บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยฉบับนี้เป็นงานวิจัยที่ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพการกำหนดหัวข้อข่าวให้กับเอกสารโดยคำนึงถึงปัจจัยต่าง ๆ ที่มีผลกระทบต่อประสิทธิภาพการกำหนดหัวข้อข่าว ได้แก่ จำนวนเอกสารที่ใช้ในการเรียนรู้ จำนวนคุณลักษณะที่เหมาะสม และค่า Threshold ที่ใช้ในการกำหนดค่าน้ำหนักด้วยวิธี TFIDF ในการกำหนดหัวข้อข่าวนั้น เราจะใช้เทคนิคการกำหนดคุณลักษณะเพื่อให้ได้กลุ่มคำที่สามารถใช้ระบุเพื่อเป็นตัวแทนของเอกสารในแต่ละประเภทกลุ่มเอกสาร แล้วประยุกต์ใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนในการจัดกลุ่มเอกสารเพื่อระบุหัวข้อให้กับแต่ละเอกสารว่าเป็นเอกสารประเภทใด

โดยกระบวนการทำงานของงานวิจัยฉบับนี้ ได้รวบรวมเอกสารข่าวจากเว็บไซต์ข่าวต่าง ๆ ที่เป็นภาษาอังกฤษ แบ่งข้อมูลข่าวออกเป็น 6 ประเภทข่าว ได้แก่ การเมือง กีฬา สุขภาพ ธุรกิจ บันเทิง และสภาพอากาศ ในขั้นตอนการทำงานนั้นแบ่งออกเป็นสองส่วนคือ ขั้นตอนของการเรียนรู้และการทดสอบหรือการจัดกลุ่มเอกสาร โดยทั้งสองขั้นตอนมีการทำงานคือ การเลือกเนื้อความข่าวจากเอกสาร HTML ตัดคำฟุ่มเฟือย และแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์ หลังจากได้คำทั้งหมดแล้วจึงนำมาหาค่าความถี่และค่าน้ำหนักด้วยวิธีการต่าง ๆ ได้แก่ TFIDF TFICF IG และ CHI แล้วทำการกำหนดคุณลักษณะตามเงื่อนไข ๆ เพื่อหาคุณลักษณะที่เหมาะสมเพื่อกำหนดให้เป็นตัวแทนของกลุ่มเอกสารแต่ละประเภท จากนั้นสร้างเวกเตอร์ของแต่ละเอกสารตามคุณลักษณะข้างต้นในรูปแบบ Vector Space Model แล้วนำข้อมูลเหล่านี้เข้าสู่ขั้นตอนการเรียนรู้เพื่อสร้างโมเดลสำหรับการจัดกลุ่มเอกสารโดยใช้อัลกอริทึมเวกเตอร์ซัพพอร์ตแมชชีน จากการเรียนรู้นี้ผลลัพธ์ที่ได้คือโมเดลการเรียนรู้การจัดกลุ่มเอกสาร

โมเดลการเรียนรู้การจัดกลุ่มเอกสาร จะถูกนำมาทดสอบกับเอกสารกลุ่มใหม่ที่อยู่ในรูปแบบของเวกเตอร์เรียบร้อยแล้ว โดยมีการวัดประสิทธิภาพด้วยค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) และค่าเอฟ (F-Measure)

จากผลการทดสอบในงานวิจัยฉบับนี้สามารถสรุปได้ว่า

(1) ค่า Threshold ที่ใช้ในการกำหนดค่าต่ำสุดของความถี่ของคำที่ปรากฏในกลุ่มเอกสาร เพื่อนำมาใช้กำหนดในการเลือกคุณลักษณะนั้น พบว่าจากค่า Threshold ที่ใช้ในการทดสอบ คือ 3, 4, 5 และ 6 ค่า Threshold ที่มีค่าเท่ากับ 5 ให้ค่าความถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าวดีที่สุดในกลุ่ม Threshold ที่กำหนด โดยมีค่าเอฟเท่ากับเฉลี่ยเท่ากับ 92.94 %

(2) จำนวนเอกสารที่ใช้ในการเรียนรู้มีผลต่อความค่าถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าว เมื่อเพิ่มจำนวนเอกสารที่ใช้ในการเรียนรู้มากขึ้น เมื่อคำนวณค่าน้ำหนักด้วยวิธี TFIDF (กำหนด Threshold=5) ค่าเอฟมีค่าเท่ากับ 93.05 % เมื่อใช้เอกสารในเอกสารทดสอบชุดที่ 3 ซึ่งมีจำนวนเอกสาร

เท่ากับ 6,000 เอกสาร โดยค่าเอฟเมื่อเทียบกับเอสการทดสอบชุดที่ 1 ซึ่งมีค่าเท่ากับ 93% ซึ่งมีจำนวนเอกสาร 3,000 เอกสาร เมื่อไม่มีการกำหนดจำนวนคุณลักษณะ และในทำนองเดียวกันเมื่อมีการระบุจำนวนคุณลักษณะเมื่อจำนวนเอกสารที่ใช้ในการเรียนรู้เพิ่มขึ้นค่าเอฟก็มีค่าที่สูงขึ้น ในการคำนวณค่าน้ำหนักด้วยวิธี TFIDF โดยเฉลี่ยเพิ่มขึ้นประมาณ 2% เมื่อกำหนดเงื่อนไขการเลือกคุณลักษณะ (n=500 และ n=1,500) โดยเมื่อค่าเอฟมีค่าที่สูงนั้นหมายความว่ามีกำหนดหัวข้อข่าวให้กับเอกสารที่ถูกต้องมาก

(3) จำนวนคุณลักษณะมีผลต่อค่าความถูกต้องหรือประสิทธิภาพของกำหนดหัวข้อข่าว เมื่อค่าเอฟสูงขึ้นหมายความว่า มีการระบุประเภทเอกสารที่ถูกเพิ่มมากขึ้นด้วย โดยจากการทดลองเพิ่มจำนวนคำสำคัญในแต่ละประเภทเอกสารครั้งละ 500 คำ พบว่า เมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น ค่าความถูกต้องของการกำหนดหัวข้อข่าวจะเพิ่มมากขึ้นด้วย โดยในกลุ่มเอกสารทดสอบชุดที่ 1 จำนวน 3000 เอกสาร เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นเฉลี่ยประมาณ 0.8% TFICF เพิ่มขึ้นเฉลี่ยประมาณ 1.8% IG เพิ่มขึ้นเฉลี่ยประมาณ 3.1% และ CHI เพิ่มขึ้นประมาณ 4.3% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 1.88% เมื่อทดลองในกลุ่มเอกสารทดสอบชุดที่ 2 จำนวน 4,200 เอกสาร เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นประมาณ 1% TFICF เพิ่มขึ้นประมาณ 2.4% IG เพิ่มขึ้นประมาณ 4.7% และ CHI เพิ่มขึ้นประมาณ 4.1% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 2.43% และในกลุ่มเอกสารทดสอบชุดที่ 3 เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นประมาณ 0.9% TFICF เพิ่มขึ้นประมาณ 3.4% IG เพิ่มขึ้นประมาณ 5.2% และ CHI เพิ่มขึ้นประมาณ 4.8% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 2.93% ดังนั้น ค่าความถูกต้องของทุกกลุ่มเอกสารทดสอบเพิ่มขึ้นโดยเฉลี่ย 2.41%

(4) จากการเปรียบเทียบค่าความถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าว โดยการกำหนดคุณลักษณะด้วยการคำนวณค่าน้ำหนักด้วยวิธี TFIDF TFICF IG และ CHI นั้น จากการทดลองพบว่าวิธี TFIDF ให้ค่าความถูกต้องมากกว่าวิธีอื่น ๆ โดยค่าเอฟมีค่าประมาณ 93.05% เมื่อใช้ทุกคำที่ผ่านการตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ กำหนดเป็นคุณลักษณะ และเมื่อทำการระบุจำนวนคุณลักษณะในเอกสารทดสอบทุกชุดจำนวน 3 ชุดนั้นค่าเอฟก็มีค่าเฉลี่ยเท่ากับ 92.93% ซึ่งสูงกว่าการคำนวณด้วยวิธี TFICF IG และ CHI ที่มีค่าเอฟเฉลี่ยเท่ากับ 84.56% 82.52% และ 85.41% ตามลำดับ