บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

การจำแนกความรู้สึก (Sentiment Classification) [1] คือการจำแนกเอกสารตามขั้นความรู้สึกซึ่งโดยทั่วไปอาจจะจำแนกเป็นความรู้สึกที่เป็นบวก (Positive) ความรู้สึกที่เป็นลบ (Negative) และความรู้สึกที่เป็นกลาง (Neutral) โดยการจำแนกความรู้สึกนั้น ได้รับการศึกษามาอย่างต่อเนื่อง เพราะการประยุกต์ใช้ในหลายลักษณะ แต่โดยทั่วไปมักจะนิยมใช้ในการจำแนกความรู้สึกที่มีการแสดงไว้ในรูปแบบข้อความ (Text) [1] เช่น ประยุกต์ใช้ในการจัดอันดับความรู้สึกจากข้อความแสดงความคิดเห็นของผู้คนที่มีต่อสินค้าและบริการ การประยุกต์ใช้เพื่อวิเคราะห์ความรู้สึกของผู้เรียน การประยุกต์ใช้เพื่อวิเคราะห์ความรู้สึกของผู้คนในเรื่องการเมือง เป็นต้น

อย่างไรก็ตาม แม้ว่าการจำแนกความรู้สึกจะได้รับการศึกษาและความสนใจมาอย่างต่อเนื่อง แต่ยังมีปัญหาที่พบในการจำแนกความรู้สึกหลายประเด็น ประเด็นที่น่าสนใจและยังคงได้รับการศึกษาเพื่อการแก้ปัญหาอยู่คือ ปัญหาความไม่สมดุลของข้อมูลในการจำแนกความรู้สึก (Imbalanced Sentiment Classification) โดยทั่วไปที่พบมากคือปัญหาความไม่สมดุลของข้อมูลในคลาส (Class Imbalance Data) [2–5]

ซึ่งปัญหาความไม่สมดุลของข้อมูลในคลาสนั้น เกิดจากกลุ่มตัวอย่างที่ใช้ในการเรียนรู้มีข้อมูลไม่สมดุลกัน โดยกลุ่มที่มีข้อมูลมากกว่าจะเรียกว่า “ข้อมูลกลุ่มหลัก (Majority Class)” ขณะที่กลุ่มตัวอย่างที่มีข้อมูลจำนวนน้อยกว่าจะเรียกว่า “ข้อมูลกลุ่มรอง (Minority Class)” เมื่อนำเอาชุดข้อมูลในลักษณะนี้ไปเรียนรู้เพื่อสร้างตัวจำแนกความรู้สึก (Sentiment Classifier) ข้อมูลใหม่ๆ ที่อ่านเข้ามาเพื่อวิเคราะห์เพื่อจำแนกกลุ่มด้วยตัวจำแนกความรู้สึกดังกล่าว ก็มีแนวโน้มที่จะทำนายกลุ่มของข้อมูลนั้นไปยังทิศทางของข้อมูลกลุ่มหลักที่ใช้ในการเรียนรู้ตัวจำแนกความรู้สึก

เทคนิคหลายๆ เทคนิคได้ถูกนำเสนอเพื่อใช้ในการควบคุมปัญหาความไม่สมดุลของข้อมูลในการจำแนกความรู้สึก เช่น Resampling Methods [4] สำหรับวิธีการนี้จะเป็นการประยุกต์เอาวิธีสุ่มตัวอย่างซึ่งเป็นวิธีการทางสถิติ เพื่อสร้างข้อมูลสำหรับการสอน โดยมีจุดประสงค์เพื่อให้จำนวนสมาชิกในข้อมูลทั้งสองกลุ่มมีความสมดุลกัน ซึ่งประกอบด้วย 2 วิธีการใหญ่ๆ คือ Oversampling [6] และ Undersampling [6] โดยวิธีการ Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้นให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก และในทางตรงข้ามวิธีการ Undersampling จะทำการสุ่มเลือกข้อมูลสำหรับการสอนจากข้อมูลในกลุ่มหลัก ให้ได้จำนวนที่

ใกล้เคียงกับจำนวนข้อมูลในกลุ่มรอง โดยทั่วไปมักประยุกต์วิธีการแบบ Undersampling แต่ก็จะเกิดปัญหาข้อมูลไม่เพียงพอต่อการเรียนรู้

โดยทั่วไปแล้ว เทคนิคด้าน Resampling Methods มักจะประยุกต์ใช้การคัดเลือกคุณลักษณะ (Feature Selection) [7] เข้ามาช่วยเพื่อควบคุมปัญหาความไม่สมดุลของข้อมูลในคลาส โดยเป็นการคัดเลือกคุณลักษณะเด่นๆ ของข้อมูลในแต่ละคลาสเพื่อเป็นตัวแทนเอกสาร และใช้ในการสร้างตัวจำแนกจำแนกความรู้สึก แต่ก็พบปัญหาคือ บ่อยครั้งพบว่าคุณลักษณะในแต่ละคลาสคือคุณลักษณะเดียวกัน ดังนั้นอาจจะเป็นการยากในการนำมาใช้เพื่อการจำแนกความรู้สึก

อย่างไรก็ตาม เมื่อไม่นานมานี้ หลายงานวิจัยที่นำเสนอเทคนิคการให้น้ำหนักคำ (Term Weighting) เข้ามาช่วยในการแก้ปัญหาความไม่สมดุลของข้อมูลในการจำแนกความรู้สึก [8], [9] และพบว่าเทคนิคการให้น้ำหนักคำแบบมีผู้สอน (Supervised Term Weighting: STW) มีแนวโน้มที่จะทำให้เกิดประสิทธิภาพในการจำแนกความรู้สึกที่ดีขึ้น

ดังนั้นในโครงงานปริญญานิพนธ์ฉบับนี้ จึงได้นำเสนอการศึกษาการแก้ปัญหาความไม่สมดุลของข้อมูลในการจำแนกความรู้สึกด้วยเทคนิคการให้น้ำหนักคำแบบมีผู้สอนอย่างน้อย 3 เทคนิค พร้อมทั้งทำการเปรียบเทียบการเทคนิคการให้น้ำหนักคำแบบไม่มีผู้สอน (Unsupervised Term Weighting) ที่นิยมใช้ในการจำแนกเอกสารความรู้สึกนั่นคือ *tf-idf* (Term Frequency-Inverse Document Frequency) (Salton, Wong, & Yang, 1975) ภายใต้ตัวจำแนกความรู้สึกอย่างน้อย 3 ตัว

1.2 วัตถุประสงค์ของโครงงาน

นำเสนอกระบวนการสำหรับการจำแนกความรู้สึกที่มีข้อมูลไม่สมดุลโดยมีเครื่องมือหลักคือเทคนิคการให้น้ำหนักคำแบบมีผู้สอน

1.3 ขอบเขตของโครงงาน

1) นำเสนอกระบวนการสำหรับการจำแนกความรู้สึกที่มีข้อมูลไม่สมดุลโดยมีเครื่องมือหลักคือเทคนิคการให้น้ำหนักคำแบบมีผู้สอน (Supervised Term Weighting)
2) เป็นการศึกษาการจำแนกแบบ 2 กลุ่ม โดยการศึกษาความไม่สมดุลระหว่าง Majority Class และ Minority Class ใน 3 ระดับของการพิจารณาคือ
    (1) 100:10
    (2) 100:20
    (3) 100:30

3) ข้อมูลที่ใช้ในการทดสอบในการจำแนกความรู้สึกที่ไม่สมดุล ในโครงงานปริญญานิพนธ์นี้เป็นข้อความรีวิวสินค้าอิเล็กทรอนิคที่รวบรวมมาจากเว็บไซต์ Amazon เอกสารจะอยู่ในรูปแบบ XML
4) หนึ่งเอกสารมีคำมากกว่า 30 คำ และไม่เกิน 300 คำ ใช้ทั้งหมด 50,000 เอกสาร
5) ศึกษาเชิงเปรียบเทียบการให้น้ำหนักคำด้วยรูปแบบเทคนิคการให้น้ำหนักคำแบบมีผู้สอนอย่างน้อย 3 ตัว โดยเปรียบเทียบกับเทคนิค *tf-idf* ซึ่งเป็นเทคนิคการให้น้ำหนักแบบไม่มีผู้สอนที่นิยมใช้
6) ศึกษาเชิงเปรียบเทียบอัลกอริทึมการเรียนรู้แบบมีผู้สอนที่ใช้ในการสร้างตัวจำแนกความรู้สึกอย่างน้อย 3 ตัว
7) การวัดประสิทธิภาพการจำแนกความรู้สึกที่ไม่สดดุลจะประเมินด้วยค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟ (F-measure: F1)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้กระบวนการในการจำแนกข้อความแสดงความรู้สึกที่มีข้อมูลแบบไม่สมดุล

1.5 อุปกรณ์และเครื่องมือที่ใช้ในการดำเนินงาน

Hardware: คอมพิวเตอร์รุ่น Intel® Core™ I5-9400F CPU @ 2.90 GHz , RAM 16 GB BUS 2666 MHz, SSD SATA 240 GB
Operating System: Windows 10 Pro
Programming Language: Java, Xml
Application Tools: Eclipse IDLE for java

1.6 แผนการดำเนินงาน

โครงงานปริญญานิพนธ์ฉบับนี้ ดำเนินงาน ณ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างเดือน พฤษภาคม 2563 ถึง เมษายน 2563 ดังที่แสดงในตารางที่ 1.1

ตารางที่ 1.1 แผนการดำเนินงาน

กิจกรรม	เดือน											
	พ.ค.	มิ.ย.	ก.ค.	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1. ศึกษาเทคนิคและรวบรวมข้อมูล												
2. ออกแบบกระบวนการวิจัย												
3. เตรียมข้อมูล												
4. พัฒนาโปรแกรม												
5. การวัดประสิทธิภาพ												
6. เขียนเอกสารฉบับสมบูรณ์												