Science and Technology
RMUTT Journal
Vol.6 No.1 (2016) : 83-93
www.sci.rmutt.ac.th/stj
Online ISSN 2229-1547

การจำแนกความคิดเห็นทางการเมืองบนเครือข่ายสังคมออนไลน์ โดยใช้วิธีการจำแนกแบบความสัมพันธ์

Opinion Classification of Politics on Social Network using Associative Classification

พนิดา ทรงรัมย์

หน่วยวิจัยการประมวลชั้นสูงสำหรับงานด้านปัญญาประดิษฐ์ การประมวลผลภาพและหุ่นยนต์ (POLAR)
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
E-mail: panida.s@msu.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอการจำแนกความคิดเห็นทางการเมืองในช่วงที่มีการปฏิวัติในประเทศไทยด้วยวิธีการจำแนกแบบความสัมพันธ์ (Associative Classification) โดยจำแนกความคิดเห็นจากข้อความคิดเห็นบนเฟสบุ๊คที่ถูกเขียนขึ้นด้วยภาษาไทยซึ่งเป็นภาษาที่มีความซับซ้อน งานวิจัยนี้ได้ทำการสกัดคุณลักษณะของข้อความโดยใช้คำเชิงบวกและคำเชิงลบที่รวบรวมจากข้อความที่อยู่บนเครือข่ายสังคมออนไลน์ และทำการจำแนกความคิดเห็นโดยพิจารณาความสัมพันธ์ของคุณลักษณะด้วยวิธีการจำแนกแบบความสัมพันธ์ ทำให้ได้กฎที่ใช้ในการจำแนกความคิดเห็นที่สามารถบ่งบอกถึงความเชื่อมั่นของความคิดเห็นในเชิงบวกหรือเชิงลบ ผลการทดลองแสดงให้เห็นว่าวิธีการจำแนกแบบความสัมพันธ์สามารถจำแนกความคิดเห็นทางการเมืองได้ถูกต้องถึง 77.75%

คำสำคัญ: การจำแนกความคิดเห็น การจำแนกแบบความสัมพันธ์ เหมืองความคิดเห็น เครือข่ายสังคมออนไลน์

Received: September 19, 2015
Revised: February 04, 2016
Accepted: February 29, 2016

Abstract

This paper presents opinion classification of politics during the government revolution in Thailand using associative classification. The opinions are classified from Facebook statuses written in Thai which are complex. Features of the statuses are extracted by using positive and negative words that are collected from social networking websites. Using feature association based on associative classification leads to the resulting rules for opinion classification with specifying the confidence of either positive or negative opinion. The experimental results show that associative classification can give accuracy to 77.75% for political opinion classification.

Keywords: Opinion classification, Associative classification, Opinion mining, Social network

1. บทนำ

เว็บไซต์เครือข่ายสังคมออนไลน์ (Social Networking Websites) เป็นเว็บไซต์ที่ใช้ในการติดต่อสื่อสารระหว่างบุคคลทั่วโลก ทำให้สามารถพูดคุย แบ่งปัน แลกเปลี่ยนแนวคิด ความรู้ หรือเรื่องที่ตัวเองสนใจได้อย่างไม่มีขีดจำกัด เช่น การเมือง วัฒนธรรม เทคโนโลยี สินค้า และการศึกษา เป็นต้น [1] ปัจจุบันมีเว็บไซต์เครือข่ายสังคมออนไลน์จำนวนมากถูกสร้างขึ้นเพื่อใช้ในการแลกเปลี่ยนและแบ่งปันข้อมูลกัน เช่น เฟสบุ๊ค (Facebook), ทวิตเตอร์ (Twitter) และ ลิงค์อิน (LinkedIn) เป็นต้น เฟสบุ๊คเป็นเว็บไซต์หนึ่งที่ได้รับความนิยมมากที่สุดในโลก โดยมีผู้เข้าใช้เฟสบุ๊ค ประมาณ 900,000,000 ต่อเดือน [2] เนื่องจากมีเครื่องมือหลากหลายที่ช่วยอำนวยความสะดวกในการแลกเปลี่ยนข้อมูลกัน เช่น การแช็ต อัพโหลดรูป ส่งไฟล์ การสร้างกลุ่ม การโพสข้อความ เป็นต้น ทำให้ข้อมูลที่อยู่บนเฟสบุ๊คมีความหลากหลายและมีจำนวนมากมายมหาศาล นอกจากนี้ข้อมูลดังกล่าวยังเป็นข้อมูลที่ผู้ใช้แสดงออกได้อย่างอิสระและทุกคนสามารถเข้าใช้งานได้ไม่ว่าจะอยู่ประเทศใดก็ตาม [3] ทำให้ข้อมูลเหล่านี้มีประโยชน์เป็นอย่างมากทางด้านการตลาด ด้านสังคม ด้านการศึกษา ด้านการเมืองและด้านอื่นๆ

นอกจากเว็บไซต์เฟสบุ๊คจะถูกนำมาใช้เป็นเครื่องมือที่ใช้ในการติดต่อสื่อสารแล้ว เฟสบุ๊คยังเป็นแหล่งแสดงความรู้สึก ความคิดเห็นของผู้ใช้ คนส่วนใหญ่มักแสดงความคิดเห็นหรือความรู้สึกผ่าน Status ของตัวเองบนเฟสบุ๊ค ซึ่ง Status ดังกล่าวเป็นข้อความสั้นๆแต่แสดงออกถึงความรู้สึก หรือความคิดเห็นของผู้ใช้ ทำให้มีนักวิจัยจำนวนมากสนใจที่จะนำข้อความคิดเห็นบนเฟสบุ๊คมาวิเคราะห์โดยใช้วิธีการที่เรียกว่าการทำเหมืองความคิดเห็น (Opinion Mining) [4] เช่น Akaichi และคณะ [5, 6] ได้นำเสนอการจำแนกความคิดเห็นของผู้ใช้งานเฟสบุ๊คจาก Status โดยใช้วิธี Naïve Bayes [7] และ SVM (Support Vector Machine) [8] จุดประสงค์ของงานวิจัยนี้ก็คือ เพื่อวิเคราะห์พฤติกรรมและความคิดเห็นของผู้ใช้งานบนเฟสบุ๊ค จำนวน 260 คนในประเทศ Tunisia ในช่วงที่มีการปฏิวัติ โดยรวบรวมคลังคำศัพท์ 3 ชนิดจากเครือข่ายสังคมออนไลน์เพื่อใช้ในการสกัดคุณลักษณะ คลังคำศัพท์ตัวแรกสำหรับเก็บตัวย่อที่แสดงถึงความรู้สึก เช่น lol, gr8, cu เป็นต้น คลังคำศัพท์ที่สองสำหรับเก็บ

สัญลักษณ์แสดงอารมณ์ เช่น ;(, :D ;), เป็นต้น คลังคำศัพท์ที่สามสำหรับเก็บคำอุทาน เช่น Wow, Haha, Oh dear เป็นต้น จากนั้นทำการพิจารณาว่าคำศัพท์แต่ละคำสื่อความหมายเชิงบวกหรือเชิงลบ การเตรียมข้อมูลในงานวิจัยนี้เริ่มจากการขจัดคำหยุด การหารากคำ การแทนคำที่พบด้วย 1 และคำที่ไม่พบด้วย 0 ใช้รูปแบบ n-gram และชนิดของคำ (Part-of-speed Tags) ในการสกัดคุณลักษณะ ทำการกำหนดค่า n-gram จำนวน 7 รูปแบบเพื่อหาค่า n-gram ที่เหมาะสมในการจำแนกข้อมูล ได้แก่ 1) unigram 2) bigram 3) trigram 4) unigram รวมกับ bigram 5) unigram รวมกับ trigram 6) bigram รวมกับ trigram และ 7) การรวมกันระหว่าง unigram bigram และ trigram ผลจากการทดลองแสดงให้เห็นว่า Naïve Bayes มีความถูกต้องสูงสุด 69.42% เมื่อใช้ bigram กำหนดคุณลักษณะ ในขณะที่ SVM ให้ความถูกต้องสูงกว่า Naïve Bayes เมื่อใช้ unigram เป็นคุณลักษณะ โดยมีค่าความถูกต้องถึง 72.74% งานวิจัยนี้พยายามสกัดคุณลักษณะหลายๆวิธี แต่ค่าความถูกต้องที่ได้มีค่าไม่สูงมากนัก

Shrivatava และPant [3] ได้พัฒนาตัวจำแนกความคิดเห็นจาก Status ที่เขียนลงในเว็บไซต์เฟสบุ๊ค เพื่อจำแนกความคิดเห็น 3 ขั้ว คือ GOOD BAD และ AVERAGE โดยพัฒนาโปรแกรม Facebook puller เพื่อทำการรวบรวมข้อมูลจาก Status ของผู้ใช้เฟสบุ๊คจำนวน 2,000 รายการ หลังจากนั้นทำการจำแนกข้อมูลโดยใช้พจนานุกรมที่บรรจุคำพ้องเสียง และโปรแกรม LIBSVM [5] ถูกนำมาใช้เพื่อสร้างตัวแบบและทดสอบตัวแบบ ผลจากการทดสอบแสดงให้เห็นว่าค่าเฉลี่ยของความถูกต้องในการจำแนกความคิดเห็นของผู้ใช้เฟสบุ๊คอยู่ที่ 70.5% งานวิจัยนี้พัฒนาโปรแกรมที่สามารถดึงข้อความคิดเห็นด้านต่างๆได้ อัตโนมัติตามที่ผู้ใช้ระบุคำค้น และสามารถจำแนกข้อความคิดเห็นออกเป็น 3 ขั้ว ทำให้ง่าย และสะดวกในการค้นหาคุณลักษณะแต่ละขั้ว แต่อย่างไรก็ตาม งานวิจัยนี้เสียเวลาในการตรวจสอบคำพ้องเสียง

Ortigosa และคณะ [9] นำเสนอวิธีการสำหรับวิเคราะห์ความคิดเห็นของผู้ใช้เฟสบุ๊คในประเทศสเปน งานวิจัยนี้ได้นำเสนอวิธีการที่ชื่อว่า SenBuk เพื่อดึงข้อมูลที่อยู่บน Status บนเฟสบุ๊ค และทำการจำแนกความนึกคิดของผู้ใช้เฟสบุ๊คด้วยวิธีการที่เกิดจากการรวมกันระหว่างวิธีการที่ใช้พจนานุกรม (Lexicon-based) และการเรียนรู้ของเครื่อง (Machine Learning) โดยได้สร้างรายการของอารมณ์ (List of emotion) ที่รวบรวมข้อมูลจาก Wikipedia เพื่อช่วยในการจำแนก วิธีการที่นำเสนอในงานวิจัยนี้มีทั้งหมด 4 วิธี คือ 1) การใช้พจนานุกรมอย่างเดียว (Lexicon-based approach) 2) การใช้ Decision tree ร่วมกับ Lexicon-based Tagging 3) การใช้ Naïve Bayes ร่วมกับ Lexicon-based Tagging และ 4) การ SVM ร่วมกับ Lexicon-based Tagging จากการทดลองในงานวิจัยนี้แสดงให้เห็นว่าการจำแนกข้อมูลด้วย SVM ร่วมกับ Lexicon-based Tagging ให้ค่าความถูกต้องสูงที่สุดคือ 83.27% ซึ่งแสดงให้เห็นว่าการใช้พจนานุกรมร่วมกับการเรียนรู้ของเครื่องในการจำแนกความคิดเห็นจะให้ค่าความถูกต้องที่สูง

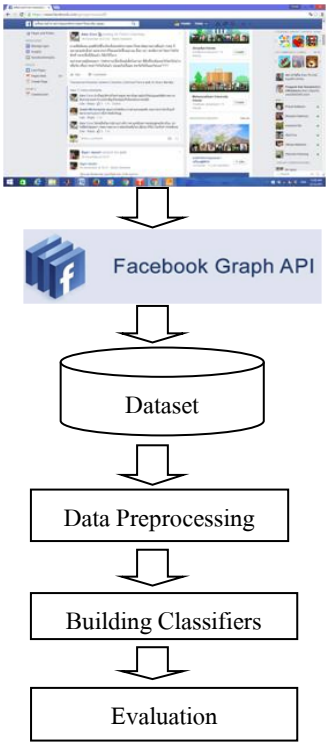
Keeshin และคณะ [10] ได้เสนองานวิจัยเพื่อจำแนกเพศของผู้ใช้งานเฟสบุ๊คจาก Status งานวิจัยนี้ได้ทดลองวิธีการที่เหมาะสมในการจำแนกเพศของผู้ใช้เฟสบุ๊คจากข้อความทั้งหมด 170,000 ข้อความ ซึ่งรวบรวมโดยใช้ Facebook Graph API จากการทดลองแสดงให้เห็นว่าข้อความที่ไม่ได้สกัดคุณลักษณะให้ค่าความถูกต้องแค่ 62% เมื่อทดสอบกับโปรแกรม MaxEnt จากนั้นงานวิจัยนี้จึงได้การพัฒนาอัลกอริทึมขึ้นมาเองเพื่อจำแนกเพศผู้ใช้งานเฟสบุ๊ค จากผลการทดลองแสดงให้เห็น

ว่าอัลกอริทึมให้ความถูกต้องเพียง 53.2% บนข้อมูลที่ไม่ได้มีการเตรียมข้อมูล จึงได้ทำการปรับค่า Learning Rate และค่าน้ำหนักเพื่อให้ได้ความถูกต้องมากขึ้น แต่อย่างไรก็ตามความถูกต้องเพิ่มขึ้นแค่ 56.6% ซึ่งเป็นค่าที่น้อยมาก จึงได้ทำการทดลองจำแนกเพศโดยใช้ Window algorithm และ Naïve Bayes ผลจากการทดลองแสดงให้เห็นว่า Naïve Bayes ให้ค่าความถูกต้องมากที่สุด 67.7% ในการจำแนกเพศผู้ใช้งานเฟสบุ๊ค งานวิจัยนี้ให้ค่าความถูกต้องที่ไม่สูง เนื่องจากผู้หญิงและผู้ชายเขียนข้อความคิดเห็นคล้ายคลึงกัน ทำให้ยากในการจำแนก

Irfan และคณะ [11] ได้ทำการศึกษาขั้นตอนการเตรียมข้อมูลและอัลกอริทึมสำหรับการทำเหมืองข้อความบนเครือข่ายออนไลน์ โดยแบ่งการเตรียมข้อมูลออกเป็น 3 ขั้นตอน คือ การสกัดคุณลักษณะ การคัดเลือกคุณลักษณะ และการแทนเอกสาร สำหรับการสกัดคุณลักษณะแบ่งออกเป็น 3 แบบ คือ 1) Morphological Analysis เป็นการสกัดคุณลักษณะโดยใช้คำซึ่งประกอบไปด้วยการขจัดคำหยุด การหารากคำ เป็นต้น 2) Syntactical Analysis เป็นการสกัดข้อมูลโดยใช้โครงสร้างไวยากรณ์ภาษา และ 3) Syntactical Analysis เป็นการสกัดคุณลักษณะโดยใช้ชนิดของคำ และการตรวจสอบโครงสร้างของประโยค Parsing ส่วนการคัดเลือกคุณลักษณะถูกนำมาใช้เพื่อลดความซ้ำซ้อนของข้อมูลและทำให้เวลาในการประมวลผลน้อยลง ซึ่งการคัดเลือกคุณลักษณะส่วนใหญ่จะใช้วิธีให้ค่าน้ำหนักของคำและคัดเลือกคำที่มีค่าน้ำหนักมากกว่าค่าแบ่งเกณฑ์ ซึ่งค่าน้ำหนักสามารถคำนวณได้จากความถี่ Latent Semantic Indexing และ Random Mapping ส่วนขั้นตอนสุดท้ายคือการแทนข้อมูล โดยส่วนใหญ่จะแทนในรูปแบบของเวกเตอร์ (Vector Space) งานวิจัยนี้ได้สรุปแบ่งอัลกอริทึมที่ใช้ในการทำเหมืองข้อความคิดเห็นออกเป็น 3 กลุ่ม คือ อัลกอริทึมที่อยู่บนพื้นฐานการเรียนรู้ของเครื่อง อัลกอริทึมที่อยู่บนพื้นฐานออนโทโลยี (Ontology) และอัลกอริทึมที่เกิดจากการรวมกันของหลายอัลกอริทึม (Hybrid Approach) งานวิจัยนี้แสดงให้เห็นว่าอัลกอริทึมที่เกิดจากการรวมกันหลายอัลกอริทึมให้ค่าความถูกต้องมากกว่าการใช้อัลกอริทึมเดี่ยวๆ

จากงานวิจัยที่เกี่ยวข้องมีการศึกษาอย่างแพร่หลายในการวิเคราะห์ความคิดเห็นด้านต่างๆ จากข้อความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์เฟสบุ๊ค โดยมีการวิเคราะห์ข้อความคิดเห็นจากหลายภาษา งานวิจัยนี้นำเสนอการวิเคราะห์ความคิดเห็นทางการเมืองจากข้อความภาษาไทย ซึ่งมีความซับซ้อน แตกต่างจากภาษาอื่น และใช้ข้อความคิดเห็นทางด้านการเมืองในช่วงที่มีการรัฐประหารในประเทศไทย ซึ่งข้อความคิดเห็นส่วนใหญ่เป็นข้อความเชิงเสียดสีทำให้เกิดความยากในการจำแนกความคิดเห็นมากกว่าข้อความคิดเห็นทางด้านอื่น ดังจะเห็นได้จากงานวิจัยของ Akaichi และคณะ [5,6] สามารถจำแนกความคิดเห็นทางการเมืองได้ค่าความถูกต้องเพียง 72.74% ดังนั้นงานวิจัยนี้จึงนำเสนอการจำแนกความคิดเห็นโดยใช้ความสัมพันธ์ของคุณลักษณะเพื่อเพิ่มประสิทธิภาพความถูกต้องในการจำแนกความคิดเห็นทางด้านการเมืองที่อยู่บนเครือข่ายสังคมออนไลน์ และทำการสกัดคุณลักษณะของข้อความโดยใช้พจนานุกรมที่รวบรวบคำเชิงบวกและคำเชิงลบที่เกี่ยวกับการเมืองที่ยู่บนเครือข่ายสังคมออนไลน์ ซึ่งการใช้พจนานุกรมที่รวบรวมขึ้นร่วมกับการเรียนรู้ของเครื่องนั้นให้ค่าความถูกต้องที่สูง ดังจะเห็นได้จากงานวิจัยของ Ortigosa และคณะ [9]

2. วิธีดำเนินการวิจัย

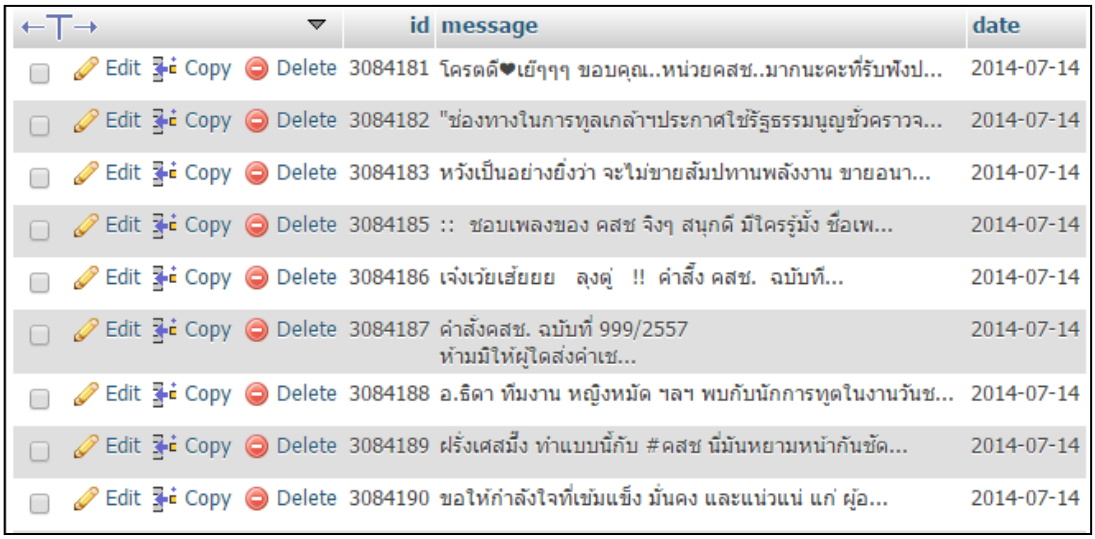


รูปที่ 1 ขั้นตอนวิธีการดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกความคิดเห็นที่มีต่อการรัฐประหารในประเทศไทยจากข้อความคิดเห็นที่อยู่บนเว็บไซต์เครือข่ายสังคมออนไลน์เฟสบุ๊คด้วยวิธีการจำแนกแบบความสัมพันธ์ กระบวนการจำแนกความคิดเห็นในงานวิจัยนี้จะเริ่มจากการรวบรวมข้อความคิดเห็นบนเฟสบุ๊คโดยใช้ Facebook Graph API การเตรียมข้อมูล การสร้างตัวจำแนกความคิดเห็น และการประเมินประสิทธิภาพ ภาพรวมของขั้นตอนวิธีการดำเนินการวิจัยแสดงได้ดังรูปที่ 1

2.1 การเก็บรวบรวมข้อมูล

งานวิจัยนี้ได้เก็บรวบรวมข้อความคิดเห็นบนเฟสบุ๊ค ซึ่งเป็นข้อความแสดงความคิดเห็นเกี่ยวกับการดำเนินงานของ คสช (คณะรักษาความสงบเรียบร้อยแห่งชาติ) ระหว่างวันที่ 22 พฤษภาคม พ.ศ 2557 ถึง 17 กรกฎาคม พ.ศ. 2557 สาเหตุที่ใช้ข้อความที่อยู่บนเฟสบุ๊คเป็นข้อมูลในการวิจัยเนื่องจากในปัจจุบันประชาชนส่วนใหญ่แสดงความคิดเห็นบนเฟสบุ๊คและข้อความที่แสดงออกบนเฟสบุ๊คเป็นข้อความที่ผู้ใช้สามารถแสดงออกได้อย่างอิสระ ไม่ต้องมีการบังคับ ทำให้ข้อมูลที่แสดงออกเป็นข้อมูลความคิดเห็นโดยแท้จริง งานวิจัยนี้ดึงข้อความคิดเห็นที่อยู่ในเฟสบุ๊คโดยการใช้ Facebook Graph API ร่วมกับการเขียนภาษา PHP เพื่อรวบรวมข้อความคิดเห็นเกี่ยวกับ คสช. ไว้ในฐานข้อมูลดังรูปที่ 2 จากนั้นทำการกรองเอาเฉพาะข้อมูลที่เกี่ยวข้องกับความคิดเห็น คสช. โดยข้อมูลที่รวบรวมมีทั้งหมด 467 ข้อความ แบ่งเป็นข้อความเชิงบวก 259 ข้อความและข้อความเชิงลบ 208 ข้อความ ซึ่งพิจารณาจากผู้เชี่ยวชาญ 3 คน ว่าข้อความแต่ละข้อความคือข้อความเชิงบวก หรือข้อความเชิงลบ นอกจากนี้ยังได้รวบรวมคำที่เป็นเชิงบวก 113 คำ และคำที่เป็นเชิงลบ 113 คำ ที่เกี่ยวข้องกับการเมืองเพื่อใช้ในการสกัดคุณลักษณะ



	id	message	date
Edit Copy Delete	3084181	โครตดี❤เย้ๆๆๆ ขอบคุณ..หน่วยคสช..มากนะคะที่รับฟังป...	2014-07-14
Edit Copy Delete	3084182	"ช่องทางในการทูลเกล้าฯประกาศใช้รัฐธรรมนูญชั่วคราวจ...	2014-07-14
Edit Copy Delete	3084183	หวังเป็นอย่างยิ่งว่า จะไม่ขายสัมปทานพลังงาน ขายอนา...	2014-07-14
Edit Copy Delete	3084185	:: ชอบเพลงของ คสช จิงๆ สนุกดี มีใครรู้มั้ง ชื่อเพ...	2014-07-14
Edit Copy Delete	3084186	เจ่งเว้ยเฮ้ยยย ลุงตู่ !! คำสั่ง คสช. ฉบับที่...	2014-07-14
Edit Copy Delete	3084187	คำสั่งคสช. ฉบับที่ 999/2557 ห้ามมิให้ผู้ใดส่งคำเช...	2014-07-14
Edit Copy Delete	3084188	อ.ธิดา ทีมงาน หญิงหมัด ฯลฯ พบกับนักการทูตในงานวันช...	2014-07-14
Edit Copy Delete	3084189	ฝรั่งเศสมึ้ง ทำแบบนี้กับ #คสช นี่มันหยามหน้ากันชัด...	2014-07-14
Edit Copy Delete	3084190	ขอให้กำลังใจที่เข้มแข็ง มั่นคง และแน่วแน่ แก่ ผู้อ...	2014-07-14

รูปที่ 2 ตัวอย่างข้อความคิดเห็นในฐานข้อมูล

2.2 การเตรียมข้อมูล

ก่อนที่จะนำข้อมูลเข้าสู่กระบวนการจำแนก จำเป็นจะต้องแปลงข้อมูลให้อยู่ในรูปแบบทรานแซคชั่น โดยแต่ละทรานแซคชั่นจะประกอบไปด้วยเซตรายการและคลาสดังตารางที่ 1

ตารางที่ 1 ตัวอย่างข้อมูลที่จะนำไปจำแนก

ข้อความ	เซตรายการ	คลาส
1	1 2 4 5 6	P
2	1 4 5 6	N
3	2 4 5	P
4	2 3 4 5	P
5	3 5 6	N

โดยคลาสประกอบไปด้วย 2 คลาส คือ P (ความคิดเห็นเชิงบวก) และ N (ความคิดเห็นเชิงลบ) แต่ละเซตรายการสร้างมาจากข้อความหนึ่งข้อความ โดยจะพิจารณาว่าข้อความประกอบไปด้วยคุณลักษณะลำดับที่เท่าไรบ้าง เช่น สมมติคำคุณลักษณะประกอบไปด้วย 6 คุณลักษณะโดยเรียงตามลำดับดังนี้ 1) อิทธิพล 2) โกง 3) ตกต่ำ 4) สงบ 5) ยั่งยืน 6) เจริญ และข้อความแสดงความคิดเห็นคือ “ผู้มีอิทธิพลเหล่านั้นมักจะโกงกินบ้านเมือง ทำให้บ้านเมืองไม่มีความสงบความเจริญ” จากข้อความแสดงความคิดเห็นดังกล่าวมีคำคุณลักษณะที่ 1 2 4 6 ซึ่งก็คือเซตรายการนั้นเอง

2.3 การสร้างตัวจำแนก

เมื่อได้ข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลแล้วจะนำข้อมูลดังกล่าวไปจำแนก ในงานวิจัยนี้ประยุกต์ใช้วิธีการจำแนกแบบความสัมพันธ์ในการจำแนกความคิดเห็นเชิงบวกหรือเชิงลบ ซึ่งเป็นการรวมกันระหว่างการสืบค้นกฎความสัมพันธ์ (Association Rule) [13] และการจำแนกข้อมูล [14] เพื่อสร้างกฎที่ใช้ในการจำแนกข้อมูลที่เข้าใจง่ายและให้ค่าความถูกต้องสูง โดยใช้ความสัมพันธ์ของข้อมูลที่เกิดร่วมกันบ่อย กฎที่ใช้ในการจำแนกอยู่ในรูปแบบของ $r : X \rightarrow c$ โดยที่ X คือ เซตรายการและ c คือ คลาส กฎจะถูกนำไปใช้ในการจำแนกเมื่อค่าสนับสนุน (Support: $supp(r)$) ของกฎมีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum Support Threshold: min_supp) และค่าความเชื่อมั่น (Confidence: $conf(r)$) ของกฎมีค่า

มากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence: *min_conf*) โดยค่าสนับสนุนและค่าความเชื่อมั่นสามารถคำนวณได้ดังสมการที่ (1) และ (2)

$$supp(r) = \frac{|g(Xc)|}{|g(D)|} \times 100 \quad (1)$$

$$conf(r) = \frac{|g(Xc)|}{|g(X)|} \times 100 \quad (2)$$

โดยที่

$|g(X)|$ คือ จำนวนทรานแซคชั่นที่มี X

$|g(Xc)|$ คือ จำนวนทรานแซคชั่นที่มี X เกิดร่วมกับ c

$|g(D)|$ คือ จำนวนทรานแซคชั่นทั้งหมด

อัลกอริทึมที่ใช้ในการสร้างตัวจำแนกในงานวิจัยนี้ คือ CBA [14] เป็นอัลกอริทึมที่ไม่ซับซ้อนแต่ให้ค่าความถูกต้องสูงในการจำแนก โดยขั้นตอนของอัลกอริทึมประกอบไปด้วย 2 ขั้นตอนหลัก คือ การสร้างกฎความสัมพันธ์บนพื้นฐานของอัลกอริทึม Apriori [13] และการสร้างตัวจำแนกจากกฎ

ขั้นตอนที่ 1 การสร้างกฎความสัมพันธ์ซึ่งมีขั้นตอนดังต่อไปนี้

สมมติกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 40% และค่าความเชื่อมั่นขั้นต่ำเท่ากับ 60%

1. เริ่มจากการหาค่าสนับสนุนของกฎที่มีเซตรายการที่มีความยาว 1 และพิจารณาว่ากฎใดบ้างที่ผ่านค่าสนับสนุนขั้นต่ำ จากนั้นตัดกฎที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำออกดังตัวอย่าง

กฎ	*supp* (%)	กฎ	*supp* (%)
1⟶P	20	4⟶N	20
1⟶N	20	5⟶P	60
2⟶P	60	5⟶N	40
3⟶P	20	6⟶P	20
3⟶N	20	6⟶N	40
4⟶P	60		

*กฎที่ถูกตัดออกคือกฎที่ระบายสีดำ

2. สำหรับกฎที่มีเซตรายการเหมือนกัน จะเลือกเฉพาะกฎที่มีค่าความเชื่อมั่นสูงที่สุด เช่น

conf(5⟶P) = (3/5)*100 = 60%

conf(5⟶N) = (2/5)*100 = 40%

กฎที่ถูกเลือก คือ 5⟶P เนื่องจากมีค่าความเชื่อมั่นสูงสุด

3. เลือกกฎที่มีค่าความเชื่อมั่นมากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำเพื่อนำไปใช้ในการจำแนก จากตัวอย่างทุกกฎมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ ดังนั้นถือว่าเป็นกฎที่สามารถนำไปใช้ในการจำแนกได้

กฎ	*conf* (%)	กฎ	*conf* (%)
2⟶P	100	5⟶P	60
4⟶P	75	6⟶N	67

4. กฎที่ผ่านค่าความเชื่อมั่นขั้นต่ำจะถูกขยายโดยเพิ่มเซตรายการทีละ 1 รายการและกลับไปตรวจสอบข้อ 1-3 ทำแบบนี้ไปเรื่อยๆจนกว่าจะไม่สามารถขยายเซตรายการได้อีก ดังนั้นจากตัวอย่างข้อมูลจะได้กฎที่จะนำไปจำแนกทั้งหมด 8 กฎ ดังนี้

กฎ	*supp* (%)	*conf* (%)
2⟶P	60	100
4⟶P	60	75
5⟶P	60	60
6⟶N	40	67
2 5⟶P	60	100
2 4 ⟶ P	60	100
4 5 ⟶P	60	75
2 4 5⟶P	60	75

ขั้นตอนที่ 2 เมื่อได้กฎสำหรับจำแนกความคิดเห็นแล้ว กฎดังกล่าวจะถูกเรียงลำดับเพื่อสร้างตัวจำแนก โดยกำหนดให้กฎ r_i อยู่ก่อนกฎ r_j ก็ต่อเมื่อ $conf(r_i)>conf(r_j)$ หรือถ้า $conf(r_i)=conf(r_j)$ ให้พิจารณา $supp(r_i)>supp(r_j)$ หรือถ้า $supp(r_i)=supp(r_j)$ ให้พิจารณา $size(r_i)<size(r_j)$ หรือถ้า $size(r_i)=size(r_j)$ ให้พิจารณา r_i ถูกสร้างก่อน r_j จากตัวอย่างเมื่อนำกฎมาเรียงจะได้ผลดังนี้

ลำดับ	กฎ	*conf* (%)	*supp* (%)
1	2 ⟶P	100	60
2	2 4 ⟶ P	100	60
3	2 5⟶P	100	60
4	4 ⟶P	75	60
5	4 5⟶P	75	60
6	2 4 5⟶P	75	60
7	6⟶N	67	40
8	5 ⟶P	60	60

2.4 การประเมินผล

งานวิจัยนี้ใช้ 10-fold cross validation ในการแบ่งข้อมูลเรียนรู้ (Training Set) และข้อมูลทดสอบ (Testing Set) เพราะผลที่ได้น่าเชื่อถือมากกว่าวิธีอื่นเนื่องจากข้อมูลทุกชุดจะถูกนำมาทดสอบเพื่อประเมินผล โดยในแต่ละรอบจะวัดประสิทธิภาพความถูกต้อง (Accuracy) ดังสมการที่ (3) จากนั้นหาค่าเฉลี่ยของความถูกต้องเพื่อดูประสิทธิภาพของตัวจำแนก นอกจากนี้ยังทำการปรับค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำเพื่อหาค่าที่เหมาะสมที่ให้ค่าความถูกต้องสูงในการจำแนก

$$Accuracy = \frac{(tp+tn)}{(tp+tn+fp+fn)} \qquad (3)$$

โดยที่

tp คือ จำนวนข้อความที่ทำนายถูกต้องว่าเป็นความคิดเห็นเชิงบวก

tn คือ จำนวนข้อความที่ทำนายถูกต้องว่าเป็นความคิดเห็นเชิงลบ

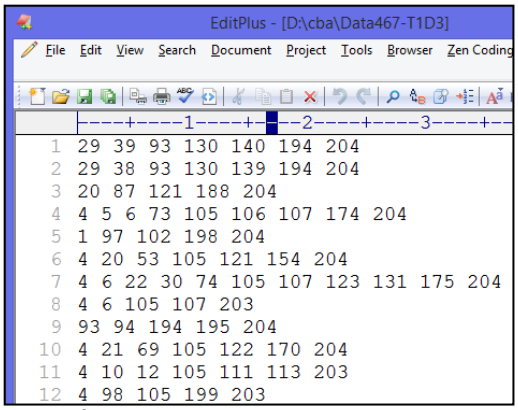
fp คือ จำนวนข้อความที่ทำนายว่าเป็นความคิดเห็นเชิงบวกแต่คำตอบคือความคิดเห็นเชิงลบ

fn คือ จำนวนข้อความที่ทำนายว่าเป็นความคิดเห็นเชิงลบแต่คำตอบคือความคิดเห็นเชิงบวก

3. ผลการวิจัยและวิจารณ์ผลการวิจัย

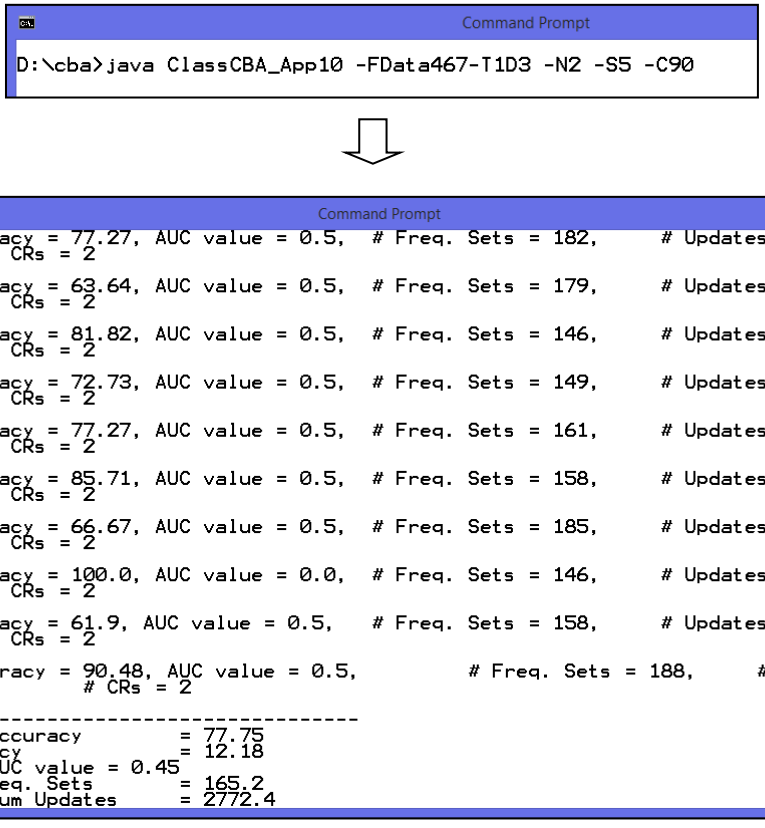
งานวิจัยนี้ทำการทดลองจำแนกความคิดเห็นทางการเมืองโดยใช้วิธีการจำแนกความสัมพันธ์ด้วยขั้นตอนวิธี CBA ซึ่งพัฒนาด้วยภาษาจาวาและสามารถดาวโหลดได้ที่ http://cgi.csc.liv.ac.uk/~frans/KDD/Software/CBA/cba.html [15] เนื่อง จากข้อมูลนำเข้าของโปรแกรม CBA ต้องเป็นตัวเลขทั้งหมดรวมถึงคลาส กระบวนการเตรียมข้อมูลจึงทำการแปลงข้อมูลนำเข้าแต่ละแถวให้เป็นหมายเลขของคุณลักษณะที่ปรากฏในแต่ละข้อความคิดเห็นและตัวเลขสุดท้าย คือ คลาส ผลจากการเตรียมข้อมูลโดยใช้คุณลักษณะที่รวบรวมคำเชิงบวก 113 คำ และคำที่เป็นเชิงลบ 113 คำ ปรากฏว่าบางคุณลักษณะไม่ได้ถูกนำมาใช้เลย คุณลักษณะดังกล่าวจึงถูกตัดทิ้งไป ดังนั้นคุณลักษณะที่นำมาใช้

จริงคือ 202 คุณลักษณะเท่านั้น โดยตัวอย่างข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลแสดงได้ดังรูปที่ 3 ตัวเลข 203 กำหนดให้เป็นความคิดเห็นเชิงลบ และตัวเลข 204 กำหนดให้เป็นความคิดเห็นเชิงบวก



รูปที่ 3 ตัวอย่างข้อมูลนำเข้าโปรแกรม CBA

จากนั้นนำข้อมูลที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วไปทดสอบกับโปรแกรม CBA ดังตัวอย่างในรูปที่ 4 และกำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำแตกต่างกันเพื่อหาค่าที่เหมาะสม โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 1%, 5%, 10%, 15%, 20% และกำหนดค่าความเชื่อมั่นขั้นต่ำเท่ากับ 50%, 60%, 70%, 80%, 90%, 100% เนื่องจากค่าความเชื่อมั่นไม่ควรจะต่ำกว่า 50%



รูปที่ 4 ตัวอย่างการประมวลผลข้อมูล

ผลจากการทดลองแสดงได้ดังตารางที่ 2 เมื่อกำหนดค่าสนับสนุนขั้นต่ำเป็น 1% จะเห็นได้ว่าค่าความเชื่อมั่นขั้นต่ำที่ 100% ให้ค่าความถูกต้องสูงสุด คือ 76.39% จากนั้นทำการปรับค่าสนับสนุนขั้นต่ำเพิ่มขึ้นเรื่อยๆ พบว่าค่าสนับสนุนขั้นต่ำที่ 5% และ 10% ให้ค่าความถูกต้องสูงถึง 77.75% เมื่อค่าความเชื่อมั่นขั้นต่ำเท่ากับ 80% และ 90% ส่วนค่าสนับสนุนขั้นต่ำที่ 15% ให้ค่าความถูกต้องสูงสุด 77.75% เมื่อค่าความเชื่อมั่นขั้นต่ำเท่ากับ 80% ส่วนค่าสนับสนุนขั้นต่ำที่ 20% ให้ค่าความถูกต้องสูงสุด 77.75% เมื่อค่าความเชื่อมั่นขั้นต่ำเท่ากับ 50% และ 60% จะเห็นได้ว่าเมื่อปรับค่าสนับสนุนขั้นต่ำที่สูงขึ้นจะทำให้ได้กฎที่มีค่าความเชื่อมั่นต่ำหรืออาจจะไม่ได้กฎในการจำแนกเลย ทำให้ค่าความถูกต้องเป็น 0 ดังนั้นค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำที่เหมาะสมในงานวิจัยนี้ คือ 10% และ 90% เพราะทำให้ได้กฎที่มีความน่าเชื่อถือมากที่สุด เนื่องจากกฎทุกกฎจะมีค่าความเชื่อมั่นถึง 90% ในการจำแนกและเป็นกฎที่เกิดขึ้นบ่อยถึง 10% จากข้อมูลทั้งหมด นอกจากนี้ยังสามารถจำแนกความคิดเห็นทางการเมืองได้ถูกต้องสูงสุดถึง 77.75%

ตารางที่ 2 ความถูกต้องของการจำแนกความคิดเห็น

min_supp(%) / *min_conf*(%)	**1**	**5**	**10**	**15**	**20**
50	70.32	70.69	62.75	68.38	**77.75**
60	71.69	70.69	62.75	68.38	**77.75**
70	70.28	71.65	70.35	71.71	41.49
80	61.84	**77.75**	**77.75**	**77.75**	0
90	60.84	**77.75**	**77.75**	70.02	0
100	**76.39**	21.67	0	0	0

4. สรุปผลการวิจัย

งานวิจัยนี้นำเสนอการจำแนกความคิดเห็นทางการเมืองที่อยู่บนเครือข่ายสังคมออนไลน์โดยใช้วิธีการจำแนกแบบความสัมพันธ์ ซึ่งให้ค่าความถูกต้องในการจำแนกสูง งานวิจัยนี้ได้ทำการสกัดคุณลักษณะของข้อมูลโดยใช้คำเชิงบวกและคำเชิงลบที่รวบรวมจากข้อความที่อยู่บนเครือข่ายสังคมออนไลน์ และทำการจำแนกโดยพิจารณาความสัมพันธ์ของคุณลักษณะโดยใช้เทคนิค CBA ผลจากการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอ สามารถจำแนกความคิดเห็นทางการเมืองที่อยู่บนเว็บไซต์เครือข่ายออนไลน์ได้ถูกต้องถึง 77.75% งานวิจัยที่จะทำต่อไปในอนาคตคือทำการคัดเลือกคุณลักษณะที่เหมาะสมที่ใช้ในการจำแนกซึ่งอาจจะนำไปสู่การจำแนกที่มีค่าความถูกต้องสูงขึ้น

5. กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้ทุนสนับสนุนงานวิจัยนี้

6. เอกสารอ้างอิง

[1] A. Pak and P. Paroubek. "*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*". in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). May 2010. 35-39.

[2] eBizMBA.com. Top 15 Most Popular Social Networking Sites | April 2014. [cited 8 April 2014]; http://www.ebizmba.com/articles/social-networking-websites.

[3] A. Shrivatava and B. Pant. "*Opinion Extraction and Classification of Real Time Facebook*". **Global Journal of Computer Science and Technology**. 12 (2012).

[4] B Liu and L.Zhang. "*A Survey of Opinion Mining and Sentiment Analysis*". In: Aggarwal CC, Zhai C, eds. **Mining Text Data.** US: Springer; 2012.415-463.

[5] J. Akaichi, Z. Dhouioui and MJL-H.Perez. "*Text Mining Facebook Status Updates for Sentiment Classification*". in Proceeding of the 17th International Conference on System Theory, Control and Computing (ICSTCC). Sinaia. 11-13 Oct. 2013. 640-645.

[6] J. Akaichi. "*Social networks' Facebook' statutes updates mining for sentiment classification*". in Proceedings of SocialCom/PASSAT/BigData/EconCom/ BioMedCom. Washington, DC, United States. 2013. 886-891.

[7] PN. Stuart Russell. "*Artificial Intelligence A Modern Approach*". New Jersey: Prentice Hall; 1995.

[8] C. Cortes and V.Vapnik. "*Support-Vector Networks*". **Mach Learn**. 20 (1995). 273-297.

[9] A. Ortigosa, M. Martín and M. Carro. "*Sentiment Analysis in Facebook and its application to e-learning*". **Compuers in Human Behavior.** 31(2014). 527-541.

[10] J. Keeshin, Z. Galant and D. Kravitz. "*Machine Learning and Feature Based Approaches to Gender Classification of Facebook Statuses*". http://thekeesh.com/cs224n/final_writeup.pdf.

[11] R. Irfan, K. King, D. Drages, S. Ewen, U. Khan, A. Madani, et al. "**A Survey on Text Mining in Social Networks**". **The Knowledge Engineering Review**. 2004.1-24.

[12] The Graph API. [cited 20 April 2014]; https://developers.facebook.com/docs/graph-api.

[13] R. Agrawal, T. Imielinski, A. Swami. "*Mining Association Rules between Sets of Items in Large Databases*". in Proceedings of the 1993 ACM SIGMOD Internatinal Conference on Mangament of Data. USA. 207-216.

[14] B. Liu. "*Integrating Classification and Association Rule Mining*". Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; New York. 80-86.

[15] F. Coenen. LUCS KDD implementation of CBA (Classification Based on Associations). http://www.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cba.html, Department of Computer Science, The University of Liverpool, UK, 2004.