

# การวิเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของอัลกอริทึม

TEXT CLASSIFICATION ANALYSIS BY STABILITY COMPARISON OF ALGORITHMS

วัชรวิวรรณ จิตต์สกุล

นักศึกษาปริญญาเอก คณะเทคโนโลยีสารสนเทศ  
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ  
Email: watchareewan.j@it.kmutnb.ac.th

ผู้ช่วยศาสตราจารย์ ดร.สุนันทา สดสี

อาจารย์ประจำภาควิชาการสื่อสารข้อมูลและเครือข่าย  
คณะเทคโนโลยีสารสนเทศ  
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ  
Email: sunantha.s@it.kmutnb.ac.th

## บทคัดย่อ

งานวิจัยนี้นำเสนอการวิเคราะห์การจำแนกข้อความ เพื่อศึกษาความเสถียรของอัลกอริทึมเพื่อการจำแนก 4 รูปแบบพื้นฐาน ได้แก่ ฐานกฎ ฐานต้นไม้ตัดสินใจ ฐานความน่าจะเป็น และฐานการเรียนรู้ กับข้อความทดสอบ 3 ชุด จากเว็บไซต์ [www.imdb.com](http://www.imdb.com), [www.yelp.com](http://www.yelp.com) และ [www.amazon.com](http://www.amazon.com) (ข้อมูล ณ วันที่ 11 พฤศจิกายน 2559) เพื่อศึกษาความเสถียรของอัลกอริทึม การวิเคราะห์เส้นโค้งคุณลักษณะสมบัตินิการทำงาน (ROC) และการทดสอบความแตกต่างของค่ากลางความถูกต้องของอัลกอริทึม (Paired-t Test) ถูกนำมาเสนอความเสถียรของอัลกอริทึมที่ศึกษา ทั้งนี้ผลการวิเคราะห์แสดงให้เห็นว่า อัลกอริทึมฐานต้นไม้ตัดสินใจ ได้แก่ Random Forest แสดงความเสถียรการจำแนกข้อความทดสอบได้สูงกว่าอัลกอริทึมที่ศึกษาอื่นๆ ด้วยค่าเฉลี่ย ROC > 0.80 และผลต่างของค่ากลางความถูกต้องของการทดสอบระหว่างการแบ่งข้อมูลออกเป็น 10 ส่วน และวิธีข้อมูลทดสอบเท่ากับ 0 อย่างมีนัยสำคัญ

**คำสำคัญ :** การจำแนกข้อความ ความเสถียรของอัลกอริทึม อัลกอริทึมเพื่อการจำแนก

## ABSTRACT

This research presents a text classification analysis aiming to compare the stability of four major types of classifiers, which are a rule based classifier, tree based classifier, probability based classifier, and learning based classifier. In this work, three-benchmark comment datasets, which were collected from “[www.imdb.com](http://www.imdb.com)”, “[www.yelp.com](http://www.yelp.com)”, and “[www.amazon.com](http://www.amazon.com)” (Accessed: 11 November 2016) were used to evaluate the stability of mentioned classifiers, by concerning a Receiver Operating Characteristic (ROC) and a paired t-test of predicted accuracy. The results showed that the tree based classifier, which is called Random Forest, presented the greatest stability performance with ROC > 0.80 and the difference of predicted accuracy between 10-fold cross validation and test dataset were “0”.

**KEYWORDS :** Text Classification, Stability of Algorithm, Classifier

## ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันข้อมูลที่เป็นเอกสารที่เป็นลักษณะอิเล็กทรอนิกส์เพิ่มขึ้นเป็นจำนวนมากในเครือข่ายอินเทอร์เน็ต โดยข้อความในเอกสารนั้นมีลักษณะไม่มีโครงสร้าง (Unstructured Text) อยู่ในรูปแบบของเว็บไซต์ อีเมล กระดานข่าวออนไลน์ ห้องสนทนา ตลอดจนสื่อสังคมออนไลน์ต่างๆ เป็นต้น ซึ่งข้อมูลลักษณะโครงสร้างไม่แน่นอนมีมากถึง 80%-90% ของข้อมูลทั้งหมด (ชูชาติ หลัไชยะศักดิ์, 2554; Kanimozhi & Venkatesan, 2015) อีกทั้งในปัจจุบัน ผู้ใช้งานในเครือข่ายอินเทอร์เน็ตมีพฤติกรรมการเข้าถึงหรือค้นหาข้อมูล เปลี่ยนแปลงไปจากอดีตที่เคยกระทำการค้นหาจากตำรา เอกสาร เปลี่ยนเป็นการค้นหาข้อมูลทางเว็บไซต์ อีกทั้งผู้ใช้งานในเครือข่ายอินเทอร์เน็ตยังมีการแลกเปลี่ยนความคิดเห็น หรือแสดงความคิดเห็นในเรื่องต่างๆ เช่น การเมือง การตลาด การศึกษา หรือบันเทิงผ่านทางเว็บไซต์ด้วย เป็นต้น

จากข้อความดังกล่าวข้างต้น แสดงให้เห็นว่าข้อความหรือความคิดเห็นในเครือข่ายอินเทอร์เน็ตมีจำนวนมากมหาศาล และมีความซับซ้อน และเพื่อให้เกิดประโยชน์ในการนำข้อความหรือความคิดเห็นมาใช้งาน การจำแนกข้อความ (Text Classification) จึงถูกนำมาใช้ในการวิเคราะห์ หรือจัดกลุ่มข้อความหรือความคิดเห็น

ในปัจจุบัน ผู้วิจัยจำนวนมากนำเสนอกระบวนการจำแนกข้อความในเครือข่ายอินเทอร์เน็ต เช่น การใช้อัลกอริทึมวิธีการจำแนกแบบสัมพันธ์ (Associative Classification) สำหรับการจำแนกความคิดเห็นทางการเมือง (พนิดา ทรงรัมย์, 2559) การจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมเบย์ (Naïve Bayes) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) สำหรับจำแนกข่าวภาษาอังกฤษด้านอาชญากรรม (ทิชากร เนตรสุวรรณ และไกรศักดิ์ เกษร, 2558) หรือวิเคราะห์ข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์ (กานดา แฝ้ววัฒนากุล, 2555) และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) ในการจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา (ราชวิทย์ ทิพย์เสนา และคณะ, 2557) เป็นต้น

จากงานวิจัยข้างต้นจะเห็นได้ว่า มีหลายอัลกอริทึมที่ใช้ในการจำแนกข้อมูล ซึ่งแต่ละอัลกอริทึมมีลักษณะพื้นฐาน

การทำงานที่ต่างกัน ดังนั้นงานวิจัยนี้จึงมีแนวคิดในการหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับการจำแนกข้อความที่ต่างกัน โดยการเลือกใช้อัลกอริทึมที่มีลักษณะพื้นฐานที่ต่างกัน เช่น ฐานกฎ (Rule based Classifier) ฐานต้นไม้ตัดสินใจ (Tree based Classifier) ฐานความน่าจะเป็น (Probability based Classifier) และฐานการเรียนรู้ (Learning based Classifier)

สำหรับโครงสร้างงานวิจัยนี้ ประกอบด้วยส่วนที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ส่วนที่ 3 วิธีการดำเนินงานวิจัย ส่วนที่ 4 ผลการดำเนินงาน และส่วนที่ 5 สรุปผลงานวิจัย

## ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในส่วนของทฤษฎีและงานวิจัยที่เกี่ยวข้องนั้น ผู้วิจัยทำการศึกษาอัลกอริทึมเกี่ยวกับการจำแนกข้อมูล โดยทำการศึกษาอัลกอริทึมที่มีลักษณะพื้นฐานที่ต่างกัน ในที่นี้ศึกษา 4 พื้นฐาน ได้แก่ ฐานกฎ ฐานต้นไม้ตัดสินใจ ฐานความน่าจะเป็น และฐานการเรียนรู้ ดังนี้

### 1. อัลกอริทึมฐานกฎ

กฎความสัมพันธ์ (Association Rules) เป็นการหาความสัมพันธ์ของข้อมูลสองชุดหรือมากกว่าสองชุดขึ้นไปภายในกลุ่มข้อมูลที่มีขนาดใหญ่ โดยทำการหารูปแบบที่เกิดขึ้นบ่อย (Frequent Pattern) และใช้ในการวิเคราะห์ความสัมพันธ์หรือทำนายปรากฏการณ์ต่าง ๆ

ในการหาความสัมพันธ์นั้นจะมีขั้นตอนวิธีการหาหลายวิธี ตัวอย่างหนึ่งของ Association Rules คือ Market Basket Analysis ใช้ในการหาความสัมพันธ์ของสินค้าที่ลูกค้ามักจะซื้อพร้อมกัน เพื่อใช้ในการจัดรายการส่งเสริมการขาย (Trupti A. Kumbhare, et al., 2014)

กฎความสัมพันธ์ (Association Rule) สามารถเขียนได้ในรูปเซตของ Item ที่เป็นเหตุ ไปสู่เซต ของ Item ที่เป็นผล โดยกำหนดให้

$I = \{ i_1, i_2, \dots, i_m \}$  เป็น set ของ Items

$D = \{ t_1, t_2, \dots, t_m \}$  เป็น set ของ Transaction ซึ่งแต่ละ Transaction ใน D จะมีหมายเลข Transaction ID ที่ไม่ซ้ำกันและกำหนดให้  $t$  เป็น Subset ของ  $I$

กฎการเชื่อมต่อกัน หรือ Conjunctive Rule นั้น เป็น อัลกอริทึมประเภทหนึ่งของกฎความสัมพันธ์ ซึ่งกฎที่ใช้นั้นเป็น ลักษณะของการให้เหตุผลแบบอุปนัย ซึ่งเป็นการสรุปผลจากการค้นหาความเป็นจริงที่ได้จากการสังเกต หรือการทดลองซ้ำหลาย ๆ ครั้ง จากการแบ่งเป็นกรณีย่อยๆ หลังจากนั้นนำมาสรุปให้ได้ผลลัพธ์ (Mohd Fauzi bin Othman & Thomas Moh Shan Yau, 2007)

## 2. อัลกอริทึมฐานต้นไม้

อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ที่มีลักษณะคล้ายกับต้นไม้ จะมีการสร้างกฎต่างๆ ขึ้นเพื่อใช้ในการตัดสินใจ ซึ่งต้นไม้ตัดสินใจนั้นมีการทำงานแบบ Supervised Learning คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ก่อนล่วงหน้าเรียกว่า Training Set ได้ อัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ (Qiang, Qin, et al., 2002; Quadri & Kalyankar, 2010)

รูปแบบของลักษณะโครงสร้างต้นไม้ประกอบไปด้วย โหนด (Node) และกิ่ง (Branch) ซึ่งแต่ละโหนดจะถูกแทนด้วยคุณลักษณะ (Feature) ของชุดข้อมูลที่นำมาเรียนรู้และทดสอบ แต่ละกิ่งของต้นไม้แสดงผลในการทดสอบ และโหนดใบ (Leaf Node) แสดงคลาสที่ผู้ใช้กำหนด ส่วนเกณฑ์การเลือกคุณลักษณะเพื่อนำมาเป็นโหนดของต้นไม้ นั้นมาจากการคำนวณค่าเกนสารสนเทศ (Information Gain) โดยพิจารณาคุณลักษณะที่มีค่าเกนสารสนเทศหรือมีค่าเอนโทรปี (Entropy) ต่ำ หมายความว่า คุณลักษณะนั้นมีความสามารถในการจำแนกหมวดหมู่สูง (Changki Lee & Gary Geunbae Lee, 2006)

อัลกอริทึม Random Forest เป็นอัลกอริทึมประเภทหนึ่งของอัลกอริทึมต้นไม้ตัดสินใจที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (Unpruned) หรือต้นไม้ถดถอย (Regression Trees) ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกจะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ เรียกข้อมูลส่วนนี้ว่า Out-of-Bag (OOB) ซึ่งวิธีการนี้เรียกว่า Bagging ผลลัพธ์ที่ได้จะอิสระจากต้นไม้ตัดสินใจแต่ละต้นถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด

อัลกอริทึม Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบเพื่อประมาณความผิดพลาดเพราะข้อมูล OOB นั้นถูกนำมาใช้ทดสอบต้นไม้ตัดสินใจแล้ว (Leo Breiman, 2001)

## 3. อัลกอริทึมฐานความน่าจะเป็น

Naive Bayes ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของทฤษฎีเบย์ (Bayes' Theorem) โดยจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัว กับตัวแปรตาม เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นของแต่ละความสัมพันธ์ในทางทฤษฎีแล้วการทำนายผลของ Naive Bayes ถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง ทฤษฎีเบย์ กำหนดให้  $P(H)$  คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์  $H$  และ  $P(H|E)$  คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์  $H$  เมื่อเกิดเหตุการณ์  $E$  จากตัวแปรที่กำหนด และแนวความคิดของ Bayes' Theorem นั้น สามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆ ดังสมการที่ 1

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (1)$$

Bayesian Logistic Regression เป็นการนำเอาการวิเคราะห์ถดถอยโลจิสติกส์ (Logistic Regression Analysis) เข้ามาช่วยเพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เพื่อทำนายว่า จะเกิดเหตุการณ์หนึ่งขึ้นหรือไม่ หรือมีโอกาสเกิดขึ้นเพียงใด โดยมีการกำหนดค่าตัวแปรตัวหนึ่ง หรือหลายตัวที่คาดว่าจะส่งผลต่อการเกิดเหตุการณ์นั้นๆ ในการวิเคราะห์ถดถอยโลจิสติกส์นั้น ต้องมีการประมาณค่าพารามิเตอร์ ซึ่งใช้วิธี Maximum likelihood (Brian Kulis, 2012)

## 4. อัลกอริทึมฐานการเรียนรู้

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นอัลกอริทึมหนึ่งที่ใช้พื้นฐานการเรียนรู้ ซึ่งใช้วิธีการจำแนกกลุ่มข้อมูลที่อาศัยระนาบการตัดสินใจมาใช้ในแบ่งกลุ่มข้อมูลออกเป็นสองส่วน โดยใช้หลักการสร้างเส้นแบ่งกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด โดยที่ซัพพอร์ตเวกเตอร์แมชชีนจะใช้แมปฟังก์ชัน (Mapping Function) สำหรับแปลงข้อมูลจากโดเมนเดิม Input Space ไปยังโดเมนที่เรียกว่า Feature Space และสร้าง

เคอร์เนลฟังก์ชัน(Kernel Function) บน Feature Space เพื่อใช้ในการวัดความคล้ายกันของข้อมูล สำหรับเคอร์เนลฟังก์ชัน ได้แก่ Linear, Polynomail Kernel, Radial Basic Function และ Sigmoid (Ali, Shamsuddin & Ismail, 2011)

จากอัลกอริทึมทั้ง 4 พื้นฐาน มีงานวิจัยที่เกี่ยวข้องดังตัวอย่างในตารางที่ 1

## วิธีดำเนินการวิจัย

การศึกษาอัลกอริทึมที่เหมาะสมที่สุดสำหรับการจำแนกข้อความที่แตกต่างกัน โดยการเลือกใช้อัลกอริทึมที่มีลักษณะพื้นฐานที่ต่างกันมีการดำเนินงาน ดังนี้

**ตารางที่ 1** ตารางแสดงงานวิจัยที่ใช้อัลกอริทึมทั้ง 4 พื้นฐานในการจำแนกข้อความ

งานวิจัย	ปี	พื้นฐาน	อัลกอริทึม	ผลงานวิจัย
จำแนกความคิดเห็นทางการเมืองบนเครือข่ายสังคมออนไลน์ ใช้วิธีการจำแนกแบบความสัมพันธ์ (พนิดา ทรงรัมย์, 2559)	2559	Rule-Base	Association Rule	สามารถจำแนกความคิดเห็นทางการเมืองทั้งเชิงบวกและเชิงลบได้ถูกต้องถึง 77.75%
การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคเหมืองข้อมูล (จุฑาทิพย์ ทิพย์พูล และ นิเวศ จิระวิจิตชัย, 2559)	2559	Tree Structure Probability Learning	Decision Tree Naïve Bayes K-Nearest Neighbor	ผลการเปรียบเทียบ พบว่า นาอ์ฟเบย์มีประสิทธิภาพในการจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมได้ดีที่สุด
จำแนกข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ (ทิชากร เนตรสุวรรณ และ ไกรศักดิ์ เกษร, 2558)	2558	Learning	Neural Network	วัดประสิทธิภาพการจำแนกข่าวภาษาอังกฤษ โดยค่าความเที่ยงตรง และค่าระลอก ซึ่งได้ค่าประสิทธิภาพที่สูงมากกว่า 80%
การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ (ราชวิทย์ ทิพย์เสนา และคณะ, 2557)	2557	Probability Learning Tree Structure	Naïve Bayes K-Nearest Neighbor Decision Tree	ผลการเปรียบเทียบประสิทธิภาพ พบว่า อัลกอริทึมเพื่อนบ้านใกล้ที่สุด ให้ประสิทธิภาพในการจำแนกดีที่สุด
การวิเคราะห์เหมืองข้อเสนอนะจากบทวิจารณ์รายการโทรทัศน์ (กานดา แม้วฒนากุล, 2555)	2555	Tree Structure Probability Learning	Decision Tree Naïve Bayes Support Vector Machine	ผลการวัดประสิทธิภาพ พบว่า อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนสามารถจำแนกข้อเสนอนะได้ดีที่สุด
การเปรียบเทียบประสิทธิภาพการจำแนกประเภทปัญหาสำหรับระบบถามตอบโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนนาอ์ฟเบย์และเคเนียร์เรสต์เนเบอร์ (วลัยลักษณ์ สุขสมบูรณ์ และสมชาย ปรากฏเจริญ, 2553)	2553	Probability Learning	Naïve Bayes K-Nearest Neighbor Support Vector Machine	ผลการวัดประสิทธิภาพ การเปรียบเทียบการจำแนก พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องดีที่สุด
Implementing News Article Category Browsing Based on Text Categorization Technique (Choochart Haruechaiyasak, et al., 2008)	2008	Probability Tree Structure Learning	Naïve Bayes Decision Tree Support Vector Machine	ผลการเปรียบเทียบพบว่า อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพดีที่สุด

## 1. การเลือกอัลกอริทึม

เนื่องจากผู้วิจัยต้องการศึกษาอัลกอริทึมรูปแบบใดเหมาะสมกับการจำแนกข้อมูลลักษณะที่แตกต่างกัน ดังนั้นจึงทำการเลือกอัลกอริทึมที่อยู่ในพื้นฐานที่แตกต่างกัน ดังนี้ ฐานกฎ (Rule based Classifier) ฐานต้นไม้ตัดสินใจ (Tree based Classifier) ฐานความน่าจะเป็น (Probability based Classifier) และฐานการเรียนรู้ (Learning based Classifier) โดยพิจารณาจากอัลกอริทึมที่มีการพัฒนาใหม่ล่าสุด แสดงดังตารางที่ 2

ตารางที่ 2 การเลือกอัลกอริทึม

Algorithm-based	Selected Algorithm
Rule based Classifier	Conjunctive Rule (Mohd Fauzi bin Othman & Thomas Moh Shan Yau, 2007)
Tree based Classifier	Random Forest (Leo Breiman, 2001)
Probability based Classifier	Bayesian Logistic Regression (Brain Kulis, 2012)
Learning based Classifier	Support Vector Machine (Ali, W., Shamsuddin & Ismail A.S., 2011)

## 2. ข้อมูลที่ใช้ในการทดสอบ และขั้นตอนการเตรียมข้อมูล

ข้อมูลที่ใช้ในการทดสอบนำมาจากฐานข้อมูล UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml>) (ข้อมูล ณ วันที่ 11 พฤศจิกายน 2559) จำนวน 3 ชุด ได้แก่ (1) ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์จากเว็บไซต์ [www.imdb.com](http://www.imdb.com) มีจำนวน 1,000 ระเบียบ (2) ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารจากเว็บไซต์ [www.yelp.com](http://www.yelp.com) มีจำนวน 3,726 ระเบียบ และ (3) ข้อมูลแสดงความคิดเห็นเกี่ยวกับสินค้าจากเว็บไซต์ [www.amazon.com](http://www.amazon.com) จำนวน 15,004 ระเบียบ ซึ่งข้อมูลทั้ง 3 ชุดเป็นลักษณะข้อความแสดงความคิดเห็น และมีการจำแนกข้อความออกเป็น 2 กลุ่ม คือ

“1” แทนการแสดงความคิดเห็นเชิงบวก และ “0” แทนการแสดงความคิดเห็นในเชิงลบ การเตรียมข้อมูลทำโดยการกรองเฉพาะข้อความที่นำเสนอความคิดเห็น ตัวอย่างข้อมูล เช่น Delicious and I will absolutely be back! “1”, Battery lasts only a few hours. “0” เป็นต้น

วิธีการทดสอบ การทดสอบแบ่งออกเป็น 2 รูปแบบ คือ รูปแบบที่ 1 ใช้วิธี 10-fold cross validation คือ ทำการแบ่งข้อมูลออกเป็น 10 ส่วนเท่ากัน หลังจากนั้นใช้ข้อมูลส่วนหนึ่งเป็นตัวทดสอบประสิทธิภาพของโมเดล และรูปแบบที่ 2 ใช้วิธีข้อมูลทดสอบ (Test Dataset) ซึ่งทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลที่ใช้สำหรับสอน (Training Set) 70% และข้อมูลสำหรับทดสอบ (Test Set) 30% โดยทำการทดสอบทั้ง 2 รูปแบบ ด้วยข้อมูล 3 ชุด กับ 4 อัลกอริทึม ที่เลือกไว้และทำซ้ำแต่ละรูปแบบจำนวน 5 ครั้ง

## 3. การประเมินประสิทธิภาพ

เพื่อศึกษาความเสถียรของอัลกอริทึม การวิเคราะห์ Receiver Operating Characteristic (ROC) และการทดสอบความแตกต่างของค่ากลางความถูกต้องของอัลกอริทึม (Paired-t Test) ถูกนำมาเสนอความเสถียรของอัลกอริทึมที่ศึกษา โดยมีสมมุติฐานของการทดสอบ คือ วิธีการวัดแบบ 10-fold cross validation และวิธีการวัดแบบ Test Dataset นั้นให้ผลต่างของค่ากลางความถูกต้องไม่แตกต่างกันอย่างมีนัยสำคัญ

## ผลการวิจัย

หัวข้อนี้นำเสนอผลการดำเนินงานวิจัยแบ่งเป็น 3 ส่วน คือ (1) การกำหนดค่าพารามิเตอร์ต่างๆ ของแต่ละอัลกอริทึม (2) การหาประสิทธิภาพและการวิเคราะห์ค่า ROC เพื่อคัดเลือกอัลกอริทึมที่ดีที่สุด และ (3) การทดสอบ Pair t-Test อัลกอริทึม ที่ทำการคัดเลือก โดยพิจารณาค่าความถูกต้องสำหรับการทดสอบจากการวัดแบบ Cross-validation 10 folds และ Test Dataset

### 1. การกำหนดค่าพารามิเตอร์

ทำการกำหนดค่าพารามิเตอร์ที่ใช้ในการทดสอบอัลกอริทึมทั้ง 4 อัลกอริทึม แสดงดังตารางที่ 3

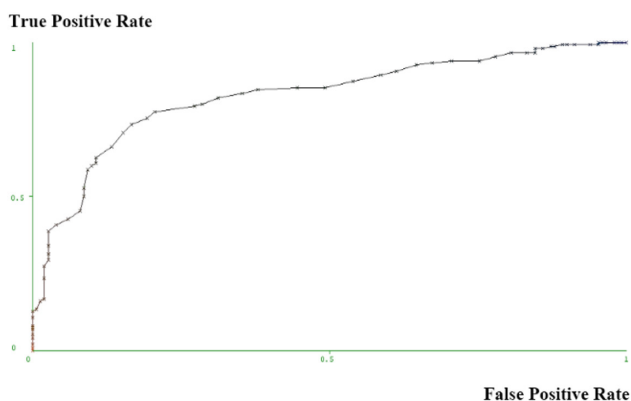
ตารางที่ 3 การกำหนดค่าพารามิเตอร์ต่างๆ ที่ใช้ในการทดสอบอัลกอริทึม

Algorithms	Stemmer	Considered Parameters
- Association Rule (Conjunctive Rule)	Iterated Lovins Stemmer	seed, folds, numtree,
- Decision Tree (Random Forest)	Lovins Stemmer	hyper parameter Selection,
- Naïve Bayes (Bayesian Logistic Regression)	Null Stemmer Snowball Stemmer	prior Class, thre shold, hyper parameter Value,
- Support Vector Machine (LibSVM)		kernel, gramma

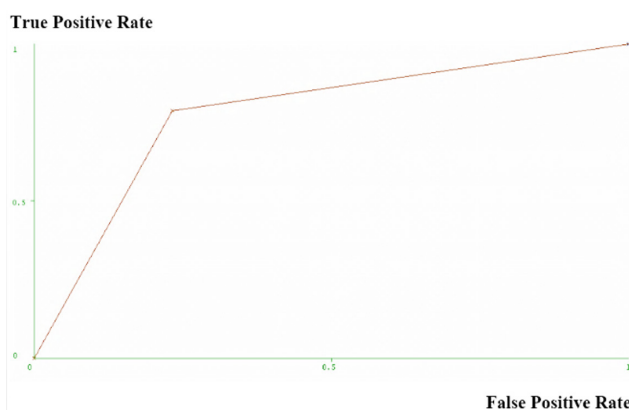
## 2. การหาประสิทธิภาพและการวิเคราะห์ค่า ROC

ทำการทดสอบโดยปรับค่าพารามิเตอร์ตามที่กำหนดไว้ในแต่ละอัลกอริทึม และเปรียบเทียบผลลัพธ์ พบว่า อัลกอริทึมที่ให้ผลลัพธ์ดีที่สุดสองอันดับแรกจากการทดสอบข้อมูลทั้ง 3 ชุด คือ อัลกอริทึม Random Forest และอัลกอริทึม Bayesian Logistic Regression ดังตารางที่ 4 ถึง 9 ซึ่งจากตารางพบว่า ในข้อมูลชุดที่ 1 และ 3 อัลกอริทึม Random Forest ให้ผลลัพธ์ที่ดีที่สุด ส่วนในข้อมูลชุดที่ 2 อัลกอริทึม Bayesian Logistic Regression ให้ผลลัพธ์ที่ดีที่สุด

### การวิเคราะห์ค่า ROC



ภาพที่ 1 กราฟ ROC ของข้อมูลชุดที่ 1 โดยอัลกอริทึม Random Forest



ภาพที่ 2 กราฟ ROC ของข้อมูลชุดที่ 1 โดยอัลกอริทึม Bayesian Logistic Regression

ในการวิเคราะห์ค่า ROC เมื่อพิจารณาในแต่ละชุดข้อมูลจากภาพที่ 1 กราฟ ROC ของข้อมูลชุดที่ 1 โดยอัลกอริทึม Random Forest แสดงให้เห็นว่า อัลกอริทึม Random Forest มีประสิทธิภาพมากกว่า เนื่องจากแนวโน้มของกราฟมีความชันมาก





ตารางที่ 7 ผลลัพธ์การวัดประสิทธิภาพด้วยอัลกอริทึม Bayesian Logistic Regression สำหรับข้อมูลชุดที่ 2

Parameter		Result													
threshold	priorClass	hyperparameterSelection	Stemmer	Correct	Incorrect	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Weighted Avg.					
										TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0.8	Gussian	Norm-based	LovinsStemmer	85.667	14.333	0.143	0.379	28.667	74.962	0.857	0.143	0.857	0.857	0.857	0.857

ตารางที่ 8 ผลลัพธ์การวัดประสิทธิภาพด้วยอัลกอริทึม Random Forest สำหรับข้อมูลชุดที่ 3

Parameter		Result												
numtree	seed	Stemmer	Correct	Incorrect	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Weighted Avg.					
									TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
100	2	LovinsStemmer	83.000	17.000	0.340	0.382	67.900	76.353	0.830	0.170	0.830	0.830	0.830	0.898

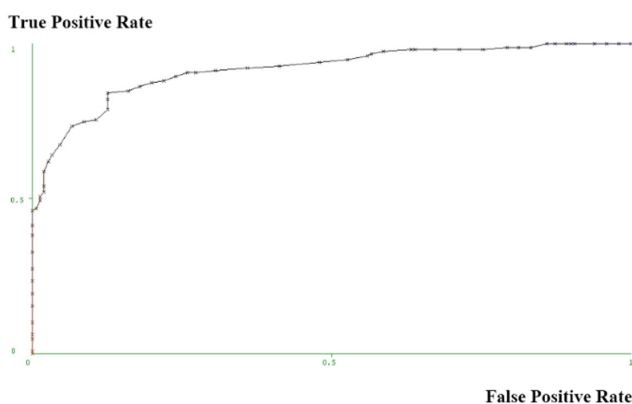
ตารางที่ 9 ผลลัพธ์การวัดประสิทธิภาพด้วยอัลกอริทึม Bayesian Logistic Regression สำหรับข้อมูลชุดที่ 3

Parameter		Result													
threshold	priorClass	hyperparameterSelection	Stemmer	Correct	Incorrect	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Weighted Avg.					
										TP Rate	FP Rate	Precision	Recall		

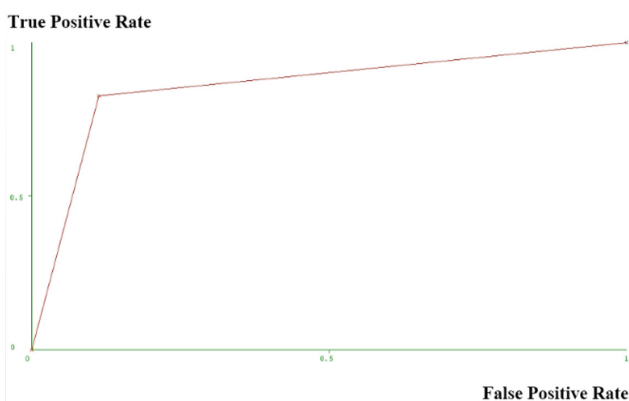


ตั้งแต่ค่า False Positive Rate มีค่า 0.4 ทำให้ค่า True Positive Rate มีค่า สูงถึง 0.8 และเส้นกราฟจะเริ่มคงที่โดยที่ค่า True Positive Rate เข้าใกล้ 1 ตั้งแต่ค่า False Positive Rate มีค่า 0.6 แสดงว่า จากกราฟแสดงให้เห็นว่า ไม่ว่าค่า False Positive Rate จะเป็นอย่างไร ตั้งแต่ 0.6 เป็นต้นไป จะทำให้ค่า True Positive Rate เข้าใกล้ 1 ซึ่งตรงกับวัตถุประสงค์ของงานวิจัยที่มุ่งเน้นความเสถียรของอัลกอริทึม

เมื่อพิจารณาจากภาพที่ 2 กราฟ ROC ของข้อมูลชุดที่ 1 โดยอัลกอริทึม Bayesian Logistic Regression จะเห็นว่า ค่า True Positive Rate กับ ค่า False Positive Rate แปรผันไปในทิศทางเดียวกัน หมายความว่าเมื่อค่า True Positive Rate สูง ค่า False Positive Rate ก็สูงด้วยเช่นกัน



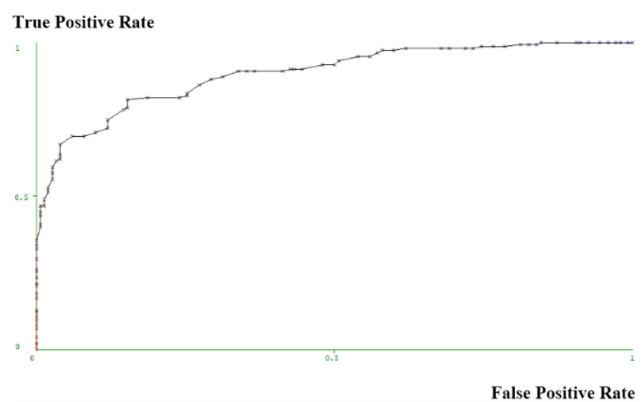
ภาพที่ 3 กราฟ ROC ของข้อมูลชุดที่ 2 โดยอัลกอริทึม Random Forest



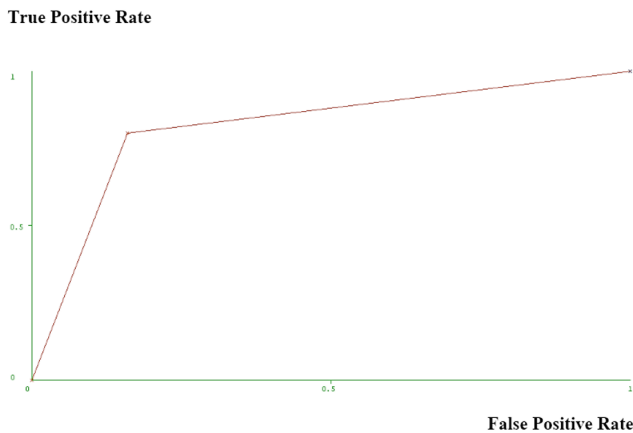
ภาพที่ 4 กราฟ ROC ของข้อมูลชุดที่ 2 โดยอัลกอริทึม Bayesian Logistic Regression

เมื่อพิจารณาข้อมูลชุดที่ 2 จากภาพที่ 3 กราฟ ROC ของข้อมูลชุดที่ 2 โดยอัลกอริทึม Random Forest แสดงให้เห็นว่าอัลกอริทึม Random Forest มีประสิทธิภาพดีกว่าอัลกอริทึม Bayesian Logistic Regression เนื่องจากแนวโน้มของกราฟ เริ่มมีความชันสูงขึ้นตั้งแต่ค่า False Positive Rate มีค่า 0.2 ทำให้ค่า True Positive Rate มีค่าสูงถึง 0.8 และเส้นกราฟจะเริ่มคงที่ โดยที่ค่า True Positive Rate เข้าใกล้ 1 ตั้งแต่ค่า False Positive Rate มีค่า 0.5 แสดงว่า ไม่ว่าค่า False Positive Rate จะเป็นค่าอะไร ตั้งแต่ 0.5 เป็นต้นไป ก็ทำให้ได้ค่า True Positive Rate ที่เข้าใกล้ 1 ซึ่งตรงกับวัตถุประสงค์ของงานวิจัยที่มุ่งเน้นความเสถียรของอัลกอริทึม

เมื่อพิจารณาจากภาพที่ 4 กราฟ ROC ของข้อมูลชุดที่ 2 โดยอัลกอริทึม Bayesian Logistic Regression ค่า True Positive Rate กับ ค่า False Positive Rate แปรผันไปในทิศทางเดียวกัน คือ ค่า True Positive Rate ใกล้เคียง 1 จะทำให้ค่า False Positive Rate ใกล้เคียง 1 ด้วยหมายความว่าหากต้องการค่า True Positive Rate ที่สูง ค่า False Positive Rate ก็สูงด้วยเช่นกัน



ภาพที่ 5 กราฟ ROC ของข้อมูลชุดที่ 3 อัลกอริทึม Random Forest



ภาพที่ 6 กราฟ ROC ของข้อมูลชุดที่ 3 อัลกอริทึม Bayesian Logistic Regression

เมื่อพิจารณาข้อมูลชุดที่ 3 จากภาพที่ 5 กราฟ ROC ของข้อมูลชุดที่ 3 โดยอัลกอริทึม Random Forest กราฟแสดงให้เห็นว่าอัลกอริทึม Random Forest มีประสิทธิภาพสูงกว่าอัลกอริทึม Bayesian Logistic Regression เนื่องจากแนวโน้มของกราฟมีความชันสูงตั้งแต่ค่า False Positive Rate มีค่า 0.3 ทำให้ค่า True Positive Rate มีค่าสูงถึง 0.8 และเส้นกราฟจะเริ่มคงที่ โดยที่ค่า True Positive Rate เข้าใกล้ 1 ตั้งแต่ค่า False Positive Rate มีค่า 0.6 แสดงให้เห็นว่าค่า False Positive Rate ตั้งแต่ 0.6 เป็นต้นไปจะช่วยให้ได้ค่า True Positive Rate ที่เข้าใกล้ 1 ซึ่งตรงกับวัตถุประสงค์ของงานวิจัยที่มุ่งเน้นความเสถียรของอัลกอริทึม

เมื่อพิจารณาจากภาพที่ 6 กราฟ ROC ของข้อมูลชุดที่ 3 โดยอัลกอริทึม Bayesian Logistic Regression ค่า True Positive Rate กับ ค่า False Positive Rate แปรผันไปในทิศทางเดียวกัน คือ ค่า True Positive Rate ใกล้เคียง 1 จะทำให้ค่า False Positive Rate ใกล้เคียง 1 ด้วย หมายความว่า หากต้องการค่า True Positive Rate ที่สูง ค่า False Positive Rate ก็สูงด้วยเช่นกัน เช่นเดียวกับการทดสอบในชุดข้อมูลชุดที่ 1 และ 2 เมื่อเปรียบเทียบผลการวัดประสิทธิภาพระหว่าง อัลกอริทึม Random Forest และ อัลกอริทึม Bayesian Logistic Regression พบว่า อัลกอริทึม Random Forest แสดงผลลัพธ์ที่ดี

### 3. การทดสอบ Pair t-Test ของอัลกอริทึม Random Forest

การทดสอบ Pair t-Test ถูกนำมาพิจารณาค่าความถูกต้องสำหรับการทดสอบจากการวัดแบบ 10 folds-cross validation และ Test Set เพื่อยืนยันความเสถียรของอัลกอริทึม Random Forest

จากผลการทดสอบ Pair t-Test ของอัลกอริทึม Random Forest กับทั้ง 3 ชุดข้อมูลทดสอบ พบว่า วิธีการวัดแบบ 10 folds-cross validation และวิธีการวัดแบบ Test Dataset นั้นให้ผลต่างของค่ากลางความถูกต้องเท่ากับ 0 อย่างมีนัยสำคัญในข้อมูลทดสอบจากผลการทดสอบแสดงว่าอัลกอริทึม Random Forest นั้นมีความเสถียรเนื่องจากไม่ว่าจะวัดด้วยวิธีการไหนก็ให้ค่าผลลัพธ์ที่ไม่แตกต่างกัน

### อภิปรายผล

งานวิจัยนี้นำเสนอการวิเคราะห์การจำแนกข้อความเพื่อศึกษาความเสถียรของอัลกอริทึมเพื่อการจำแนก 4 รูปแบบพื้นฐาน ได้แก่ ฐานกฎ เลือกอัลกอริทึม Conjunctive Rule ฐานต้นไม้ตัดสินใจ เลือกอัลกอริทึม Random Forest ฐานความน่าจะเป็น เลือกอัลกอริทึม Bayesian Logistic Regression และฐานการเรียนรู้ เลือก Support Vector Machine กับข้อความทดสอบนำมาจากฐานข้อมูล UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml>) จำนวน 3 ชุด ได้แก่ ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์จากเว็บไซต์ [www.imdb.com](http://www.imdb.com) ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารจากเว็บไซต์ [www.yelp.com](http://www.yelp.com) และ ข้อมูลแสดงความคิดเห็นเกี่ยวกับสินค้าจากเว็บไซต์ [www.amazon.com](http://www.amazon.com) ซึ่งข้อมูลทั้ง 3 ชุดเป็นลักษณะข้อความแสดงความคิดเห็น แบ่งออกเป็น 2 กลุ่ม คือ “1” แทนการแสดงความคิดเห็นเชิงบวก และ “0” แทนการแสดงความคิดเห็นเชิงลบ ทำการปรับค่าพารามิเตอร์ตามที่กำหนด ใช้การทดสอบ 2 รูปแบบ คือ วิธี 10-fold cross validation และวิธี Test Dataset ผลลัพธ์ พบว่าอัลกอริทึมที่ให้ผลลัพธ์ที่ดีที่สุดสองอันดับแรกของข้อมูลทั้ง 3 ชุด คือ อัลกอริทึม Random Forest และอัลกอริทึม Bayesian Logistic Regression ต่อมาทำการวิเคราะห์ ROC พบว่าอัลกอริทึม Random Forest นั้นให้ประสิทธิภาพที่ดีกว่าเนื่องจากแนวโน้ม

**ตารางที่ 10** ตารางแสดงผลลัพธ์ของการทดสอบ Pair t-Test สำหรับข้อมูลชุดที่ 1

Paired Samples Test					
	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Pair 1 SuppliedTestSet – CrossValidation10Folds	0.03373	2.39496	0.154	119	0.878

**ตารางที่ 11** ตารางแสดงผลลัพธ์ของการทดสอบ Pair t-Test สำหรับข้อมูลชุดที่ 2

Paired Samples Test					
	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Pair 1 SuppliedTestSet – CrossValidation10Folds	0.35873	1.61983	2.426	119	0.071

**ตารางที่ 12** ตารางแสดงผลลัพธ์ของการทดสอบ Pair t-Test สำหรับข้อมูลชุดที่ 3

Paired Samples Test					
	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Pair 1 SuppliedTestSet – CrossValidation10Folds	-.00516	1.47175	-.038	119	0.969

ของกราฟมีความชันตั้งแต่ค่าตั้งแต่ 0.2-0.4 ก็ทำให้ได้ค่า True Positive Rate เข้าใกล้ 1 ซึ่งตรงกับวัตถุประสงค์ของงานวิจัย ที่มุ่งเน้นความเสถียรของอัลกอริทึม และเพื่อยืนยันความเสถียรของอัลกอริทึม Random Forest ได้ทำการทดสอบความแตกต่างของค่ากลาง ความถูกต้องของอัลกอริทึม (Paired-t Test) ทั้งนี้ ผลการวิเคราะห์แสดงให้เห็นว่า อัลกอริทึมฐาน Random Forest แสดงความเสถียรการจำแนกข้อความได้สูงกว่า อัลกอริทึมที่ศึกษาอื่นๆ ด้วยค่าเฉลี่ย ROC > 0.80 และ ผลต่างของค่ากลางความถูกต้องเท่ากับ 0 เมื่อทดสอบระหว่างวิธี 10-fold cross validation และวิธี Test Dataset อย่างมีนัยสำคัญ ซึ่งงานวิจัยนี้สามารถนำไปประยุกต์ใช้ในเรื่องของการสนับสนุนการตัดสินใจ การให้คำแนะนำได้

## เอกสารอ้างอิง

- กานดา แผ้ววัฒนากุล. 2555. “การวิเคราะห์เหมืองข้อมูลจากบทวิจารณ์รายการโทรทัศน์.” วิทยานิพนธ์ปริญญาโทบริหารเทคโนโลยีสารสนเทศ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.
- จุฑาทิพย์ ทิพย์พูล และนิเวศ จิระวิชิตชัย. 2559. “การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคเหมืองข้อมูล.” **วารสารวิทยาศาสตร์และเทคโนโลยี มทร.ธัญบุรี** 6, 1: 102-109.
- ชูชาติ หลูไชยศักดิ์. 2554. **Material: Text mining**. สืบค้นวันที่ 11 พฤศจิกายน 2559 จาก <http://www2.it.kmutnb.ac.th/teacher/choochart>
- ทิชากร เนตรสุวรรณ และไกรศักดิ์ เกษร. 2558. “การจำแนกข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความ.” **รายงานสืบเนื่องการประชุมวิชาการระดับชาติ NCCIT2015 ครั้งที่ 11 วันที่ 2-3 กรกฎาคม 2558 ณ โรงแรมอโนมากรุงเทพ**, หน้า 61-65.
- พนิดา ทรงรัมย์. 2559. “การจำแนกความคิดเห็นทางการเมืองบนเครือข่ายสังคมออนไลน์ โดยใช้วิธีการจำแนกแบบสัมพันธ์.” **วารสารวิทยาศาสตร์และเทคโนโลยี มทร.ธัญบุรี** 6, 1: 83-93.

ราชวิทย์ ทิพย์เสนา ฉัตรเกล้า เจริญผล และแกมกาญจน์ สมประเสริฐศรี. 2557. “การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา.” **วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม** 33, 5: 493-502.

วลัยลักษณ์ สุขสมบูรณ์ และสมชาย ปราการเจริญ. 2553. “การเปรียบเทียบประสิทธิภาพการจำแนกประเภทปัญหาสำหรับระบบถามตอบโดยใช้ซอฟต์แวร์เวกเตอร์แมชชีน นาอ์ฟเบย์และเคเนียร์เรสต์เนเบอร์.” **รายงานสืบเนื่องการประชุมวิชาการ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ครั้งที่ 7 วันที่ 7-8 ธันวาคม 2553 ณ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน**, หน้า 747-756.

Ali, W., Shamsuddin and Ismail A.S. 2011. “Web Proxy Cache Content Classification based on Support Vector Machine.” **Journal of Artificial Intelligence** 4, 1: 100-109.

Brian Kulis. 2012. **Lecture 6: Bayesian Logistic Regression**. CSE 788.04: Topics in Machine Learning.

Changki Lee and Gary Geunbae Lee, 2006. “Information Gain and Divergence - base Feature Selection for Machine Learning-based Text Categorization.” **Journal of Information Processing and Management: an International Journal - Special issue: Formal methods for information retrieval** 42, 1: 155-165.

Choochart tHaruechaiyasak, Wittawat Jitkritum, Chatchawal Sangkeetrakarn, and Chaianun Damrongrat. 2008. “Implementing News Article Category Browsing Based on Text Categorization Technique.” In: **IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology**, pp. 143-146.

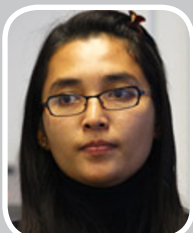
- Kanimozhi K. V. and Venkatesan M. 2015. "Unstructured Data Analysis - A Survey." **International Journal of Advanced Research in Computer and Communication Engineering** 4, 3: 223-225.
- Leo Breiman, 2001. "Random Forests." **Machine Learning** 45, 1: 5-32.
- Mohd Fauzi bin Othman and Thomas Moh Shan Yau. 2007. "Comparison of Different Classification Techniques Using WEKA for Breast Cancer." In: **3rd Kuala Lumpur International Conference on Biomedical Engineering**, pp. 520-523.
- Quadri M. and Kalyankar. D. N. V. 2010. "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques." **Global Journal of Computer Science and Technology** 10, 2: 2-5.
- Qiang Ding, Qin Ding, and William Perrizo. 2002. "Decision tree classification of spatial data streams using Peano Count Trees." In: **Proceedings of the 2002 ACM symposium on Applied computing**, pp. 413-417.
- Trupti A. Kumbhare and Santosh V. Chobe 2014. "An Overview of Association Rule Mining Algorithms." **International Journal of Computer Science and Information Technologies** 5, 1: 927-930.



#### วัชรวัชรณ จิตต์สกุล

สำเร็จการศึกษาปริญญาโท (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ พ.ศ. 2549 และสำเร็จการศึกษาปริญญาตรี (วศ.บ.) สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี พ.ศ. 2545

ปัจจุบันปฏิบัติงานในตำแหน่งนักวิจัย คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ



#### ผู้ช่วยศาสตราจารย์ ดร.สุนันทา สดสี

สำเร็จการศึกษาปริญญาเอก (Dr.-Ing) สาขาวิศวกรรม Communication Network มหาวิทยาลัย FernUniversität in Hagen สหพันธ์สาธารณรัฐเยอรมนี พ.ศ. 2555 สำเร็จการศึกษาปริญญาเอก (ปร.ด.) สาขาวิชาเทคโนโลยีสารสนเทศ (นานาชาติ) มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ พ.ศ. 2554 สำเร็จการศึกษาปริญญาโท (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ(นานาชาติ) มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ พ.ศ.2548 และสำเร็จการศึกษาปริญญาตรี (วศ.บ.) สาขาวิชาวิศวกรรมโทรคมนาคม สถาบันเทคโนโลยีพระจอมเกล้า เจ้าคุณทหารลาดกระบัง พ.ศ. 2545

ปัจจุบันปฏิบัติงานในตำแหน่งรองคณบดีฝ่ายวิชาการและวิจัย คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ