การระบุตัวผู้เขียนข้อความออนไลน์ภาษาไทยด้วยซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจ

รังสิพรรณ มฤคทัต*
รองศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดล

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0-2889-2138 ต่อ 6251 อีเมล: rangsipan.mar@mahidol.ac.th
รับเมื่อ 8 กันยายน 2557 ตอบรับเมื่อ 15 ธันวาคม 2557
DOI: 10.14416/j.kmutnb.2014.12.006

บทคัดย่อ

ปัญหาหนึ่งที่มาพร้อมกับการใช้สื่อสังคมออนไลน์ในประเทศไทยคือ การโพสต์ข้อความล่อลวง หมิ่นประมาท หรือเผยแพร่ข้อมูลข่าวสารที่เป็นเท็จ ผู้เขียนข้อความอาจใช้ชื่อปลอมหรือแอบอ้างเป็นคนอื่น แต่รูปแบบลีลาการเขียนบางอย่างที่เป็นรสนิยมส่วนตัวหรือเกิดจากความเคยชิน เช่น การใช้คำเรียกตัวเอง คำลงท้ายประโยค เครื่องหมายวรรคตอน ยังปรากฏร่องรอยอยู่และสามารถตรวจจับได้ งานวิจัยนี้จึงคัดเลือกคุณลักษณะในการเขียนข้อความออนไลน์ภาษาไทยจำนวน 53 คุณลักษณะและใช้คุณลักษณะเหล่านี้ในการระบุตัวผู้เขียนข้อความนิรนาม โดยวิธีการที่เลือกใช้คือการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจ เมื่อทดสอบกับข้อความขนาดสั้น (ความยาวเฉลี่ย 144 คำ) ซัพพอร์ตเวกเตอร์แมชชีนให้อัตราความถูกต้องเฉลี่ย 79% ต้นไม้ตัดสินใจให้อัตราความถูกต้องเฉลี่ย 75% เมื่อทดสอบกับข้อความขนาดยาวขึ้น (ความยาวเฉลี่ย 312 คำ) ทั้งสองวิธีให้อัตราความถูกต้องเฉลี่ย 88% และ 82% ตามลำดับ

คำสำคัญ: ข้อความออนไลน์ การระบุตัวผู้เขียน การจำแนก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ

การอ้างอิงบทความ: รังสิพรรณ มฤคทัต, "การระบุตัวผู้เขียนข้อความออนไลน์ภาษาไทยด้วยซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจ," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 25, ฉบับที่ 1, หน้า 103-111, ม.ค. - เม.ย. 2558. http://dx.doi.org/10.14416/j.kmutnb.2014.12.006

Author Identification of Thai Online Messages Using Support Vector Machine and Decision Tree

Rangsipan Marukatat*
Associate Professor, Department of Computer Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand

* Corresponding Author, Tel. 0-2889-2138 ext. 6251, E-mail: rangsipan.mar@mahidol.ac.th
Received 8 September 2014; Accepted 15 December 2014
DOI: 10.14416/j.kmutnb.2014.12.006

Abstract

One problem that comes with the use of online social media in Thailand is the posting of deceptive, abusive, or hoax messages. The authors of such messages may use fake accounts or impersonate innocent persons. But some of their writing styles, influenced by individual preferences or habits, such as the use of first-person pronouns, sentence-ending words, or punctuations can still be traced and detected. In this research, fifty-three writing attributes of Thai online messages were selected and used to identify the authors of anonymous messages. The identification methods were based on classification by support vector machine and decision tree. When testing with short messages (average length of 144 words), support vector machine yielded an average accuracy of 79% whereas decision tree yielded an average accuracy of 75%. When testing with long messages (average length of 312 words), both methods yielded average accuracies of 88% and 82%, respectively.

Keywords: Online Messages, Author Identification, Classification, Support Vector Machine, Decision Tree

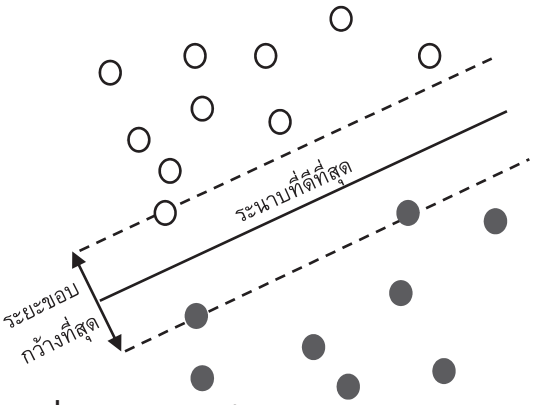
Please cite this article as: R. Marukatat, "Author Identification of Thai Online Messages Using Support Vector Machine and Decision Tree," *J. KMUTNB*, Vol. 25, No. 1, pp. 103-111, Jan. - Apr. 2015 (in Thai). http://dx.doi.org/10.14416/j.kmutnb.2014.12.006

1. บทนำ

จากรายงานของสำนักงานสถิติแห่งชาติ [1] พบว่าผู้ใช้อินเทอร์เน็ตในประเทศไทยมีจำนวนมากถึง 18 ล้านคน หรือคิดเป็นร้อยละ 29 ของประชากรทั้งหมด โดยการใช้สื่อสังคมออนไลน์เช่น เฟซบุ๊ก ทวิตเตอร์ และเว็บบอร์ด เป็นกิจกรรมหลักอย่างหนึ่งของผู้ใช้อินเทอร์เน็ตกว่า 10 ล้านคน ปัญหาที่มาพร้อมกับการใช้สื่อเหล่านี้คือการโพสต์ข้อความล่อลวง หมิ่นประมาท หรือข้อมูลเท็จ ซึ่งสามารถแพร่กระจายไปอย่างรวดเร็ว โดยผู้เขียนข้อความอาจใช้ชื่อปลอมหรือแอบอ้างเป็นบุคคลอื่น การระบุตัวผู้เขียนข้อความนอกจากจะดูจาก IP Address แล้ว รูปแบบลีลาเฉพาะตัวในการเขียนข้อความก็เป็นสิ่งที่บ่งชี้ถึงตัวผู้เขียนได้เช่นกัน งานวิจัยเกี่ยวกับรูปแบบลีลาการเขียนส่วนใหญ่วิเคราะห์ข้อความภาษาอังกฤษ [2]-[7] ซึ่งมีบางจุดที่แตกต่างจากข้อความออนไลน์ภาษาไทย เช่น

- ภาษาไทยเขียนคำติดกันและแบ่งประโยคด้วยการเว้นวรรค ซึ่งต่างจากภาษาอังกฤษ
- การใช้คำเรียกตัวเองและคำลงท้ายประโยคในภาษาไทย ซึ่งบ่งบอกเพศของผู้เขียน
- การใช้ภาษาไทยปนอังกฤษและการสะกดคำไทยด้วยอักขระอังกฤษ (ภาษาคาราโอเกะ) โดยเฉพาะเมื่อพิมพ์ข้อความจากโทรศัพท์มือถือหรือแท็บเล็ต
- การใช้ Emoticon หน้าตรงเช่น ^_^ ซึ่งได้รับอิทธิพลจาก Kaomoji และเป็นที่นิยมในแถบเอเชีย [8]
- การใช้สัญลักษณ์อื่นๆ เช่น 555 แทนเสียงหัวเราะในภาษาไทย

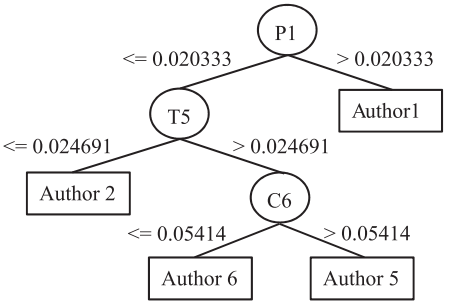
การระบุตัวผู้เขียนข้อความจากรูปแบบลีลาการเขียนสามารถกระทำได้ด้วยวิธีการจำแนก โดยเริ่มจากการสร้างต้นแบบการจำแนกจากข้อมูลฝึกหรือข้อความตัวอย่างที่รวบรวมจากกลุ่มผู้เขียน คลาสแต่ละคลาสคือผู้เขียนแต่ละคน ตัวแปรอิสระคือคุณลักษณะต่างๆ ที่สกัดจากข้อความ เช่น การใช้คำและสัญลักษณ์ต่างๆ ความยาวเฉลี่ยของประโยคและย่อหน้า ความหลากหลายในการใช้คำ เป็นต้น จากนั้นจึงนำต้นแบบซึ่งอาจอยู่ในรูปกฎ ต้นไม้ตัดสินใจ หรือฟังก์ชันคณิตศาสตร์ ไปจำแนกข้อความนิรนามว่าตรงกับคลาสหรือผู้เขียนคนใด วิธีจำแนกที่ได้รับความนิยมมีหลายวิธี เช่น ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) โครงข่ายประสาทเทียม (Neural Network) เพื่อนบ้านใกล้ที่สุด (K-nearest Neighbor หรือ KNN) และต้นไม้ตัดสินใจ (Decision Tree) เป็นต้น



รูปที่ 1 การแบ่งข้อมูลในซัพพอร์ตเวกเตอร์แมชชีน

สมมติว่าข้อมูลฝึกได้มาจากผู้เขียน 2 คน ข้อมูลแต่ละเรคคอร์ดเปรียบได้กับจุดในปริภูมิหลายมิติ โดยที่แต่ละมิติแทนคุณลักษณะ SVM อาศัยฟังก์ชันเคอร์เนล (Kernel Function) ในการแปลงข้อมูลให้อยู่ในปริภูมิที่เหมาะสม จากนั้นจึงปรับพารามิเตอร์จนได้ระนาบเกิน (Hyperplane) ที่ดีที่สุดในการแยกสมาชิกทั้ง 2 คลาสออกจากกัน ระนาบที่ดีที่สุดนี้ควรเป็นระนาบเชิงเส้นที่ทำให้เกิดระยะขอบกว้างที่สุดดังรูปที่ 1 วิธีฝึก SVM ที่นิยมใช้คือ Sequential Minimal Optimization (SMO) [9]

สำหรับต้นไม้ตัดสินใจ วิธีการสร้างต้นไม้ที่นิยมใช้คือ C4.5 [10] ซึ่งพิจารณาคุณลักษณะทีละคุณลักษณะ เลือกคุณลักษณะที่จำแนกข้อมูลได้ดีที่สุดเป็นโหนดภายใน (Internal Node) การสร้างต้นไม้จะดำเนินไปจนได้โหนดใบที่ประกอบด้วยสมาชิกเพียงคลาสใดคลาสหนึ่ง หรือมีการปนเปื้อนน้อยที่สุด คุณลักษณะที่ไม่มีผลต่อการจำแนกจึงไม่ถูกเลือกเป็นโหนดในต้นไม้ เมื่อจบกระบวนการสร้างต้นไม้แล้วอาจมีการตัดเล็มกิ่งขนาดเล็กออก เพื่อไม่ให้ต้นไม้มีความเฉพาะเจาะจงกับข้อมูลฝึก



รูปที่ 2 ตัวอย่างต้นไม้ตัดสินใจ

มากเกินไป (เกิด Overfitting) จากรูปที่ 2 เรคคอร์ดที่มีค่า P1 <= 0.020333, T5 > 0.024961, และ C6 > 0.05414 จะถูกจำแนกเป็น Author 5

งานวิจัยนี้จึงได้ศึกษาและนำคุณลักษณะที่ de Vel et al. [6] และ Cheng et al. [7] ใช้ในการวิเคราะห์ข้อความภาษาอังกฤษ มาดัดแปลงให้เหมาะสมกับการวิเคราะห์ข้อความออนไลน์ภาษาไทย และเปรียบเทียบผลการระบุตัวผู้เขียนด้วย SVM และต้นไม้ตัดสินใจ เพื่อคัดเลือกวิธีที่เหมาะสมที่จะนำไปพัฒนาเป็นซอฟต์แวร์สำหรับการระบุตัวผู้เขียนข้อความออนไลน์ภาษาไทย โดยผู้ใช้เป้าหมายคือเจ้าหน้าที่ที่สืบสวนอาชญากรรมไซเบอร์

เหตุที่เลือกเปรียบเทียบสองวิธีนี้เนื่องจากงานวิจัยส่วนใหญ่รายงานว่า SVM ให้อัตราความถูกต้องสูงกว่าวิธีอื่นๆ [3]-[7] แต่ต้นแบบของ SVM คือระนาบแบ่งข้อมูลซึ่งเป็นฟังก์ชันคณิตศาสตร์ค่อนข้างซับซ้อน เมื่อนำไปใช้จริง เช่นในการหาตัวผู้เขียนข้อความล่อลวงเหยื่อ เจ้าหน้าที่สืบสวนอาจไม่สามารถอธิบายวิธีจำแนกข้อมูลที่นำไปสู่การชี้ตัวบุคคลใดบุคคลหนึ่งว่า (น่าจะ) เป็นผู้เขียนข้อความได้อย่างละเอียด ในขณะที่ต้นไม้ตัดสินใจเป็นต้นแบบที่ไม่ซับซ้อน สามารถอธิบายวิธีจำแนกข้อมูลได้ง่ายกว่า จึงอาจเป็นอีกทางเลือกหนึ่งที่เหมาะกับการนำไปใช้จริง

2. การสกัดคุณลักษณะจากข้อความออนไลน์

การสกัดคุณลักษณะต่างๆ จากข้อความออนไลน์ในงานวิจัยนี้ใช้โปรแกรมที่พัฒนาขึ้นเอง โดยโมดูลการตัดคำภาษาไทยอาศัยโค้ดจากโปรแกรม LexTo (GNU LGPL 2.1) [11] ซึ่งตัดคำแบบอ้างอิงพจนานุกรม คุณลักษณะที่สกัดได้มีทั้งสิ้น 53 คุณลักษณะ ดังสรุปในตารางที่ 1 ประกอบด้วย คุณลักษณะเกี่ยวกับอักขระเดี่ยว (C1-C9) คุณลักษณะเกี่ยวกับคำ (W1-W10) คุณลักษณะเกี่ยวกับเครื่องหมายวรรคตอน (P1-P10) คุณลักษณะเกี่ยวกับเครื่องหมายแสดงอารมณ์ (E1-E10) คุณลักษณะเกี่ยวกับโครงสร้างข้อความ (S1-S5) และคุณลักษณะเกี่ยวกับเนื้อความ (T1-T9) โดยคุณลักษณะเกี่ยวกับอักขระไทย Emoticon ประเภทต่างๆ และคำประเภทต่างๆ เป็นส่วนที่ทำเพิ่มเติมจาก [6], [7]

ตารางที่ 1 คุณลักษณะที่สกัดจากข้อความออนไลน์

คุณลักษณะ	คำอธิบาย
C1	จำนวนอักขระทั้งหมด
C2	จำนวนอักขระไทย
C3	จำนวนอักขระอังกฤษ A-Z, a-z
C4	จำนวนตัวพิมพ์ใหญ่ A-Z
C5	จำนวนอักขระเลข 0-9
C6	จำนวนช่องว่าง (Whitespace)
C7	จำนวนแท็บ (Tab Space)
C8	จำนวนตัวจบบรรทัด (Line Break)
C9	จำนวนอักขระพิเศษเช่น ! " # $ %
W1	จำนวนคำทั้งหมด
W2	ค่าเฉลี่ยจำนวนอักขระใน 1 คำ
W3	จำนวนคำที่ไม่ซ้ำกัน (Unique Word)
W4	จำนวนคำที่มีอักขระมากกว่า 8 ตัว
W5	จำนวนคำที่มีอักขระน้อยกว่า 3 ตัว
W6	ค่า Yule's Characteristic K
W7	ค่า Simpson's Diversity Index
W8	ค่า Sichel's S
W9	ค่า Honore's Vocabulary Richness
W10	ค่าเอนโทรปี
P1	จำนวนมหัพภาค (.)
P2	จำนวนจุลภาค (,)
P3	จำนวนทวิภาค (:)
P4	จำนวนอัฒภาค (;)
P5	จำนวนยัติภังค์ (-), ยัติภาค (—)
P6	จำนวนปรัศนี (?)
P7	จำนวนอัศเจรีย์ (!)
P8	จำนวนวงเล็บทุกชนิด
P9	จำนวนไม้ยมก
P10	ไม้ยมกอยู่ติดตัวอักษรหรือไม่

ตารางที่ 1 คุณลักษณะที่สกัดจากข้อความออนไลน์ (ต่อ)

คุณลักษณะ	คำอธิบาย
E1	จำนวน ... (ติดกัน 3 ตัวขึ้นไป)
E2	จำนวน ??? (ติดกัน 3 ตัวขึ้นไป)
E3	จำนวน !!! (ติดกัน 3 ตัวขึ้นไป)
E4	จำนวน ๆๆๆ (ติดกัน 3 ตัวขึ้นไป)
E5	จำนวน 555 (ติดกัน 3 ตัวขึ้นไป)
E6	จำนวนคำที่อักขระสุดท้ายติดกัน 3 ตัวขึ้นไป เช่น มาก<u>กกก</u>
E7	จำนวน Emoticon หน้าเอียงมีจมูก เช่น :-)
E8	จำนวน Emoticon หน้าเอียงไม่มีจมูก เช่น :)
E9	จำนวน Emoticon หน้าตรงมีแก้ม เช่น (T_T)
E10	จำนวน Emoticon หน้าตรงไม่มีแก้ม เช่น T_T
S1	จำนวนบรรทัดทั้งหมด
S2	จำนวนบรรทัดว่าง
S3	จำนวนประโยคทั้งหมด
S4	ความยาวเฉลี่ยของบรรทัดที่ไม่ใช่บรรทัดว่าง
S5	ความยาวเฉลี่ยของประโยค
T1	จำนวนคำฟุ่มเฟือย
T2	จำนวนคำลงท้ายที่บอกเพศหญิง
T3	จำนวนคำลงท้ายที่บอกเพศชาย
T4	จำนวนคำเพี้ยนรูป คำสแลง
T5	จำนวนคำสบถ
T6	จำนวนคำอุทาน
T7	จำนวนคำเรียกตัวเองที่บอกเพศหญิง
T8	จำนวนคำเรียกตัวเองที่บอกเพศชาย
T9	จำนวนคำเรียกตัวเองไม่ระบุเพศ

จากตาราง คุณลักษณะ W6-W10 เป็นค่าที่บ่งบอกความหลากหลายในการใช้คำ มีสูตรการคำนวณดังนี้

$$\text{Yule's K} = 10^4\left(-\frac{1}{N}+\sum_{i=1}^{V} V_i\left(\frac{i}{N}\right)^2\right) \quad (1)$$

$$\text{Simpson's D} = \sum_{i=1}^{V} V_i\left(\frac{i}{N}\right)\left(\frac{i-1}{N-1}\right) \quad (2)$$

$$\text{Sichel's S} = \frac{V_2}{V} \quad (3)$$

$$\text{Honore's R} = \frac{100\log_{10} N}{1-\left(\frac{V_1}{V}\right)} \quad (4)$$

$$\text{Entropy} = \sum_{i=1}^{N} V_i\left(-\log_{10}\frac{i}{N}\right)\left(\frac{i}{N}\right) \quad (5)$$

โดย N เป็นจำนวนคำทั้งหมด V เป็นจำนวนคำที่ไม่ซ้ำกัน (Unique Word) V_i เป็นจำนวนคำที่ไม่ซ้ำกันซึ่งปรากฏในข้อความ i ครั้ง ถ้าข้อความชิ้นหนึ่งมีความหลากหลายในการใช้คำมาก ค่า Yule และ Simpson จะต่ำ ในขณะที่ค่า Honore และเอนโทรปีจะสูง

3. วิธีการวิจัย

3.1 การเตรียมข้อมูล

งานวิจัยนี้เลือกข้อความที่โพสต์โดยผู้ใช้ 4 คนในเว็บบอร์ด pantip.com โดยเลือกมาคนละ 50 ข้อความ และใช้ข้อความที่โพสต์โดยผู้ดูแลแฟนเพจอีก 2 แฟนเพจ คนละ 50 ข้อความเช่นเดียวกัน

Author 1 เป็นหญิง โพสต์ข้อความเกี่ยวกับนักแสดงและความงามใน pantip.com

Author 2 เป็นหญิง โพสต์ข้อความเกี่ยวกับนักแสดงใน pantip.com

Author 3 เป็นหญิง โพสต์ข้อความเกี่ยวกับนักแสดงและชีวิตครอบครัวใน pantip.com

Author 4 เป็นชาย โพสต์ข้อความเกี่ยวกับการเมืองใน pantip.com

Author 5 เป็นชาย โพสต์ข้อความล้อเลียนเสียดสีสังคมและการเมืองในแฟนเพจ

Author 6 เป็นชาย โพสต์ข้อความล้อเลียนเสียดสีสังคมและการเมืองในแฟนเพจ

ข้อความทั้งหมดโพสต์ในพื้นที่สาธารณะ บุคคลทั่วไปสามารถอ่านได้ โดยไม่จำเป็นต้องล็อกอินเข้าเว็บบอร์ดและเฟซบุ๊ก และข้อความนั้นๆ ไม่ใช่การคัดลอกคำพูดของบุคคลอื่น ช่วงเวลาที่โพสต์ข้อความตั้งแต่ พ.ศ. 2551-2557 ผู้วิจัยกำหนดนามสมมติเช่น Author 1 แทนชื่อล็อกอิน ไม่มีการเก็บ IP Address ชื่อแฟนเพจ และข้อมูลส่วนตัวใดๆ ของผู้เขียนข้อความ

ข้อความที่รวบรวมจากผู้เขียนแต่ละคน แบ่งออกเป็นข้อความสั้นและข้อความยาวอย่างละ 25 ข้อความ

ความยาวเฉลี่ยของข้อความแสดงในตารางที่ 2 จากการสำรวจเบื้องต้นพบว่า ถ้าข้อความสั้นเกินไป (สั้นกว่า 150 คำโดยประมาณ) คุณลักษณะส่วนใหญ่จะมีค่าเป็นศูนย์ ทำให้วิเคราะห์รูปแบบลีลาการเขียนไม่ได้มากนัก แต่ถ้ายาวเกินไป (ยาวกว่า 300 คำโดยประมาณ) ในเนื้อหามักมีการคัดลอกคำพูดของบุคคลอื่นมาด้วย รูปแบบลีลาการเขียนที่ปรากฏจึงไม่ใช่ของเจ้าตัวอย่างแท้จริง

ตารางที่ 2 ความยาวเฉลี่ยของข้อความ

ผู้เขียน	ความยาวเฉลี่ย (คำ)	
	ข้อความสั้น	ข้อความยาว
Author 1	165	369
Author 2	131	293
Author 3	142	291
Author 4	146	299
Author 5	151	319
Author 6	133	303
เฉลี่ย	144	312

เมื่อสกัดคุณลักษณะจากข้อความแต่ละชุด จนได้เซตข้อมูลที่ประกอบด้วย 150 เรคคอร์ด การทดสอบการระบุตัวผู้เขียนด้วยเซตข้อมูลหนึ่งๆ จะทำซ้ำ 10 รอบ ในแต่ละรอบ สุ่มข้อมูล 20 เรคคอร์ดจากผู้เขียนแต่ละคนเป็นข้อมูลฝึก (รวมข้อมูลฝึกทั้งสิ้น 120 เรคคอร์ด) ส่วนที่เหลือเป็นข้อมูลทดสอบ

3.2 เครื่องมือวิจัย

งานวิจัยนี้ใช้โมดูล SMO และ J48 ของโปรแกรม Weka [12] ในการจำแนกด้วย SVM และต้นไม้ตัดสินใจแบบ C4.5 ตามลำดับ โดยกำหนดพารามิเตอร์ดังนี้

1. SMO: Normalize ข้อมูล (Filter Type = Normalize Training Data) และใช้เคอร์เนล Polynomial กำลังสอง ซึ่งเป็นเคอร์เนลพื้นฐานสำหรับการจำแนก

2. J48: เนื่องจากข้อมูลฝึกมีจำนวนน้อย จึงแตกโหนดแบบ Binary Split ปรับขนาดของโหนดใบด้วยฟังก์ชัน Laplace และเล็มกิ่งแบบ Subtree Raising แต่ไม่ใช้ Reduced Error Pruning (ไม่มีการกันข้อมูลฝึกส่วนหนึ่งไว้สำหรับการเล็มกิ่ง)

เมื่อใช้ตัวจำแนกแต่ละตัวจำแนกข้อมูลทดสอบ จะวัดอัตราความถูกต้องโดยรวม และคำนวณค่า F-Measure ในการระบุตัวผู้เขียนแต่ละคน (ผู้เขียน X) ดังนี้

$$\text{อัตราความถูกต้อง} = \frac{\text{จำนวนเรคคอร์ดที่จำแนกถูก}}{\text{จำนวนเรคคอร์ดทั้งหมด}} \quad (6)$$

$$\text{F-Measure(X)} = \frac{2 \times \text{precision(X)} \times \text{recall(X)}}{\text{precision(X)} + \text{recall(X)}} \quad (7)$$

โดยอัตราเรียกคืนหรือ Recall(X) เป็นสัดส่วนของข้อความที่ X เขียนและตัวจำแนกระบุว่าเป็นของ X ได้ถูกต้อง ส่วนอัตราเที่ยงหรือ Precision(X) เป็นสัดส่วนของข้อความที่ตัวจำแนกระบุว่า X เขียนและ X เป็นผู้เขียนจริงๆ

4. ผลการทดลอง

4.1 อัตราความถูกต้องโดยรวม

รูปที่ 3 และ 4 แสดงกราฟเปรียบเทียบอัตราความถูกต้องในการระบุตัวผู้เขียนด้วย SVM และต้นไม้ตัดสินใจ เมื่อทดสอบกับข้อมูลแต่ละชุด ในภาพรวมพบว่า SVM มีอัตราความถูกต้องเฉลี่ยสูงกว่าต้นไม้ตัดสินใจ และเมื่อทดสอบกับข้อความขนาดยาวได้อัตราความถูกต้องเฉลี่ยสูงกว่าเมื่อทดสอบกับข้อความขนาดสั้น

4.1.1 จากการทดสอบกับข้อมูลแต่ละชุด

H_0 : อัตราความถูกต้องเฉลี่ยของ SVM และต้นไม้ตัดสินใจไม่แตกต่างกัน

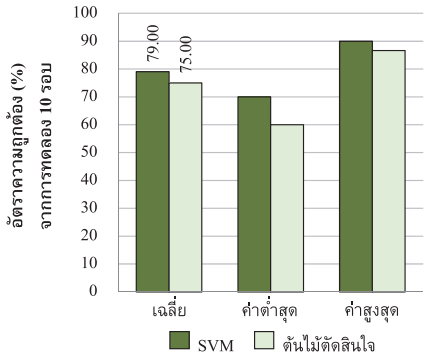
H_1 : อัตราความถูกต้องเฉลี่ยของทั้งสองวิธีแตกต่างกัน

โดยกำหนด $\alpha = 0.05$ และใช้สถิติ Paired Samples T-test พบว่าในกรณีข้อความสั้น อัตราความถูกต้องเฉลี่ยของทั้งสองวิธีไม่แตกต่างกันที่นัยสำคัญ 0.05 ($t = 1.857$, $df = 9$, p-value = 0.096) แต่ในกรณีของข้อความยาว อัตราความถูกต้องเฉลี่ยของทั้งสองวิธีแตกต่างกันอย่างมีนัยสำคัญ ($t = 3.139$, $df = 9$, p-value = 0.012)

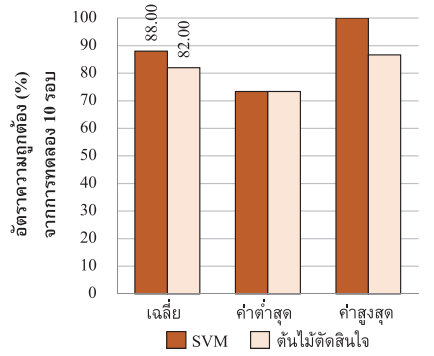
4.1.2 จากการใช้ตัวจำแนกแต่ละตัว

H_0 : อัตราความถูกต้องเฉลี่ยเมื่อทดสอบกับข้อความสั้นและข้อความยาวไม่แตกต่างกัน

H_1 : อัตราความถูกต้องเฉลี่ยเมื่อทดสอบกับ



รูปที่ 3 กราฟเปรียบเทียบอัตราความถูกต้องในการระบุตัวผู้เขียนข้อความสั้น



รูปที่ 4 กราฟเปรียบเทียบอัตราความถูกต้องในการระบุตัวผู้เขียนข้อความยาว

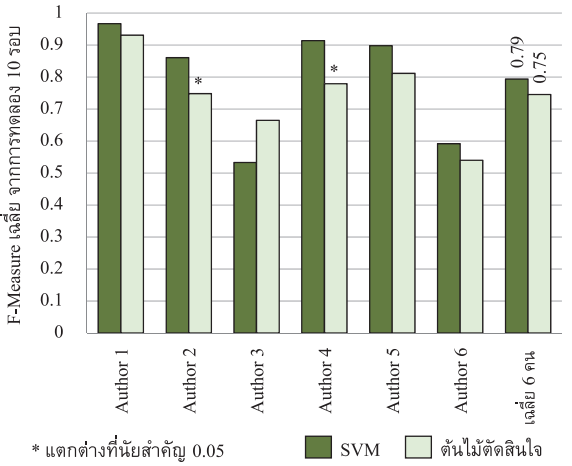
ข้อความสั้นและข้อความยาวแตกต่างกัน

โดย $\alpha = 0.05$ และใช้สถิติ Independent Samples T-test พบว่าอัตราความถูกต้องเฉลี่ยเมื่อทดสอบกับข้อความสั้นและข้อความยาวแตกต่างกันอย่างมีนัยสำคัญ ไม่ว่าจะใช้ตัวจำแนกใด (สำหรับ SVM ได้ค่า $t = -3.819$, $df = 18$, p-value = 0.005 สำหรับต้นไม้ตัดสินใจ ได้ค่า $t = -2.596$, $df = 18$, p-value = 0.018)

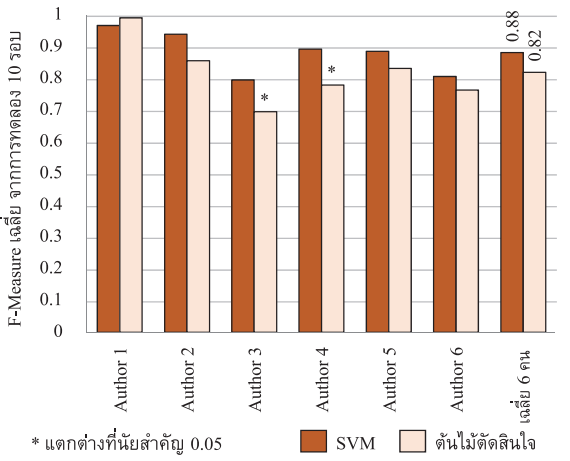
4.2 ผลการระบุตัวผู้เขียนแต่ละคน

สมมติฐานสำหรับผู้เขียน X และเซตข้อมูล Y

H_0 : F-Measure เฉลี่ยของ SVM และต้นไม้ตัดสินใจไม่แตกต่างกัน



รูปที่ 5 กราฟเปรียบเทียบค่า F-Measure เฉลี่ยในการระบุตัวผู้เขียนข้อความสั้นแต่ละคน



รูปที่ 6 กราฟเปรียบเทียบค่า F-Measure เฉลี่ยในการระบุตัวผู้เขียนข้อความยาวแต่ละคน

H_1 : F-Measure เฉลี่ยของทั้งสองวิธีแตกต่างกัน

โดย $\alpha = 0.05$ และใช้สถิติ Paired Samples T-test ในการทดสอบสมมติฐานแต่ละชุด

รูปที่ 5 และรูปที่ 6 แสดงกราฟเปรียบเทียบค่า F-Measure เฉลี่ยในการระบุตัวผู้เขียนแต่ละคน จากผลการทดสอบสมมติฐานพบว่า ในกรณีข้อความสั้น การระบุตัวผู้เขียนคนที่ 2 และ 4 ด้วย SVM ให้ค่า F-Measure เฉลี่ยแตกต่างจาก (สูงกว่า) เมื่อระบุตัวผู้เขียนคนดังกล่าวด้วย

ต้นไม้ตัดสินใจ อย่างมีนัยสำคัญ

ส่วนในกรณีข้อความยาว การระบุตัวผู้เขียนคนที่ 3 และ 4 ด้วย SVM ให้ค่า F-Measure เฉลี่ยแตกต่างจาก (สูงกว่า) การระบุตัวผู้เขียนคนดังกล่าวด้วยต้นไม้ตัดสินใจอย่างมีนัยสำคัญเช่นเดียวกัน

เมื่อเปรียบเทียบรูปแบบการเขียนที่สังเกตด้วยตากับผลการระบุตัวผู้เขียนข้างต้น มีข้อสังเกตคือ

1. ถ้าผู้เขียนมีรูปแบบการเขียนเป็นเอกลักษณ์เฉพาะตัวเช่น ผู้เขียนคนที่ 1 และ 5 มีการใช้คำอุทาน คำสบถ Emoticon และสัญลักษณ์พิเศษต่าง ๆ แตกต่างจากคนอื่นชัดเจน การระบุตัวสามารถทำได้ง่ายไม่ว่าจะใช้ SVM หรือต้นไม้ตัดสินใจ

2. ถ้าผู้เขียนมีรูปแบบการเขียนไม่แตกต่างจากคนอื่นมากนัก เช่น ผู้เขียนคนที่ 2, 3, และ 4 การระบุตัวด้วย SVM ให้ผลลัพธ์ดีกว่าต้นไม้ตัดสินใจ

3. เมื่อข้อความตัวอย่างมีขนาดสั้น คุณลักษณะที่สกัดได้จำนวนหนึ่งมีค่าเป็น 0 จึงไม่มีผลต่อการจำแนก เมื่อข้อความตัวอย่างยาวขึ้น คุณลักษณะที่มีค่าไม่เป็น 0 มีจำนวนมากขึ้นและมีผลต่อการจำแนกมากขึ้น ทำให้ทั้ง SVM และต้นไม้ตัดสินใจระบุตัวผู้เขียนได้ดีขึ้น

5. อภิปรายผลและสรุป

5.1 เปรียบเทียบผลลัพธ์กับงานวิจัยอื่น

ในงานวิจัยนี้ พบว่า SVM สามารถระบุตัวผู้เขียนได้ดีกว่าต้นไม้ตัดสินใจ โดยเฉพาะเมื่อผู้เขียนมีรูปแบบลีลาการเขียนคล้ายคลึงกัน (จำแนกข้อมูลได้ยาก) ผลลัพธ์ที่ได้สอดคล้องกับงานวิจัยอื่นเช่น [3]-[7] ซึ่งแสดงในตารางที่ 3 โดยงานวิจัย [4]-[6] จำแนกผู้เขียนเพียงไม่กี่คน ส่วน [7] จำแนกเพียงเพศของผู้เขียนเท่านั้น แต่มีข้อสังเกตคือ Sun et al. [3] สามารถจำแนกผู้เขียน 20 คน ด้วยอัตราความถูกต้องเฉลี่ยสูงถึง 92.96% งานวิจัยดังกล่าววิเคราะห์ชุดอักขระที่อยู่ติดกัน 1-5 ตัว (Character n-gram, $1 \leq n \leq 5$) สกัดคุณลักษณะได้มากกว่าหนึ่งหมื่นคุณลักษณะ จึงวิเคราะห์ข้อมูลได้ละเอียดกว่างานวิจัยนี้มาก

นอกจากนี้ Luyckx [13] ยังทดสอบการหาตัวผู้เขียนจากเซตของผู้เขียนขนาดต่าง ๆ พบว่าเมื่อเซตมีขนาดใหญ่ขึ้นเช่นจาก 2 คนเป็น 5 คน อัตราความถูกต้องเฉลี่ยลดลงถึง 18.25%

ดังนั้น ในทางปฏิบัติที่มีผู้โพสต์ข้อความในเว็บบอร์ดเป็นจำนวนมาก เจ้าหน้าที่สืบสวนควรคัดกรองให้เหลือผู้ต้องสงสัยเพียงไม่กี่คนก่อน แล้วจึงวิเคราะห์หาตัวผู้เขียนจากกลุ่มผู้ต้องสงสัยนั้น จะช่วยให้ผลลัพธ์มีความแม่นยำมากขึ้น วิธีการที่เหมาะสมในการคัดกรองผู้ต้องสงสัยเป็นอีกหัวข้อหนึ่งที่ควรมีการวิจัยต่อในอนาคต

ตารางที่ 3 ผลลัพธ์จากงานวิจัยอื่น

งานวิจัย	ตัววัด	การจำแนก
[3]	อัตราความถูกต้องเฉลี่ย 92.96% (SVM)	ผู้เขียน 20 คน
[4]	อัตราความถูกต้องเฉลี่ย 91.43% (SVM) 77.14% (ต้นไม้ตัดสินใจ)	 ผู้เขียน 3 คน ผู้เขียน 3 คน
[5]	F-Measure เฉลี่ย 77.28% (SVM)	ผู้เขียน 7 คน
[6]	F-Measure เฉลี่ย 85.57% (SVM)	ผู้เขียน 3 คน
[7]	อัตราความถูกต้องเฉลี่ย 85.13% (SVM)	เพศของผู้เขียน (ชายหรือหญิง)
[13]	อัตราความถูกต้องเฉลี่ย 90.45% (SVM) 72.20% (SVM) 64.71% (SVM)	 ผู้เขียน 2 คน ผู้เขียน 5 คน ผู้เขียน 13 คน

5.2 จุดเด่นและข้อจำกัดของต้นไม้ตัดสินใจ

จุดเด่นของต้นไม้ตัดสินใจคือ ต้นแบบการจำแนกมีความเป็นมิตรกับผู้ใช้งาน (User Friendliness) เจ้าหน้าที่สืบสวนสามารถทำความเข้าใจและอธิบายถึงที่มาของการระบุตัวผู้เขียนได้ง่าย แต่การระบุตัวผู้เขียนด้วยวิธีนี้ยังให้อัตราความถูกต้องและ F-Measure ต่ำกว่า SVM ดังนั้น การนำต้นไม้ตัดสินใจไปใช้จริง ควรมีปัจจัยอื่นเป็นตัวช่วย เช่น ข้อความตัวอย่างต้องยาวพอสมควร และอาจใช้ป่าตัดสินใจ (Decision Forest) ซึ่งประกอบด้วยต้นไม้ตัดสินใจหลายต้นช่วยกันจำแนกข้อมูล [14] เพื่อให้ผลลัพธ์มีความแม่นยำน่าเชื่อถือมากขึ้น

สำหรับแผนงานในอนาคตแบ่งเป็น 2 ส่วน ส่วนแรกได้แก่ การสกัดคุณลักษณะในการเขียนข้อความออนไลน์ภาษาไทยเพิ่มเติม เพื่อวิเคราะห์รูปแบบลีลาการเขียนให้ละเอียดขึ้น ส่วนที่สองได้แก่ การศึกษาและทดสอบวิธีการระบุตัวผู้เขียนแบบอื่นนอกเหนือจากการจำแนก เช่นใช้หลักการจัดกลุ่ม (Clustering) และคำนวณคะแนนความคล้ายกันระหว่างข้อความนิรนามกับกลุ่มข้อความตัวอย่างของผู้ต้องสงสัยแต่ละคน วิธีนี้จะเปิดโอกาสให้เจ้าหน้าที่สืบสวนเลือกผู้ต้องสงสัยมากกว่าหนึ่งคนที่ได้คะแนนสูงพอๆ กันไปวิเคราะห์ต่อได้ ทำให้การสืบหาตัวผู้เขียนมีความยืดหยุ่นมากขึ้น

6. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยจากสำนักงานคณะกรรมการวิจัยแห่งชาติ ประจำปีงบประมาณ 2556 โปรแกรมสกัดคุณลักษณะจากข้อความออนไลน์พัฒนาโดยนายทัพพสาร อร่ามบุญพงศ์ นายรัฐนันท์ นลินทัศไนย และนายรอบรู้ สมเกียรติเจริญ

เอกสารอ้างอิง

[1] National Statistical Office (Thailand), *The 2013 Information and Communication Technology Survey in Household*, ISSN 1686-4212.

[2] L. Pearl and M. Steyvers, "Detecting authorship deception: a supervised machine learning approach using author writeprints," *Literary and Linguistic Computing*, vol. 7, no. 2, pp. 183-196, 2012.

[3] J. Sun, Z. Yang, and S. Liu, "Applying stylometric analysis techniques to counter anonymity in cyberspace," *Journal of Networks*, vol. 7, no. 2, pp. 259-266, 2012.

[4] R. Zheng, Y. Qin, Z. Huang, and H. Chen, "Authorship analysis in cybercrime investigation," in *ISI 2003, LNCS 2665*, H. Chen et al. (Eds). Springer-Verlag, 2003, pp. 59-73.

[5] J. Diederic, J. Kindermann, E. Leopol, and G. Paass, "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, no. 1-2, pp. 109-123, 2003.

[6] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining email content for author identification forensics," *SIGMOD Record*, vol. 30, no. 4, pp.55-64, 2001.

[7] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp.78-88, 2011.

[8] S. Bedrick, R. Beckley, B. Roark, and R. Sproat, "Robust kaomoji detection in Twitter," in *Proc. of the 2012 Workshop on Language in Social Media*, Montreal, Canada, 2012, pp. 56-64.

[9] J. Platt, "Fast training of support vector machine using sequential minimal optimization," in *Advances in Kernel Methods – Support Vector Learning*, B. Schoelkopf et al. (Eds), 1998, pp. 185-208.

[10] R. Quinlan, *C4.5: programs for machine learning,* Morgan Kaufmann Publishers, 1993.

[11] National Electronics and Computer Technology Center (Thailand), LexTo: Thai Lexeme Tokenizer. [Online]. Available: http://www.sansarn.com/lexto/.

[12] University of Waikato. New Zealand. Weka 3: data mining software in Java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka.

[13] K. Luyckx, "Scalability Issues in Authorship Attribution," PhD. Thesis, Universiteit Antwerpen, Belgium, 2010.

[14] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Journal of Pattern Recognition*, vol. 44, no. 2, pp. 330-349, 2011.