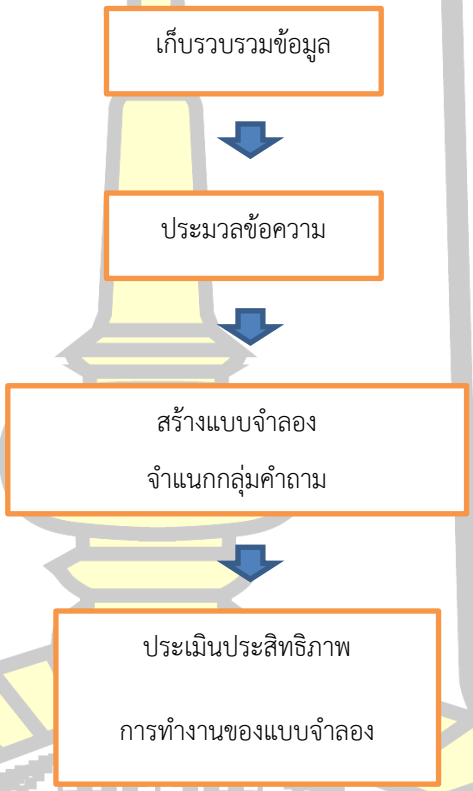
รุ่งกานต์ สุขลิ้ม [20] การเปรียบเทียบการวิเคราะห์ข้อมูลรูปแบบการเรียนรู้ของนักศึกษาปริญญาตรีสาขาคอมพิวเตอร์มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ตามหลักการเดวิด คอล์บ โดยใช้แบบจำลอง J48, SVM, DecisionTable และ Naïve Bayes. จากผลการวิจัยครั้งนี้ผู้วิจัยจะนำผลการวิเคราะห์ข้อมูลไปใช้เป็นกฎในการแบ่งรูปแบบการเรียนรู้ของผู้เรียนเพื่อทำนาย

รูปแบบการเรียนรู้ก่อนเข้าสู้บทเรียนโดยระบบการเรียนการสอนที่พัฒนาขึ้น จะสามารถตอบสนองผู้เรียนได้ตรงตามรูปแบบ การเรียนรู้ของแต่ละคนได้เป็นอย่างดี

รังสิพรรณ มฤคทัต [21] ปัญหาหนึ่งที่มาพร้อมกับการใช้สื่อสังคมออนไลน์ในประเทศไทยคือการโพสต์ข้อความล่อลวง หมิ่นประมาทหรือเผยแพร่ข้อมูลข่าวสารที่เป็นเท็จ ผู้เขียนข้อความอาจใช้ชื่อปลอมหรือแอบอ้างเป็นคนอื่นแต่รูปแบบลีลาการเขียนบางอย่างที่เป็นรสนิยมส่วนตัวหรือเกิดจากความเคยชิน เช่น การใช้คำเรียกตัวเอง คำลงท้ายประโยค เครื่องหมายวรรคตอน ยังปรากฏร่องรอยอยู่และสามารถตรวจจับได้งานวิจัยนี้จึงคัดเลือกคุณลักษณะในการเขียนข้อความออนไลน์ภาษาไทยจำนวน 53 คุณลักษณะและใช้คุณลักษณะเหล่านี้ในการระบุตัวผู้เขียนข้อความนิรนาม โดยวิธีการที่เลือกใช้คือการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจเมื่อทดสอบกับข้อความขนาดสั้น (ความยาวเฉลี่ย 144 คำ) ซัพพอร์ตเวกเตอร์แมชชีนให้อัตราความถูกต้องเฉลี่ย 79% ต้นไม้ตัดสินใจให้อัตราความถูกต้องเฉลี่ย 75% เมื่อทดสอบกับข้อความขนาดยาวขึ้น (ความยาวเฉลี่ย 312 คำ) ทั้งสองวิธีให้อัตราความถูกต้องเฉลี่ย 88% และ 82% ตามลำดับ

บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้จะกล่าวถึงวิธีการดำเนินงานวิจัยในการศึกษาเกี่ยวกับการจำแนกคำถามอัตโนมัติบนกระดานข่าว โดยใช้เทคนิคเหมืองข้อความ โดยจะนำเสนอกระบวนการและเทคนิคการทำเหมืองข้อความเพื่อแก้ปัญหาการจำแนกข้อความ ทดสอบประสิทธิภาพการทำงานของแต่ละเทคนิควิธี จากนั้นจะนำแบบจำลองของเทคนิควิธีการจำแนกที่ให้ประสิทธิภาพในการจำแนกดีที่สุดมาออกแบบและพัฒนาระบบเพื่อใช้งานผ่านเว็บไซต์ โดยแบ่งวิธีการทำงานออกเป็น 5 ขั้นตอนหลัก ดังภาพประกอบ 9 ภาพรวมกระบวนการทำงาน



รูปภาพ 9 ขั้นตอนการดำเนินงานวิจัย

3.1 การเก็บรวบรวมข้อมูล

งานวิจัยนี้ทำการเก็บรวบรวมข้อมูลจากกระดานข่าวและ Facebook บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม ซึ่งเป็นเว็บไซต์ของหน่วยงานมีหน้าที่ควบคุมกำกับนิสิตระดับบัณฑิตศึกษา

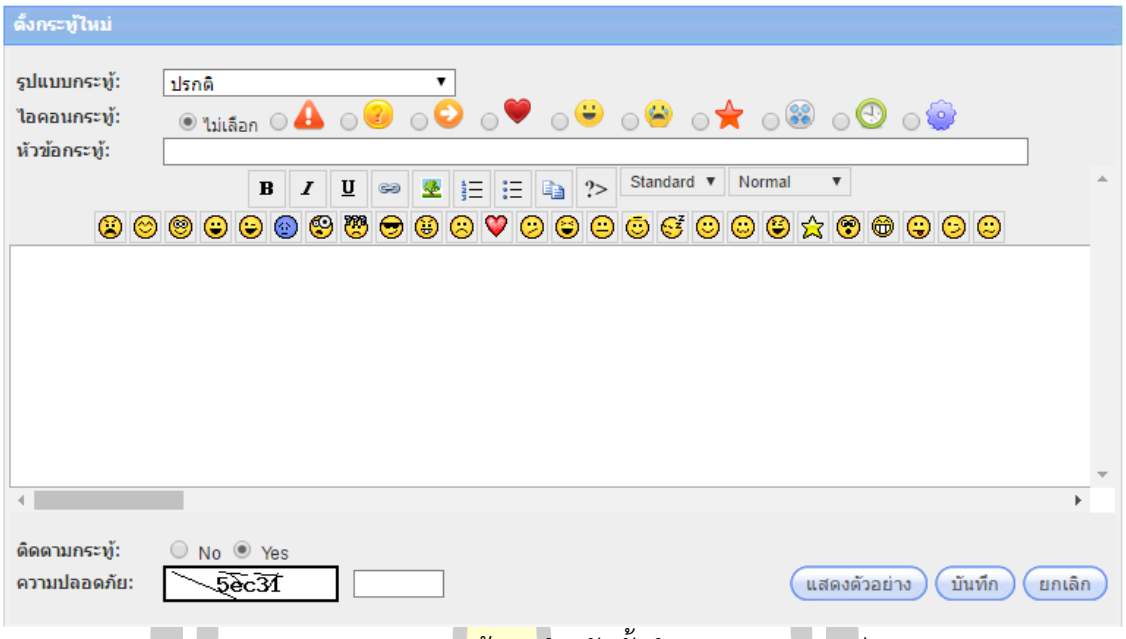
โดยเก็บข้อมูลตั้งแต่เดือนมิถุนายน 2552 ถึงปัจจุบันจำนวน 1,102 เรคคอร์ด มีผู้ที่สนใจเข้าศึกษาและนิสิตของมหาวิทยาลัยมหาสารคาม เข้ามาเยี่ยมชมและตั้งหัวข้อคำถามบนกระดานข่าวเป็นจำนวนมาก ข้อความที่ผู้สนใจตั้งคำถาม ส่วนมากจะเป็นการซักถามข้อมูลต่าง ๆ เกี่ยวกับการรับสมัครเข้าศึกษาการลงทะเบียน ค่าธรรมเนียมการศึกษา และเรื่องอื่น ๆ ที่เกี่ยวข้องกับมหาวิทยาลัยมหาสารคามผู้สนใจสามารถตั้งคำถามและเขียนบรรยายข้อความได้อย่างอิสระ โดยจะมีผู้อานวยการและบุคลากรกองบริการการศึกษา เป็นผู้ตอบคำถาม โดยเก็บข้อมูลในรูปแบบฐานข้อมูล มีรายละเอียดดังตาราง 3

ตาราง 3 ตารางเก็บข้อมูลคำถาม

ลาดับ	ชื่อคอลัมน์	รายละเอียด	ประเภท
1	TopicID	รหัสคำถาม	Int(6) PK
2	TopicTitle	หัวข้อคำถาม	Varchar(200)
3	TopicDetail	รายละเอียดของคำถาม	text
4	TopicName	ชื่อผู้ตั้งคำถาม	Varchar(60)
5	TopicContact	ข้อมูลติดต่อ	Varchar(60)
6	TopicDate	วัน เวลา ตั้งคำถาม	datetime
7	TopicReDate	วัน เวลา การตอบ	datetime
8	TopicReBy	ชื่อผู้ตอบ	Varchar(60)

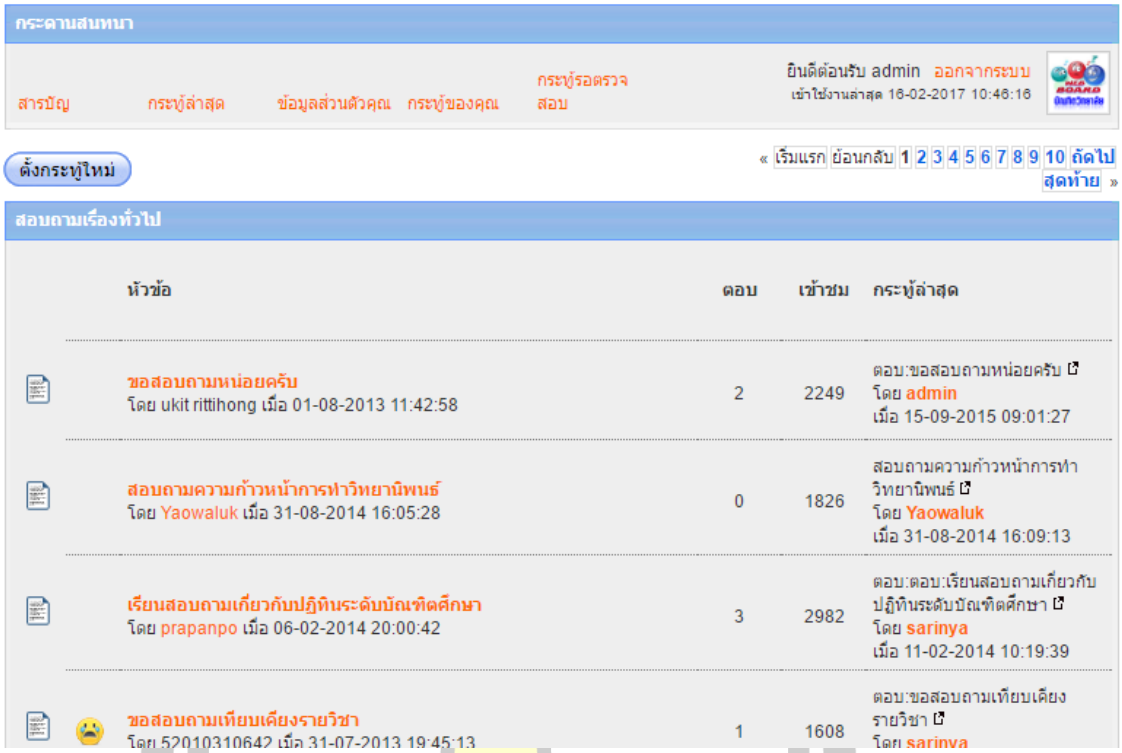
จากตารางที่ 3 ตารางเก็บข้อมูลคำถาม โครงสร้างของตารางประกอบไปด้วย 8 แอตทิบิว สำหรับแอตทิบิวที่นามาใช้วิเคราะห์คือ TopicTitle หรือ หัวข้อคำถาม เก็บข้อมูลในรูปแบบข้อความ ซึ่งคำถามจะนำมาวิเคราะห์ด้วยกระบวนการทำเหมืองข้อความ ผู้ตั้งคำถามสามารถกรอกรายละเอียดของคำถามผ่านหน้าเว็บไซต์ http://www.grad.msu.ac.th/2012/index.php/2012-01-13-17-30-35 และ Facebook https://www.facebook.com/profile.php?id=100006432858287

ภาพประกอบ 6 แสดงหน้าจอสำหรับตั้งคำถามบนกระดานข่าว



ภาพประกอบ 6 แสดงหน้าจอสำหรับตั้งคำถามบนกระดานข่าว

ภาพประกอบ 6 ตัวอย่างหัวข้อคำถามบนกระดานข่าว แสดงคำถามจากหน้าจอกระดานข่าว บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม โดยได้มาจากเว็บไซต์ ซึ่งยังไม่มีการจัดหมวดหมู่ของคำถามที่ชัดเจน

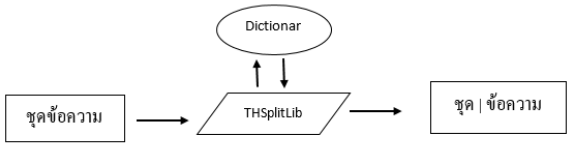


ภาพประกอบ 7 ตัวอย่างหัวข้อคำถามบนกระดานข่าว

3.2 ประมวลข้อความ

3.2.1 การตัดคำ

งานวิจัยนี้ใช้ โปรแกรม THSplitlib [12] ซึ่งเป็นโปรแกรม Open Source ในการตัดคำภาษาไทย ใช้หลักการตัดคำซึ่งดำเนินการโดยใช้พื้นฐานจากพจนานุกรม (Dictionary-Based Approach) ในการเปรียบเทียบการตัดคำกับคำที่จัดเก็บในพจนานุกรม ดัง ภาพประกอบ 8



ภาพประกอบ 8 การตัดคำ

จากภาพประกอบ 8 โดยอธิบายการทำงาน 3 ขั้นตอนดังนี้

ขั้นตอนที่ 1 เพิ่มคำถามใหม่ผ่านฟอร์มเพิ่มคำถาม

ขั้นตอนที่ 2 คำถามที่ป้อนเข้ามาจะถูกนำมาตัดคำโดยเทียบกับพจนานุกรม คำศัพท์ด้วยวิธีเทียบคำที่ยาวที่สุดที่พบในพจนานุกรม THSplitlib

ขั้นตอนที่ 3 เป็นการเปรียบเทียบคำที่ตัดแล้วกับคำสำคัญในฐานข้อมูล ว่าพบคำเหล่านั้นปรากฏในประโยคหรือไม่

ตาราง 4 ตัวอย่างผลการตัดคำภาษาไทย

นักศึกษา	พ้น	สภาพ	เนื่องจาก	เกรดเฉลี่ย
ไม่ถึง	2	อยาก	กลับ	เข้ามา
ใหม่	ใน	คณะ	เดิม	แต่
ไม่ทราบ	ข้อมูล	การ	รับสมัคร	ว่า
ต้อง	ทำ	อย่างไรบ้าง	?	รบกวน
ผู้อานวยการ	ช่วย	แนะนา	ด้วย	ค่ะ

3.2.2 การกำจัดคำหยุด

การกาจัดคำหยุดเป็นการกำจัดคำที่ไม่มีนัยสำคัญออก ซึ่งจำนวนคำหยุดที่ปรากฏในข้อความมีความถี่ค่อนข้างสูง คำหยุดอาจเป็นคุณลักษณะที่ไม่เกี่ยวข้อง อาจส่งผลกระทบในการจำแนกกลุ่มคำสำหรับคำหยุดในข้อความภาษาไทยจะทำการเปรียบเทียบคำที่พบในพจนานุกรม THSplitlib ได้แก่ คำบุพบท คำสันธาน คำสรรพนาม คำวิเศษณ์ คำอุทาน ดังตารางที่5 ตัวอย่างคำหยุด

ตาราง 5 ตัวอย่างคำหยุด

คำบุพบท	คำสันธาน	คำสรรพนาม	คำวิเศษณ์	คำอุทาน
กับ	กว่า	กระผม	ใกล้	555
ของ	คือ	ข้าพเจ้า	ครับ	อุ้ย
ซึ่ง	จึง	ฉัน	ง่าย	ขอโทษ
ด้วย	แต่	เธอ	ด้วย	ไง
โดย	ถ้า	นาย	ที่สุด	เยี่ยม
ที่	และ	ผม	ยาก	อิอิ
เพื่อ	หรือ	เรา	มาก	ฮา

3.2.3 โค้ดตัวอย่างไม่มีสาระสำคัญในประโยคออกไป

```
if($result!=""){

    $difresult=array("กับ","ของ","ซึ่ง","ไป","ด้วย","ที่","เพื่อ","และ","กว่า","คือ","จึง","แต่","ถ้า","
และ","หรือ","กระผม","ฉัน","เธอ","นาย","ผม","เรา","ใกล้
","ครับ","ง่าย","ด้วย","ที่สุด","ยาก","มาก","หลัง","หรือ","ไง","ส่วน","ส่ง","555","การ","ข้าพเจ้า","ขอ
โทษ","เยี่ยม","อิอิ","ฮา","  ");

    $result = array_diff($result, $difresult);

    print_r($result);
```

```
        sort($result);
        $iCount = count($result);
}
```

3.2.4 การเลือกคุณลักษณะหมวดหมู่คำถาม

การเลือกคุณลักษณะหมวดหมู่คำถามที่นำมาใช้ในการเรียนรู้เพื่อสร้างแบบจำลอง ได้จากการรวบรวมข้อมูลคำศัพท์บนกระดานข่าว ซึ่งการจำแนกกลุ่มข้อความได้ข้อมูลจากเจ้าหน้าที่บัณฑิตวิทยาลัย (นักวิชาการคอมพิวเตอร์ ชำนาญการพิเศษ) ซึ่งเป็นผู้เชี่ยวชาญในหน่วยงานเป็นผู้เลือกคำสำคัญ และได้จากการจัดทำแบบสอบถามผ่านเว็บไซต์ สามารถแบ่งหมวดหมู่คำถามออกเป็น 4 กลุ่ม ได้แก่ การรับเข้า สารสนเทศ บทนิพนธ์ มาตรฐานการสอบ รายละเอียดดังตารางที่ 6 หมวดหมู่คำถามบนกระดานข่าว สำหรับคำศัพท์ที่เป็นตัวแทนของแต่ละกลุ่ม รายละเอียดดังตารางที่ 7 ตารางที่ 8 ตารางที่ 9 และตารางที่ 10

ตาราง 6 หมวดหมู่คำถามบนกระดานข่าว

ลำดับ	ชื่อหมวดหมู่	คลาส	จำนวนคำสำคัญ
1	งานรับเข้า	Admissions	28
2	งานสารสนเทศ	Information	25
3	งานมาตรฐานบทนิพนธ์	thesis	28
4	งานมาตรฐานการสอบ	Standardized Exam	28

ตาราง 7 งานรับเข้า

ระบบ	เข้า	สำรอง	ระเบียบ
พิเศษ	สัมภาษณ์	ติด	กำหนดการ
สมัคร	ต่อ	หลักสูตร	คัดเลือก
คณะ	สาขา	Admissions	เลือก
รับ	แอด	เอก	นอกเวลา
รอบ	ประกาศ	เรียก	ในเวลา
ตรง	ปริญญา	โท	เกณฑ์

ตาราง 8 งานสารสนเทศ

ระบบ	iThesis	ไฟล์	system
สารสนเทศ	รหัสผ่าน	ข้อมูล	เปลี่ยนรหัส
เข้า	User	email	mendeley
generate	ชื่อใช้งาน	error	save
เชื่อมต่อ	password	เออเล่อ	เข้าใช้งาน
ฐานข้อมูล	connect	Database	ดาต้าเบส

ตาราง 9 งานมาตรบทนิพนธ์

วิทยานิพนธ์	Thesis	แก้ไข	system
บทนิพนธ์	Ts	การศึกษาค้นคว้าอิสระ	วันที่อนุมัติ
เข้า	Is	ปี	วันที่ส่ง
จัดทำ	เดือน	เสนอ	สำเร็จ
โครงร่าง	พิจารณา	วันเสนอ	เข้าใช้งาน
ส่ง	บรรณานุกรม	สมบูรณ์	เล่ม

ตาราง 10 งานมาตรฐานการสอบ

รายวิชา	วิชา	จัดสอบ	รหัสวิชา
อังกฤษ	ภาคต้น	ปลาย	ภาคเรียน
เทียบ	ภาค	ฤดูร้อน	ปิดระบบ
ผลสอบ	วันที่	เทียบโอน	เปิดระบบ
เผยแพร่	ประกาศ	นิสิต	ปริญญา
ข้อมูล	เอก	โท	ติดต่อ

3.2.5 การสร้างดัชนีคำสำคัญ

เนื่องจากคำถามยังอยู่ในรูปแบบภาษาธรรมชาติ จึงต้องแปลงข้อความให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเรียนรู้ได้ เพื่อสร้างตัวแทนเนื้อหาของเอกสาร งานวิจัยนี้จึงจัดทำดัชนีคำสำคัญในรูปแบบ TF-Weighting โดยการแทนเอกสารจะอยู่ในรูปแบบของเวกเตอร์คำ (Word Vector) ซึ่งจะคำนวณหาค่าน้ำหนักให้กับดัชนี หาได้จากสมการที่ 2-1 ผลลัพธ์การคำนวณดัชนีคำสำคัญที่ได้ดังตารางที่ 11 ตัวอย่างการทำดัชนีคำสำคัญด้วย TF-Weighting

ตาราง 11 ตัวอย่างการทำดัชนีคำสำคัญด้วย TF-Weighting

คำถาม	Word1	Word2	Word3	Word4	Word5	Word6	...	Word102	Class
Q1	0	0.2	0	0.1	0.2	0	...	0	Thesis
Q2	0.1	0.1	0.3	0	0	0	...	0	Admissions
Q3	0	0.1	0	0	0	0.2	...	0.1	Admissions
Q4	0.2	0	0	0	0	0	...	0	Admissions
Q5	0.1	0	0.1	0	0	0	...	0	Information
Q6	0	0	0	0	0.2	0	...	0	Thesis
Q7	0	0.2	0.1	0	0	0	...	0	Exam
...	0	0	0	0	0	0	...	0	Information
Q1165	0	0	0.2	0	0.2	0	...	0	Exam

3.3 สร้างแบบจำลองจำแนกกลุ่มข้อความ

การสร้างแบบจำลองในการจำแนกกลุ่มข้อความ งานวิจัยนี้ได้นำแบบจำลองข้อมูลระหว่างนาอีฟเบย์ (Naive Bayes) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)และโครงข่ายประสาทเทียม (Neural Network) มาใช้ในกระบวนการทดสอบการจำแนกข้อความ ซึ่งเป็นเทคนิคในการจำแนกข้อมูลแบบมีการเรียนรู้ ที่ได้รับความนิยมในการจำแนกประเภทข้อมูล จากงานวิจัยที่เกี่ยวข้องแสดงให้เห็นถึงประสิทธิภาพในการสร้างแบบจำลองที่ดี โดยใช้โปรแกรม Weka (Waikato Environment for Knowledge Analysis) ในการประมวลผลข้อมูล การแบ่งข้อมูลทดสอบจะแบ่งโดยใช้วิธีการประเมินผล K-fold Cross Validation กำหนด k-fold 10

3.4 ประเมินประสิทธิภาพการทำงานของแบบจำลอง

การประเมินประสิทธิภาพการทำงานของแบบจำลอง สำหรับงานวิจัยนี้ ใช้การวัดประสิทธิภาพโดยการหาค่าความเที่ยง (Precision) ค่าความระลึก (Recall) และค่าความถูกต้อง (Accuracy) โดยหาได้จากสมการที่ 2-13, 2-14 และ 2-15 ตามลำดับ และประเมินความสามารถของแบบจำลอง โดยวัดที่ประสิทธิผลของการจำแนกหมวดหมู่ตามแนวคิดทางด้านการค้นคืนสารสนเทศด้วยการวัด F-Measure ดังสมการที่ 2-16

บทที่ 4
ผลการดำเนินงานวิจัย

ผลการดำเนินการวิจัยในการเปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล ซึ่งใช้ข้อมูลจำนวน 1,102 เรคคอร์ด แอททริบิวจำนวน 102 แอททริบิวและคลาสผลลัพธ์ 4 คลาส ทดลองด้วยวิธีแบบ 10-fold cross validation สามารถแสดงประสิทธิภาพของแบบจำลองและแบบจำลองดังต่อไปนี้

4.1 ผลการจำแนกและการทดสอบประสิทธิภาพ

4.1.1 ผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคอีฟเบย์ (Naive Bayes) สามารถแสดงค่า Precision Recall และ F-measure ได้ดังตาราง 12

ตาราง 12 การจำแนกหัวข้อกระดานข่าวด้วยเทคนิคอีฟเบย์ (Naive Bayes)

Class	Precision	Recall	F-Measure
Information	0.909	0.769	0.833
Admission	0.875	0.875	0.875
Thesis Standard	0.625	0.833	0.714
Exam Standard	0.667	0.667	0.667
Average	76.90%	78.60%	77.23%

จากตารางที่ 12 แสดงผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคอีฟเบย์ (Naive Bayes) พบว่าค่า Precision มี Average อยู่ที่ 76.90% ค่า Recall มี Average อยู่ที่ 78.60% และค่า F-measure มี Average อยู่ที่ 77.23%

4.1.2 ผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคโครงข่ายประสาทเทียม (Neural Network) สามารถแสดงค่า Precision Recall และ F-measure ได้ดังตาราง 13

ตาราง 13 การจำแนกหัวข้อกระดานข่าวด้วยเทคนิคโครงข่ายประสาทเทียม (Neural Network)

Class	Precision	Recall	F-Measure
Information	0.786	0.846	0.815
Admission	0.857	0.75	0.8
Thesis Standard	0.625	0.833	0.714
Exam Standard	0	0	0
Average	**56.70%**	**60.73%**	**58.23%**

จากตารางที่ 13 แสดงผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคโครงข่ายประสาทเทียม (Neural Network) พบว่าค่า Precision มี Average อยู่ที่ 56.70% ค่า Recall มี Average อยู่ที่ 60.73% และค่า F-measure มี Average อยู่ที่ 58.23%

4.1.3 ผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) สามารถแสดงค่า Precision Recall และ F-measure ได้ดังตาราง 14

ตาราง 14 การจำแนกหัวข้อกระดานข่าวด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

Class	Precision	Recall	F-Measure
Information	0.917	0.846	0.88
Admission	0.889	1	0.941
Thesis Standard	0.714	0.833	0.769
Exam Standard	0.5	0.333	0.4
Average	**75.50%**	**75.30%**	**74.75%**

จากตารางที่ 14 แสดงผลการจำแนกหัวข้อกระดานข่าวด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)พบว่าค่า Precision มี Average อยู่ที่ 75.50% ค่า Recall มี Average อยู่ที่ 75.30% และค่า F-measure มี Average อยู่ที่ 74.75%

4.2 ผลการสร้างแบบจำลอง

ในงานวิจัยนี้ผู้วิจัยได้วัดประสิทธิภาพโดยใช้ค่า Precision, Recall, F-Measure และ Accuracy

Algorithms	Precision	Recall	F-measure	Accuracy
Neural Network	56.70%	60.73%	58.23%	73.33%
Naïve Bayes	76.90%	78.60%	77.23%	80.00%
SVM	75.50%	75.30%	74.75%	83.33%

บทที่ 5
สรุปผลและข้อเสนอแนะ

ผลการวิจัยในการจำแนกหัวข้อกระดานข่าวของบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม โดยอัตโนมัติ เนื้อหาประกอบด้วยสรุปผลการดาเนินงาน ปัญหาและอุปสรรค และข้อเสนอแนะสาหรับการวิจัย

5.1 สรุปผลการดำเนินงาน

ผลการเปรียบเทียบประสิทธิภาพแบบจำลองข้อมูลระหว่างนาอีฟเบย์ (Naive Bayes) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)และโครงข่ายประสาทเทียม (Neural Network) อัลกอริทึม โครงข่ายประสาทเทียม (Neural Network) ร้อยละ 73.53 อัลกอริทึม นาอีฟเบย์ (Naive Bayes) ร้อยละ 80.00 และอัลกอริทึม ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) 83.33

5.2 อภิปลายผล

งานวิจัยนี้มีวัตถุประสงค์เปรียบเทียบประสิทธิภาพการจำแนกหัวข้อกระดานข่าวโดยใช้เทคนิคเหมืองข้อมูล โดยใช้โปรแกรม Weka ซึ่งผลการทดสอบพบว่าอัลกอริทึม ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าที่มีประสิทธิภาพมากที่สุดถึง ร้อยละ 83.33

ข้อมูลที่ใช้ในการวิจัยนี้ เป็นข้อมูลที่เกี่ยวกับกระดานข่าวและFacebookบัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม ตั้งแต่ปี 2552-เมษายน 2561 จำนวน 1,102 Record การวิจัยได้แบ่งหมวดหมู่คำถามออกเป็น 4 กลุ่ม ได้แก่ การรับเข้า (Admission) งานระบบสารสนเทศ (Information) มาตรฐานบทนิพนธ์ (Thesis Standard) และงานมาตรการสอบ (Exam Standard) ซึ่งการนำข้อมูลเข้ามาทดลองในงานวิจัยนี้ อาจได้ค่าผลการทดลองที่แตกต่างกัน ขึ้นอยู่กับขนาดของข้อมูล ตัวแปรที่นำมาทดสอบ และเงื่อนไขในการจัดตารางสอนที่แตกต่างกัน ฉะนั้น การทดลองเพิ่มขนาดของข้อมูล เพิ่มตัวแปร และการปรับเปลี่ยนเงื่อนไข อาจจะทำให้ได้ผลการทดลองที่แตกต่างกัน

5.3 ปัญหาและอุปสรรค

จากการดำเนินงานวิจัยพบว่า เกิดปัญหาและอุปสรรค ซึ่งส่งผลกระทบต่อกระบวนการวิจัย โดยสามารถสรุปประเด็น ได้ดังนี้

5.3.1 การเก็บรวบรวมข้อมูลคำถามพบว่า มีคำที่เขียนผิด เขียนตัวย่อ และอักขระพิเศษต่าง ๆที่ผู้ตั้งคำถามสร้างขึ้นมา ส่งผลต่อจำนวนคำสำคัญ ไม่ตรงกับคลาสของกลุ่มคำ

5.3.2 การตั้งข้อคำถามที่สั้นเกินไป ในกรณีนี้จะส่งผลต่อการจำแนกกลุ่มคำถามได้ อาจทำให้การจำแนกคำถามไม่ถูกต้อง

5.3.3 คำถามหนึ่งหัวข้อสามารถจำแนกได้เพียงกลุ่มเดียวเท่านั้น ไม่สามารถจำแนกหลายกลุ่มได้

5.4 ข้อเสนอแนะสำหรับการวิจัย

แนวทางการวิจัยในอนาคตเพื่อศึกษาการจำแนกคำถามอัตโนมัติ ซึ่งเป็นข้อความภาษาไทยนั้น กระบวนการประมวลผลข้อความถือได้ว่าเป็นกระบวนการสำคัญในการจำแนกข้อความไม่ว่าจะเป็นภาษาใด ๆ ซึ่งจะส่งผลให้ประสิทธิภาพการจำแนกข้อความมีความถูกต้องมากยิ่งขึ้น งานวิจัยนี้ จึงมีข้อเสนอแนะ ดังนี้

5.4.1 ผู้วิจัยควรคำนึงถึงหลักการตัดคำที่จะนำมาใช้เพื่อให้เหมาะสมกับลักษณะคำที่จะนำมาทดสอบ เนื่องจากถ้าเลือกวิธีการตัดคำที่ไม่เหมาะสมแล้วอาจจะได้คำที่ไม่สมบูรณ์ และควรเลือกพจนานุกรมคำศัพท์ที่มีจานวนคำมากพอที่จะสามารถตัดคำที่มีความหมายได้ทุกคำ

5.4.2 การกาจัดคำหยุด ต้องคำนึงถึงความสำคัญของคำ.เนื่องจากคำหยุดบางคำอาจจะนำมาใช้เป็นคำสำคัญได้ .ขึ้นอยู่กับลักษณะของคำที่จะนำมาใช้ในการประมวลผล

5.4.4 ควรเลือกวิธีการสร้างดัชนีเอกสารที่เหมาะสมกับคุณลักษณะของคำและเทคนิควิธีที่จะใช้ทดสอบ ซึ่งการสร้างดัชนีเอกสารมีหลายรูปแบบ การเลือกวิธีการสร้างดัชนีเอกสารที่เหมาะสมจะมีผลต่อประสิทธิภาพของการจำแนกข้อความ

บรรณานุกรม

[1] วนิศา สถาพรวจนา, "การทำเหมืองข้อความและออนโทโลจีในการจำแนกความสนใจของ นักท่องเที่ยวต่อการท่องเที่ยวในประเทศไทย," ป.โท, วิทยานิพนธ์ วท.บ. (เทคโนโลยีสารสนเทศ), มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือม, กรุงเทพฯ, 2553.

[2] จักรพันธ์ จารภูมิ, "การทำเหมืองข้อความเว็บไซต์สังคมออนไลน์เพื่อวิเคราะห์กระแสการเมืองกรณีศึกษาข้อความสั้นเว็บไซต์ทวิตเตอร์," ป.โท, ปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2554.

[3] M.Sukanyal, S.Biruntha, "Techniques on Text Mining," in *IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, Syed Ammal Engineering College Ramanathapuram, 2012, pp. 269-271.

[4] วิชุดา โชติรัตน์, "ระบบวิเคราะห์ข่าวออนไลน์โดยใช้ฐานความรู้ออนโทโลยี และการทำเหมืองข้อความกรณีศึกษาข่าวในพื้นที่ 5 จังหวัดชายแดนภาคใต้ ประเทศไทย," ป.โท, วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2554.

[5] A. N. H. Cherfi, Y. Toussaint, "Towards a text mining methodology using association rule extraction," *Soft Computing* pp. 431-441, 2006.

[6] ศุภเทพ สติมั่น, "การทาเหมืองเครือข่ายสังคมออนไลน์เพื่อจำแนกประเภทความสนใจผู้ใช้งานกรณีศึกษาข้อความภาษาไทยบนเว็บไซต์ทวิตเตอร์," ป.โท, ปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2553.

[7] นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์, พยุง มีสัจ "การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ," *NIDA Development Journal* pp. 187-205, 2553.

[8] T. W, "Data and Text Mining: A business Application Approach," *Prentice Hall, United States,* 2004.

[9] W. L. Fan W, Rich S, "Tapping The Power of Text Mining," *Communication of the ACM* pp. 76-82, 2006.

[10] E. L. Aas K, "Text Categorization: a Suvey," 1999.

[11] นิเวศ จิระวิชิตชัย, "แบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ," วารสารวิชาการเทคโนโลยีอุตสาหกรรม pp. 141-149, 2556.

[12] สุวิชา เผือกอิ่ม. (2556, 2 เมษายน 2558). *THSplitLib* โปรแกรม. Available: http://www.alogik.com/thsplitlib/

[13] บุญเสริม กิจศิริกุล: จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 2548.

[14] พัชรนิกานต์ พงษ์ธนู, "วิเคราะห์ความพึงพอใจของลูกค้าจากข้อความคาแนะนำโดยการทำเหมืองความคิดเห็น," ป.โท, วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยขอนแก่น, ขอนแก่น, 2554.

[15] ชลธิศา พลทองมาก, "การวิเคราะห์ความเสี่ยงการเป็นโรคไวรัสตับอักเสบซี โดยต้นไม้การตัดสินใจ," ป.โท, การศึกษาอิสระปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยขอนแก่น, ขอนแก่น, 2553.

[16] M. P. Thammasiri D, "Ensemble Data Classification Based on Decision Tree, Artificial Neuron Network and Support Vector Machine Optimized by Genetic Algorithm," *The Journal of KMUTNB* pp. 81-90, 2011.

[17] L. Q. Huang H, Wu L, Huang T, Yuan S, "The Application Research of Topic Word List in Text Automatic Classification," *Wuhan, Chaina,* pp. 111-114, 2009.

[18] สิงห์ทัย สุขสว่างโรจน์, "ระบบถาม-ตอบภาษาไทยเพื่อใช้ในการตัดสินใจของนักศึกษามหาวิทยาลัยรามคำแหง," ป.โท, วท.บ. วิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยรามคำแหง, กรุงเทพฯ, 2558.

[19] ราชวิทย์ ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์ สมประเสริฐศรี "การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา," วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, pp. 493-502, 2557.

[20] รุ่งกานต์ สุขลิ้ม, กฤช สินธนะกุล, จรัญ แสนราช, "การเปรียบเทียบการวิเคราะห์ข้อมูลรูปแบบการเรียนรู้ของนักศึกษาปริญญาตรี สาขาคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ตามหลักการเดวิด คอล์บ โดยใช้แบบจำลอง J48, SVM," ป.โท, วท.บ. สาขาวิทยาคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2556.

[21] รังสิพรรณ มฤคทัต, "การระบุตัวผู้เขียนข้อความออนไลน์ภาษาไทยด้วยซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจ," วารสารวิชาการพระจอมเกล้าพระนครเหนือ, pp. 103-111, 2558.

พหุน ปณฺ ฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นายจักรฤษณ์ บุญสีลา
วันเกิด	วันที่ 22 สิงหาคม พ.ศ. 2531
สถานที่เกิด	อำเภอจตุรพักตรพิมาน จังหวัดร้อยเอ็ด
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 62 หมู่ 17 อำเภอจตุรพักตรพิมาน จังหวัดร้อยเอ็ด รหัสไปรษณีย์ 45180
ตำแหน่งหน้าที่การงาน	นักวิชาการคอมพิวเตอร์
สถานที่ทำงานปัจจุบัน	บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม อาคารราชนครินทร์(RN) ชั้น 2 ตำบลขามเรียง อำเภอกันทรวิชัย จังหวัดมหาสารคาม รหัสไปรษณีย์ 44150
ประวัติการศึกษา	พ.ศ. 2549 มัธยมศึกษาตอนปลาย โรงเรียนวาปีปทุม จังหวัดมหาสารคาม พ.ศ. 2552 วิทยาลัยเทคโนโลยีพณิชยการพลาญชัยร้อยเอ็ด จังหวัดร้อยเอ็ด พ.ศ. 2554 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยมหาสารคาม พ.ศ. 2562 ปริญญาวิทยาศาสตรมหาบัณฑต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม