การจำแนกความพึงพอใจบนหลักการของการถ่วงน้ำหนักคำด้วยคลาสมิวชวลอินฟอร์เมชัน

อุไรวรรณ บัวตูม[1] กัลยารัตน์ เชี่ยวชาญ[2] วชิราภรณ์ ศรีพุทธ[3] สมบัติ ฝอยทอง[4*]

[1,4]สาขาปัญญาประดิษฐ์ประยุกต์และเทคโนโลยีอัจฉริยะ คณะวิทยาศาสตร์และศิลปศาสตร์, มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี, จันทบุรี, ประเทศไทย*

[2]สาขาเทคโนโลยีสารสนเทศและวิทยาการข้อมูล คณะวิทยาศาสตร์และศิลปศาสตร์, มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี, จันทบุรี, ประเทศไทย

[3]สาขาบริหารธุรกิจ คณะวิทยาศาสตร์และศิลปศาสตร์, มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี, จันทบุรี, ประเทศไทย

*ผู้ประพันธ์บรรณกิจ อีเมล : sombut@buu.ac.th

รับต้นฉบับ : 30 พฤษภาคม 2566; รับบทความฉบับแก้ไข : 4 สิงหาคม 2566; ตอบรับบทความ : 5 สิงหาคม 2566
เผยแพร่ออนไลน์ : 27 ธันวาคม 2566

บทคัดย่อ

การพัฒนาอย่างรวดเร็วของเทคโนโลยีสารสนเทศและแพลตฟอร์มออนไลน์ทำให้การสั่งซื้อสินค้าผ่านระบบออนไลน์จึงเป็นที่นิยม ในปัจจุบันนี้เนื้อหาหรือข้อความของการโพสต์ของลูกค้าจึงมีความสำคัญต่อผู้ผลิตสินค้าในการปรับปรุงคุณภาพของสินค้าให้ตรงกับความต้องการของลูกค้า การวิเคราะห์ความพึงพอใจจากข้อความถือว่าเป็นกระบวนการอย่างหนึ่งของการประมวลผลภาษาธรรมชาติที่สามารถใช้ในการวิเคราะห์เพื่อสะท้อนความรู้สึกและทัศนคติของผู้ใช้ที่มีต่อสินค้าเชิงอารมณ์ได้ทั้งด้านบวกหรือลบ งานวิจัยทางด้านการจำแนกประเภทความพึงพอใจและการจำแนกประเภทเอกสารโดยส่วนใหญ่จะนิยมใช้การกำหนดค่าน้ำหนักคำ ด้วยค่าความถี่เอกสารผกผัน (idf) อย่างไรก็ตามการกำหนดค่าน้ำหนักคำด้วยค่าความถี่เอกสารผกผันอย่างเดียวอาจจะมีประสิทธิภาพไม่เพียงพอในการจำแนกประเภทความพึงพอใจเนื่องจากค่าน้ำหนักนี้ไม่ได้พิจารณามุมมองการจำแนกประเภทข่าวประกอบด้วย งานวิจัยนี้จึงนำเสนอการปรับค่าน้ำหนักคำแบบมีผู้สอนด้วยการนำค่ามิวชวลอินฟอร์เมชันของคลาสคำนวณร่วมกับความถี่ของคำและความถี่เอกสารผกผัน ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอมีประสิทธิภาพที่ดีกว่าการแจกแจงเทอมและการกำหนดค่าน้ำหนักคำด้วยค่าความถี่เอกสารผกผันอย่างเดียว เมื่อพิจารณาจากค่าตัวชี้วัดประสิทธิภาพ: ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าความระลึก และค่าความถ่วงดุล

คำสำคัญ **:** มิวชวลอินฟอร์เมชัน การทำเหมืองความคิดเห็น การวิเคราะห์ความพึงพอใจ การปรับค่าน้ำหนักคำ การจำแนกประเภทเอกสาร

Sentiment Classification Based on Term Weighting with Class-mutual Information

Uraiwan Buatoom[1] Kanyarat Cheawchan[2] Vajiraporn Sriput[3] Sombut Foithong[4*]

[1,4*]*Applied Artificial Intelligent and Smart Technology Program, Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Chanthaburi, Thailand*

[2]*Information Technology and Data Science Program, Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Chanthaburi, Thailand*

[3]*Business Administration Program, Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Chanthaburi, Thailand*

*Corresponding Author. E-mail address: sombut@buu.ac.th

Received: 30 May 2023; Revised: 4 August 2023; Accepted: 5 August 2023
Published online: 27 December 2023

Abstract

Online platforms and information technology are developing quickly, which boosts the popularity of online e-commerce. Nowadays, posting content from client sentiments is vital for product makers to improve product quality as much as possible to exceed customers' expectations. Sentiment analysis is a process of natural language processing that finds out the sentiment and attitudes of users towards a product, whether positive or negative. Most sentiment and text classification research use term weighting with inverse document frequency (idf). However, assigning term weights using the idf method alone may not be effective enough to classify sentiment because this weight does not consider vital information classification views. This paper presents a supervised term weighting using the class mutual information calculated with the term frequency and the inverse document frequency. Experimental results show that the proposed method performs more effectively than term distribution and the term weighting that use only the inverse document frequency when considering by the performance indicator value: Accuracy, Precision, Recall and F1-Measure.

***Keywords*:** Mutual information, Opinion mining, Sentiment analysis, Term weighting, Text classification

1) บทนำ

ด้วยการพัฒนาอย่างรวดเร็วของเทคโนโลยีสารสนเทศและแพลตฟอร์มออนไลน์ในปัจจุบัน การสั่งซื้อสินค้าผ่านระบบออนไลน์จึงเป็นที่นิยมอย่างแพร่หลายในปัจจุบัน ผู้สั่งซื้อสินค้าสามารถโพสต์แสดงความคิดเห็นคุณภาพของสินค้าและบริการของผู้จัดจำหน่าย ดังนั้นเนื้อหาการโพสต์แสดงความรู้สึกของลูกค้าจึงมีความสำคัญอย่างยิ่งต่อผู้ผลิตสินค้าและผู้สั่งซื้อสินค้าเพื่อใช้ในการปรับปรุงคุณภาพของสินค้าและบริการให้ตรงกับความต้องการของลูกค้ามากที่สุด นอกจากนี้บทวิจารณ์สินค้าและบริการที่ลูกค้าโพสต์ไว้จำนวนมากตามเว็บไซต์การทำธุรกิจผ่านสื่ออิเล็กทรอนิกส์ได้กลายเป็นแหล่งข้อมูลอันมีค่าสำหรับการวิเคราะห์ทางการตลาด บทวิจารณ์เหล่านี้เมื่อนำมาวิเคราะห์ความพึงพอใจ (เป็นเชิงบวก หรือ เชิงลบ) สามารถนำมาใช้เพื่อกำหนดกลยุทธ์ทางธุรกิจของเว็บไซต์การทำธุรกิจผ่านสื่ออิเล็กทรอนิกส์ต่าง ๆ เช่น Amazon.com และ Epinion.com [1] ซึ่งผู้ใช้ออนไลน์สามารถได้รับประโยชน์จากการอ่านความคิดเห็นที่มีประโยชน์ของผู้อื่นผ่านบทวิจารณ์ต่าง ๆ ที่โพสต์ได้

การวิเคราะห์ความพึงพอใจนั้นจะเป็นการศึกษาข่าวสารในส่วนของข้อความ อารมณ์ ความคิดเห็น หรือทัศนคติที่มีต่อสินค้าและบริการ ซึ่งเป็นข่าวสารที่มีความเป็นส่วนตัว [2] โดยทำการจัดหมวดหมู่ข้อความว่าเป็นเชิงบวก หรือ เชิงลบ ต่อผู้ผลิตและผู้ให้บริการ การวิเคราะห์ความพึงพอใจนั้นสามารถแบ่งออกได้เป็น 3 ระดับใหญ่ ๆ ได้แก่ ระดับเอกสาร ระดับประโยค และระดับแง่มุมเชิงลึก [3] โดยในระดับเอกสาร [4] จะเป็นการพิจารณาดูว่าข้อความคิดเห็นที่อยู่ในเอกสารนั้นเป็นความคิดเห็นที่เป็นเชิงบวก หรือ เชิงลบ ซึ่งในระดับนี้จะมองว่าข้อมูลในเอกสารเป็นหน่วยของข่าวสาร [5] ที่จะนำมาวิเคราะห์ ส่วนในระดับประโยค [6] นั้นเป้าหมายเพื่อจัดกลุ่มความพึงพอใจในระดับที่เล็กลงกว่าระดับเอกสารโดยพิจารณาแต่ละประโยคของผู้โพสต์ สิ่งสำคัญประการแรกที่ต้องดำเนินการ คือพิจารณาดูว่าความคิดเห็นของผู้โพสต์นั้นเป็นแบบมุมมองส่วนตัว (subjective) หรือ มุมมองที่อยู่บนข้อเท็จจริง (objective) ซึ่งถ้าประโยคนั้นเป็นความคิดเห็นในมุมมองส่วนตัว แล้วการวิเคราะห์ความพึงพอใจในระดับประโยคก็จะพิจารณาดูว่าเป็นความคิดเห็นเชิงบวก หรือ ลบ [7] ในขณะที่ระดับแง่มุมเชิงลึกก็จะแตกต่างต่างจากทั้ง 2 ระดับโดยจะมีความซับซ้อนมากกว่าในการวิเคราะห์ความพึงพอใจซึ่งมีเป้าหมายเพื่อพิจารณาจัดกลุ่มประเภทความพึงพอใจว่าเป็นแง่มุมที่เป็นเป้าหมายของเนื้อหาหรือไม่ [8]

การจำแนกประเภทความพึงพอใจเป็นงานของการจำแนกประเภทเอกสารภาษาธรรมชาติแบบอัตโนมัติเพื่อระบุหมวดหมู่ที่ถูกต้องของเอกสาร โดยการเปลี่ยนข้อความที่แสดงของเอกสารไปเป็นรูปแบบที่มีความกระชับที่สามารถทำให้เอกสารเหล่านั้นสามารถจัดหมวดหมู่โดยใช้คอมพิวเตอร์ได้ ซึ่งโดยทั่วไปแล้วนิยมใช้โมเดลของปริภูมิเวกเตอร์ (Vector space model) ในการแสดงเนื้อหาของเอกสารในรูปของเวกเตอร์ในปริภูมิของคำ (term space) ตัวอย่างเช่น $d = (w_1, w_2, w_3,..., w_k)$ โดยที่ k เป็นจำนวนของคำในเอกสาร โดยทั่วไปคำที่แตกต่างกันก็จะมีความสำคัญที่แตกต่างกันในข้อความ ดังนั้นตัวบ่งชี้ที่สำคัญ w_i จะเป็นเครื่องแสดงว่าคำ t_i มีอิทธิพลมากน้อยแค่ไหนต่อเอกสาร d วิธีการกำหนดค่าน้ำหนักคำเป็นขั้นตอนที่สำคัญที่จะปรับปรุงประสิทธิภาพการจำแนกประเภทความพึงพอใจ และการจัดหมวดหมู่ของเอกสารด้วยการกำหนดค่าน้ำหนักที่เหมาะสมให้กับคำ ถึงแม้ว่าการจัดหมวดหมู่ของเอกสารมีการศึกษาอย่างมากในช่วงหลายทศวรรษที่ผ่านมาก็ตาม แต่การปรับค่าน้ำหนักคำโดยปกติเป็นการยืมหลักการมาจากการสืบค้นเอกสาร (information retrieval) เช่น รูปแบบง่ายที่สุด คือการนำเสนอแบบทวิภาค และที่นิยมใช้กันมากที่สุดคือ ความถี่ของคำ-ความถี่เอกสารผกผัน (tf.idf) ซึ่งทั้งสองหลักการนี้เป็นการกำหนดค่าน้ำหนักคำแบบไม่มีผู้สอน (unsupervised term weighting)

ในการจำแนกประเภทเอกสารนั้นการปรับค่าน้ำหนักคำด้วย tf.idf ถือว่าประสบความสำเร็จมากที่สุดมีการนำมาประยุกต์ใช้อย่างแพร่หลาย [9]–[12] ปัญหาการจำแนกประเภทความพึงพอใจจัดว่าเป็นหัวข้อหนึ่งของปัญหาการจำแนกประเภทเอกสารบนหลักการพิจารณาเป็นหัวข้อ ซึ่งอัลกอริทึมการจำแนกประเภทเอกสารต่าง ๆ สามารถนำมาประยุกต์ใช้กับการจำแนกประเภทความพึงพอใจได้เช่นกัน ในการศึกษาการจำแนกประเภทความพึงพอใจมีงานวิจัยจำนวนมากได้นำหลักการ tf.idf [11]–[13] มาทำการปรับค่าน้ำหนักคำเช่นเดียวกับการจำแนกประเภทเอกสารเพื่อศึกษาถึงประสิทธิภาพของวิธีการต่าง ๆ ที่นำเสนอ บางงานวิจัยชี้ให้เห็นว่าหัวข้อหนึ่ง ๆ มีแนวโน้มที่จะถูกเน้นด้วยคำหลักบางคำที่เกิดขึ้น และความรู้สึกโดยรวมอาจไม่ถูกเน้นผ่านการใช้คำเดิม ๆ ซ้ำ ๆ ก็ได้ [13] แสดงว่าการปรับค่าน้ำหนักคำด้วยวิธีการ tf.idf อย่างเดียวอาจจะมีประสิทธิภาพไม่เพียงพอในการจำแนกประเภทความพึงพอใจ เนื่องจากทั้งความถี่ของคำและความถี่เอกสารผกผันจะพิจารณาเฉพาะความถี่ของคำที่เกิดขึ้น

เป็นสำคัญ ซึ่งความถี่ของคำที่บ่งชี้ถึงทัศนคติด้านบวก หรือ ด้านลบ ของข้อความก็อาจจะมีความถี่โดยส่วนใหญ่ไม่สูงมากนัก

มีหลายงานวิจัย [9], [10] ใช้วิธีการปรับค่าน้ำหนักคำแบบมีผู้สอน (supervised term weighting) โดยการใช้การแจกแจงเทอมเพื่อเพิ่มประสิทธิภาพการจำแนกประเภทเอกสารให้มีประสิทธิภาพสูงขึ้นโดยนำหลักการทางสถิติมาประยุกต์ใช้ซึ่งประกอบด้วยการคำนวณ ค่าเบี่ยงเบนมาตรฐานในเอกสารที่รวบรวม (in-collection deviation) ค่าเบี่ยงเบนภายในคลาส (intra-class deviation) และค่าเบี่ยงเบนระหว่างคลาส (inter-class deviation) อย่างไรก็ตามการแจกแจงเทอมค่อนข้างใช้เวลานานในการคำนวณเนื่องจากมีสูตรที่ซับซ้อน [9], [10] นอกจากนี้เมื่อนำเทคนิคเหล่านี้มาประยุกต์ใช้กับการวิเคราะห์ความพึงพอใจที่แต่ละบทวิจารณ์มีความยาวคำที่ไม่มากและไม่ค่อยซ้ำกัน เหมือนกับเอกสารอื่น ๆ อย่างเช่น หนังสือ หรือหนังสือพิมพ์ จึงทำให้ประสิทธิภาพการจำแนกประเภทความพึงพอใจไม่สูงมาก [13]

มิวชวลอินฟอร์เมชัน (Mutual Information) เป็นทฤษฎีที่สำคัญในการวัดข่าวสาร นิยมนำมาประยุกต์ใช้ในการหาความสัมพันธ์ระหว่างคุณลักษณะ และคลาสของข้อมูล เพื่อพิจารณาว่าคุณลักษณะใดของข้อมูล มีพลังอำนาจในการจำแนกประเภทข้อมูลสูง คุณลักษณะใดที่ไม่จำเป็นในการจำแนกข้อมูล ข้อดีของมิวชวลอินฟอร์เมชันคือ มีความทนทานต่อสัญญาณรบกวนและการแปลงพิกัด รวมทั้งมีรูปแบบการคำนวณที่ไม่ซับซ้อน นักวิจัยจำนวนมากจึงนิยมนำมาใช้ในงานด้านการเลือกลักษณะเด่นของข้อมูล (feature selection) [14]–[17] นอกจากนี้ในงานวิจัยด้านการจำแนกประเภทเอกสารยังนิยมนำมิวชวลอินฟอร์เมชันมาช่วยเพิ่มประสิทธิภาพในการสืบค้นเอกสารให้มีค่าความถูกต้องในการทำนายและค่าตัวชี้วัดต่าง ๆ สูงขึ้น โดยนำมาใช้ในการลดจำนวนคำที่ไม่จำเป็นซึ่งทำให้มิติของเอกสารลดลง และช่วยลดเวลาในการประมวลผลการสืบค้นเอกสาร [15]–[17]

จากข้อดีของมิวชวลอินฟอร์เมชันที่กล่าวมาข้างต้น ในงานวิจัยจึงนี้ได้นำเสนอแนวคิดใหม่ในการนำมิวชวลอินฟอร์เมชันมาประยุกต์ใช้ในการปรับค่าน้ำหนักคำในการสืบค้นเอกสารบทวิจารณ์ความพึงพอใจของลูกค้าโดยพิจารณาให้ความสำคัญกับคำต่าง ๆ ที่มีพลังอำนาจในการจำแนกประเภทความพึงพอใจ ซึ่งการปรับค่าน้ำหนักคำด้วยวิธีการ tf.idf เพียงอย่างเดียวนั้นอาจจะมีประสิทธิภาพไม่เพียงพอในการประยุกต์ใช้กับการจำแนกประเภทความพึงพอใจบทวิจารณ์ของผู้ใช้ที่มีต่อผลิตภัณฑ์และบริการ เนื่องจากเป็นหลักการที่ให้ความสำคัญกับความถี่ของคำที่เกิดขึ้นเท่านั้น จึงมองไม่เห็นข่าวสารที่สำคัญที่ใช้ในการจำแนกประเภทบทวิจารณ์ความพึงพอใจของลูกค้าออนไลน์ ดังนั้นงานวิจัยนี้จึงนำเสนอการปรับค่าน้ำหนักคำแบบมีผู้สอนโดยใช้คลาสของมิวชวลอินฟอร์เมชัน [14] ในการเพิ่มประสิทธิภาพการจำแนกประเภทความพึงพอใจโดยการปรับค่าน้ำหนักคำ tf.idf โดยคำนวณร่วมกับคลาสของมิวชวลอินฟอร์เมชัน หลักการที่นำเสนอนี้จะใช้เวลาในการคำนวณที่น้อยกว่าหลักการแจกแจงเทอมอย่างมาก นอกจากนี้ยังมีประสิทธิภาพที่ดีกว่าในการจำแนกความถูกต้องของการจัดประเภทความพึงพอใจว่าเป็นด้านบวก หรือด้านลบ

ส่วนที่เหลือของงานวิจัยมีดังนี้ คือ ส่วนที่ 2 เป็นบทสรุปของมิวชวลอินฟอร์เมชันและการแจกแจงเทอมทั้ง 3 ชนิด รวมถึงวิธีการอื่น ๆ ที่นำมาเปรียบเทียบ ส่วนที่ 3 แสดงปัญหาที่เกิดขึ้นของ idf และอธิบายหลักการสำหรับวิธีการที่นำเสนอ ส่วนที่ 4 เปรียบเทียบผลลัพธ์ของวิธีการที่นำเสนอกับวิธีการอื่น ๆ ด้วยการเปรียบเทียบค่าตัวชี้วัด ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าความระลึก และ ค่าความถ่วงดุล และส่วนของการสรุปผลและหัวข้อการวิจัยในอนาคตที่เป็นไปได้จะกล่าวถึงในหัวข้อที่ 5

2) ทฤษฎีพื้นฐาน

2.1) เอนโทรปีและมิวชวลอินฟอร์เมชัน (Entropy and Mutual Information)

เครื่องมือที่นักวิจัยนิยมใช้วัดความไม่แน่นอน หรือ ความยุ่งเหยิงของตัวแปรสุ่มคือ เอนโทรปี ถ้ากำหนดให้ตัวแปรสุ่มแบบไม่ต่อเนื่อง X และ Y ที่มีเซ็ตของผลลัพธ์ที่เป็นไปได้ทั้งหมด คือ Ω และ Ψ ตามลำดับ และกำหนดฟังก์ชั่นความหนาแน่นความน่าจะเป็นของ X คือ $p(x) = \Pr\{X = x\}, x\epsilon\Omega$ ดังนั้นการคำนวณหาเอนโทรปีของตัวแปรสุ่ม X สามารถนิยามเป็น

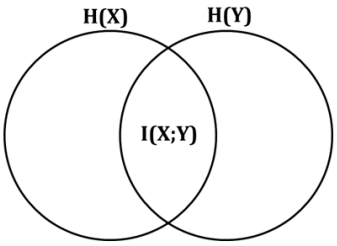
$$H(X) = -\sum_{x\epsilon\Omega} p(x)\, log\, p(x) \quad (1)$$

สำหรับตัวแปรสุ่มแบบไม่ต่อเนื่องสองตัว คือ X และ Y ทีมีการแจกแจงความน่าจะเป็นร่วมกัน p(x, y) ดังนั้นในการวัดผลรวมของข่าวสารที่ตัวแปรสุ่มหนึ่งเกี่ยวข้องกับอีกตัวแปรสุ่มจะคำนวณหาโดยมิวชวลอินฟอร์เมชัน ซึ่งสามารถคำนวณได้ดังนี้

$$I(X;Y) = \sum_{x\epsilon\Psi}\sum_{y\epsilon\Omega} p(x,y) \log\frac{p(x,y)}{p(x)\cdot p(y)} \quad (2)$$

ถ้ามิวชวลอินฟอร์เมชันระหว่าง 2 ตัวแปรมีค่ามาก (น้อย) หมายถึงว่า ตัวแปรทั้งสองจะมีความสัมพันธ์ที่ใกล้ชิดกัน (ไม่มีความสัมพันธ์ใกล้ชิดกัน) ถ้ามิวชวลอินฟอร์เมชันมีค่าเป็นศูนย์

แล้วตัวแปรสุ่มทั้งสองตัวจะไม่เกี่ยวข้องกันอย่างสมบูรณ์ ในขณะที่ รายละเอียดความสัมพันธ์ระหว่างเอนโทรปีและมิวชวลอินฟอร์เมชัน สามารถพิจารณาได้จากรูปที่ 1



รูปที่ 1 : ความสัมพันธ์ระหว่างเอนโทรปีและมิวชวลอินฟอร์เมชัน

2.2) การปรับค่าน้ำหนักคำแบบดั้งเดิม (Traditional Term Weighting)

ในงานการจำแนกประเภทเอกสารนั้นแต่ละคำของเวกเตอร์เอกสารจะเกี่ยวข้องกับค่า (น้ำหนัก) ซึ่งใช้วัดความสำคัญของคำและบ่งชี้ว่าคำนั้นมีศักยภาพในการจัดหมวดหมู่เอกสารมากน้อยแค่ไหน การกำหนดค่าน้ำหนักคำมาตรฐานที่นิยมใช้ได้แก่ ความถี่ของคำที่ปรากฏในเอกสาร (tf) แต่การใช้งาน tf อย่างเดียวอาจจะมีพลังอำนาจในการจำแนกไม่มากพอในการสืบค้นเอกสารที่ต้องการจากเอกสารที่ไม่ต้องการอื่น ๆ ดังนั้นความถี่เอกสารผกผัน (idf) จึงถูกนำมาใช้เพื่อเพิ่มพลังในการจำแนกของคำสำหรับการสืบค้นเอกสาร กำหนดให้ $D = \{d_1, d_2, d_3, \dots, d_N\}$ เป็นเซตของเอกสารทั้งหมด N เอกสาร (Corpus), $T = \{t_1, t_2, t_3, \dots, t_k\}$ เป็นเซตทั้งหมดของ k คำ ดังนั้นแต่ละเอกสารสามารถเขียนแสดงในรูปในรูปของปริภูมิเวกเตอร์ของคำ $d_j = (w_{j1}, w_{j2}, w_{j3}, \dots, w_{ji})$ เมื่อ j = 1,2,3,…., N และ i = 1,2,3, …, k โดยที่ w_{ij} แสดงถึงค่าน้ำหนักของคำ t_i ในเอกสาร d_j โดยความถี่ของคำ (tf) ในรูปแบบมาตรฐานสามารถคำนวณได้ดังนี้

$$tf_{ij} = \frac{\text{ความถี่ของคำที่ } \boldsymbol{i} \text{ ในเอกสารที่ } \boldsymbol{d_j}}{\text{จำนวนคำทั้งหมดในเอกสารที่ } \boldsymbol{d_j}} \quad (3)$$

การใช้ tf อย่างเดียวอาจจะส่งผลให้พลังในการจำแนกไม่เพียงพอเนื่องจากคำที่มีความถี่สูง ๆ ก็จะทำให้มีค่าน้ำหนักของคำในเอกสารมากตามไปด้วย แต่คำที่มีอำนาจการจำแนกสูง ๆ ก็ควรที่จะปรากฏในเอกสารจำนวนไม่มาก ซึ่งในที่นี้ก็คือการคำนวณหาค่าความถี่เอกสารผกผัน (idf) ซึ่งสามารถคำนวณได้ดังนี้

$$idf_i = log\left(\frac{N}{Num\_doc(t_i,D)}\right) \quad (4)$$

โดยที่ $Num\_doc(t_i, D)$ คือ จำนวนเอกสารในคลังเอกสารที่ประกอบด้วยคำ t_i และในงานวิจัยนี้ในการคำนวณจะใช้ลอการิทึมฐาน 2 ในการคำนวณ ในขณะที่การคำนวณหาค่าน้ำหนักของคำที่ t_i ในเอกสาร d_j ด้วย if.idf สามารถคำนวณได้ดังนี้

$$w_{ij} = tf_{ij} \cdot idf_i \quad (5)$$

การปรับค่าน้ำหนักคำด้วย tf.idf เป็นหลักการที่มาจากการสืบค้นเอกสาร (Information retrieval) และในช่วงหลายสิบปีที่ผ่านมามีงานวิจัยจำนวนมากได้นำหลักการนี้มาประยุกต์ใช้ในการจำแนกประเภทเอกสาร บางงานวิจัยก็นำเทคนิคหลักการอื่น ๆ มาประยุกต์ใช้กับ tf.idf เพื่อปรับปรุงประสิทธิภาพในการจัดหมวดหมู่เอกสาร นอกจากนี้หลักการปรับค่าน้ำหนักคำด้วย tf.idf นั้นเป็นหลักการของการปรับค่าน้ำหนักคำแบบไม่มีผู้สอนโดยจะให้ความสำคัญอย่างมากกับความถี่ของคำที่ปรากฏในเอกสารและจำนวนของเอกสารที่ปรากฏคำนั้น อย่างไรก็ตามการนำหลักการของ tf.idf มาคำนวณหาค่าน้ำหนักคำในการวิเคราะห์ความพึงพอใจ อาจจะมีประสิทธิภาพไม่มากพอในการจำแนกประเภทความพึงพอใจว่าเป็นด้านบวก หรือ ลบ เนื่องจากบทวิจารณ์นั้นเป็นประโยคที่มีความยาวจำนวนคำไม่มากและมีแนวโน้มที่ในประโยคจะถูกเน้นด้วยคำหลักบางคำและความรู้สึกโดยรวมอาจไม่ถูกเน้นผ่านการใช้คำเดิม ๆ ซ้ำ ๆ

นอกจากนี้ยังมีวิธีการปรับค่าน้ำหนักคำแบบ rtf-sisf [18] เป็นอีกวิธีการหนึ่งที่มีการปรับเปลี่ยนพัฒนามาจาก tf.idf โดยการพิจารณาความถี่สัมพัทธ์ของคำ t_i ในเอกสารที่มีการรวบรวมและมีการบวก 1 เพิ่มเข้าไปในการคำนวณ idf เพื่อป้องกันกรณีที่ค่าอินพุตของลอการิทึมเป็น 1 ซึ่งการคำนวณน้ำหนักคำแบบ rtf-sisf สามารถคำนวณได้ดังนี้

$$w_{ij} = (tf_{ij}/MaxFreq_i) \cdot A \quad (6)$$

โดยที่ $A = \log(N/(1 + Num\_doc(t_i, D)))$ และ $MaxFreq_i$ คือ จำนวนความถี่สูงสุดของคำ t_i

การปรับค่าน้ำหนักคำที่อธิบายข้างต้นนั้นเป็นรูปแบบหนึ่งของการปรับค่าน้ำหนักแบบไม่มีผู้สอนซึ่งเป็นหลักการที่สนใจหรือให้ความสำคัญกับความถี่ของคำที่เกิดขึ้นในเอกสาร โดยไม่ได้สนใจว่าคำนั้นมีชนิดของคลาสของเอกสารเป็นคลาสอะไร ซึ่งในหัวข้อถัดไปจะกล่าวถึงการปรับค่าน้ำหนักคำแบบมีผู้สอนที่จะมีการพิจารณาชนิดของคลาสของเอกสารของคำใด ๆ เพื่อช่วยเพิ่มประสิทธิภาพการจำแนกเอกสารให้ดีขึ้น

2.3) การปรับค่าน้ำหนักคำแบบมีผู้สอน (Supervised Term Weighting)

การปรับค่าน้ำหนักคำแบบมีผู้สอนนั้นจะใช้ชนิดของคลาสของเอกสารที่ใช้สำหรับการสอนในการปรับค่าน้ำหนักคำ t_i ทั้งนี้ขึ้นอยู่ว่าแต่ละคำสามารถจำแนกเอกสารได้ถูกต้องมากน้อยแค่ไหนซึ่งจะส่งผลต่อค่าน้ำหนักคำที่จะปรับเพื่อให้คำนั้นมีพลังอำนาจในการจำแนกที่สูงในการทำนายหมวดหมู่ของเอกสาร ซึ่งในงานวิจัยนี้จะอธิบายหลักการที่นำมาเปรียบเทียบประสิทธิภาพกับวิธีการที่นำเสนอดังนี้

2.3.1) การแจกแจงเทอม (Term Distributions) ในส่วนนี้เป็นการอธิบายการปรับค่าน้ำหนักคำโดยใช้การแจกแจงเทอมหรือคำ มีหลายงานวิจัยที่นำหลักการนี้มาใช้ในการจำแนกประเภทเอกสาร [9], [10], [19], [20] เพื่อปรับปรุงประสิทธิภาพของการปรับค่าน้ำหนัก tf.idf การแจกแจงเทอมที่ใช้กันมีสามชนิด ได้แก่ ค่าเบี่ยงเบนมาตรฐานระหว่างคลาส (icsd), ค่าเบี่ยงเบนมาตรฐานของคลาส (csd) และค่าเบี่ยงเบนมาตรฐาน (sd) ถ้ากำหนดให้ tf_{ijq} เป็นค่าความถี่ของคำ t_i ในเอกสาร d_j ซึ่งอยู่ในคลาส c_q โดยสามารถนิยามค่า icsd ของคำ t_i, csd ของคำ t_i ในคลาส q และ sd ของคำ t_i ตามลำดับได้ดังนี้

$$icsd_i = \sqrt{\frac{\sum_q \left[\overline{tf_{iq}} - \frac{\sum_q \overline{tf_{iq}}}{|C|}\right]}{|C|}} \quad (7)$$

$$csd_{iq} = \sqrt{\frac{\sum_{d_j \in c_q}\left[tf_{ijq} - \overline{tf_{iq}}\right]^2}{|c_q|}} \quad (8)$$

$$sd_i = \sqrt{\frac{\sum_q \sum_{d_j \in c_q}\left[tf_{ijq} - \frac{\sum_q \sum_{d_j \in q} tf_{ijq}}{\sum_q |c_q|}\right]^2}{\sum_q |c_q|}} \quad (9)$$

โดยที่ $\overline{tf_{iq}} = \frac{\sum_{d_j \in c_q} tf_{ijq}}{|c_q|}$ เป็นค่าเฉลี่ยความถี่ของเทอม t_i ในเอกสารทั้งหมดที่อยู่ภายในคลาส c_q, $|C|$ คือจำนวนคลาส และ $|c_q|$ คือ จำนวนเอกสารในคลาส c_q

การคำนวณการปรับค่าน้ำหนักคำด้วยวิธี icsd ของคำใด ๆ จะไม่ขึ้นกับคลาส คำที่มีค่า icsd สูงจะทำให้เกิดพลังในการจำแนกความแตกต่างกันระหว่างคลาส และทำให้มีอำนาจในการแยกแยะการจำแนกประเภทเอกสารที่สูงกว่าคำอื่น ๆ ดังนั้นค่า icsd จะใช้ในการส่งเสริมคำที่มีอยู่ในเกือบทุกคลาสแต่มีความถี่ของคำสำหรับคลาสเหล่านั้นค่อนข้างแตกต่างกันมาก ในขณะที่คำซึ่งมีค่า csd สูงนั้นแสดงว่าคำนี้จะปรากฏอยู่ในเอกสารส่วนใหญ่ของคลาสด้วยความถี่ที่ต่างกันค่อนข้างมาก และไม่ใช่ตัวแทนที่ดีของคลาส ในขณะที่ค่า csd ของคำนั้นมีค่าน้อยแสดงว่าอาจเกิดขึ้นจากหนึ่งในสองสาเหตุคือ การเกิดขึ้นของคำนั้นเกือบจะเท่ากันสำหรับทุกเอกสารในคลาส หรืออาจจะเป็นคำที่ไม่ค่อยเกิดขึ้นในคลาสก็ได้ สำหรับคำที่มีค่า sd สูงแสดงว่าคำนั้นจะปรากฏในเอกสารส่วนใหญ่ที่มีการรวบรวมด้วยความถี่ที่ต่างกันอย่างมาก ในทางกลับกันคำที่มีค่า sd ต่ำอาจจะเกิดจากสาเหตุสองประการ คือ การเกิดขึ้นของคำนั้นเกือบจะเท่ากันสำหรับทุกเอกสาร หรือเป็นคำที่ไม่ค่อยเกิดขึ้นในเอกสารที่มีการรวบรวม อย่างไรก็ตามการปรับค่าน้ำหนักคำด้วยการแจกแจงเทอมทั้ง 3 หลักการนี้ icsd จะใช้สำหรับในการส่งเสริมคำในขณะที่ csd และ sd จะใช้สำหรับการลดการส่งเสริมคำ โดยสามารถคำนวณหาค่าน้ำหนัก w_{ij} ของคำ t_i ของเอกสาร d_j ได้ดังนี้

$$w_{ij} = tf_{ij} \cdot idf_i \cdot icsd_i^p \quad (10)$$

$$w_{ij} = (tf_{ij} \cdot idf_i)/csd_{iq}^p \quad (11)$$

$$w_{ij} = (tf_{ij} \cdot idf_i)/sd_i^p \quad (12)$$

ซึ่งในการทดลองงานวิจัยนี้จะใช้ค่า p เป็น 0.5 และ 1 โดยเป็นค่าที่ได้จากผลการทดลองของงานวิจัย [10] ในการเปรียบเทียบเทียบประสิทธิภาพการปรับค่าน้ำหนักคำกับวิธีการอื่น ๆ ในการจำแนกประเภทความพึงพอใจ

2.3.2) ความถี่ของคำ-ความถี่ของเอกสารที่ตรงกับความต้องการ (Term Frequency-Relevance Frequency: tf.rf)

การปรับค่าน้ำหนักคำแบบมีผู้สอนด้วยวิธีการ tf.rf [21] เป็นหลักการปรับค่าน้ำหนักคำที่จะพิจารณาสัดส่วนของจำนวนเอกสารที่มีคำ t_i ประกอบอยู่และอยู่ในคลาสบวก (positive class) ต่อจำนวนเอกสารที่มีคำ t_i ประกอบอยู่และอยู่ในคลาสลบ (negative class) โดยสามารถนิยามสูตรการปรับค่าน้ำหนักคำ t_i ของเอกสาร d_j ได้ดังสมการต่อไปนี้

$$w_{ij} = tf_{ij} \cdot \log\left(2 + \frac{D_{pos}(t_i)}{\max(1, D_{neg}(t_i))}\right) \quad (13)$$

โดยที่ $D_{pos}(t_i)$ คือ จำนวนเอกสารที่ประกอบด้วยคำ t_i และอยู่ในคลาสบวก

$D_{neg}(t_i)$ คือ จำนวนเอกสารที่ประกอบด้วยคำ t_i และอยู่ในคลาสลบ

ในงานวิจัยนี้จะนำเสนอวิธีการปรับค่าน้ำหนักคำทั้งแบบมีผู้สอนโดยใช้หลักการคลาสมิวชวลอินฟอร์เมชันที่แก้ข้อเสียของวิธีการ tf.idf ที่เป็นวิธีการปรับค่าน้ำหนักคำที่ให้ความสำคัญกับความถี่ของคำที่เกิดขึ้นโดยไม่ได้มีการพิจารณาถึงความถี่ของคำนั้นมาจากคลาสใด ยิ่งเมื่อนำหลักการ tf.idf มาใช้ในการปรับค่าน้ำหนักคำสำหรับการจำแนกความพึงพอใจก็อาจจะทำให้

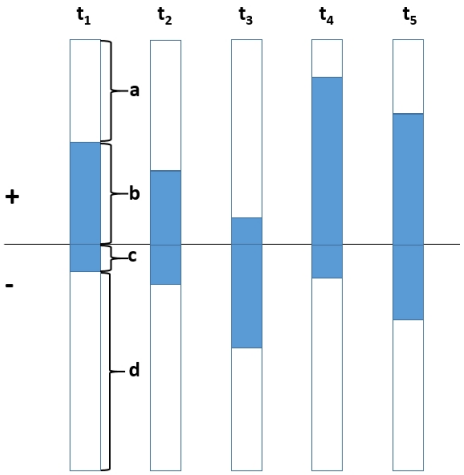
ประสิทธิภาพความแม่นยำในการจัดหมวดหมู่บทวิจารณ์ไม่สูงมากนัก เนื่องจากบทวิจารณ์ออนไลน์ของผู้ใช้โดยทั่วไปแล้วจะเป็นโพสต์ข้อความสั้น ๆ ไม่ยาวมาก และความรู้สึกที่โพสต์ลงไปโดยรวมแล้วอาจไม่เน้นการใช้คำเดิม ๆ ซ้ำ ๆ เหมือนกับในเอกสารเช่น หนังสือ หรือ หนังสือพิมพ์ ซึ่งรายละเอียดแนวคิดของวิธีการที่นำเสนอและแนวทางในการแก้ข้อเสียของ tf.idf จะอธิบายในหัวข้อถัดไป

3) การปรับค่าน้ำหนักคำด้วยคลาสมิวชวลอินฟอร์เมชัน

3.1) ปัญหาการปรับค่าน้ำหนักคำด้วยวิธีการ idf ในการจำแนกประเภทความพึงพอใจ

โดยทั่วไปแล้วการวิเคราะห์ความพึงพอใจก็จะเกี่ยวข้องกับวิเคราะห์ทิศทางความรู้สึกบนหลักพิจารณาของข้อความ การพิจารณาดูข้อความนั้นว่าเป็นแบบมุมมองความคิดเห็นส่วนตัว หรือ มุมมองที่อยู่บนเนื้อหาข้อเท็จจริง และถ้าเป็นแบบมุมมองความคิดเห็นส่วนตัวจึงจะวิเคราะห์ต่อไปว่าข้อความนั้นเป็นความพึงพอใจที่เป็นเชิงบวก หรือ เชิงลบ ซึ่งในงานวิจัยโดยทั่วไปก็จะมองว่าปัญหาการจำแนกประเภทความพึงพอใจจัดว่าเป็นรูปแบบหนึ่งของปัญหาการจำแนกประเภทเอกสารที่เกี่ยวกับหัวข้อที่สนใจ ดังนั้นกระบวนการขั้นตอนต่าง ๆ ในการจำแนกประเภทความพึงพอใจก็จะเหมือนกับการจัดหมวดหมู่ของเอกสาร รวมไปถึงการแสดงถึงรูปแบบการแสดงของเอกสารจะอยู่ในรูปแบบเวกเตอร์น้ำหนักของคำซึ่งค่าน้ำหนักของคำที่นิยมใช้ในการศึกษาทดลองวิจัยส่วนใหญ่จะเป็น idf [22], [23] อย่างไรก็ตามรูปแบบการปรับค่าน้ำหนักด้วย idf อาจจะขาดประสิทธิภาพในบางสถานการณ์ซึ่งจะนำไปสู่ประสิทธิภาพความแม่นยำในการทำนายด้วยตัวจำแนกประเภทและตัวชี้วัดอื่น ๆ ไม่สูงมากนัก [22], [23]

ความถี่ของคำ tf จะแสดงถึงความสัมพันธ์ที่ใกล้ชิดกันระหว่างคำและเนื้อหาในเอกสารที่ประกอบด้วยคำนั้น เป็นที่สังเกตได้ว่าถ้าความถี่ของคำมีค่ามากและแพร่กระจายไปยังจำนวนเอกสารเป็นจำนวนมากก็จะทำให้ไม่สามารถสืบค้นเอกสารให้ตรงกับที่ต้องการจากเอกสารทั้งหมดได้ เพื่อแก้ปัญหานี้จึงมีการนำ idf มาช่วยปรับปรุงเพื่อเพิ่มพลังในการจำแนกของคำในการสืบค้นเอกสาร แต่อย่างไรก็ตามการนำหลักการนี้มาใช้ในการจำแนกประเภทของความพึงพอใจก็อาจจะพบเจอปัญหาบางประการในการวิเคราะห์ความพึงพอใจดังในรูปที่ 2 โดยการกระจายของเอกสารจะประกอบด้วย 5 คำ คือ t_1, t_2, t_3, t_4 และ t_5 แต่ละคอลัมน์จะพิจารณาว่าเป็นการกระจายของเอกสารในคลังเอกสารสำหรับแต่ละคำ ความสูงของแต่ละแท่งก็จะแทนจำนวนเอกสารที่อยู่ในคลังเอกสารและเส้นแนวนอนจะใช้แบ่งเอกสารออกเป็น 2 กลุ่มคือ คลาสบวก (Positive) และ คลาสลบ (Negative) ความสูงแต่ละคอลัมน์ที่อยู่เหนือเส้นแนวนอนและใต้เส้นแนวนอนก็จะแทนจำนวนเอกสารที่อยู่ในคลาสบวกและคลาสลบ ตามลำดับ ความสูงของพื้นที่แรงเงาของแต่ละคอลัมน์ก็จะแทนจำนวนเอกสารที่อยู่ในคลาสบวกและลบ เพื่อให้การอธิบายเข้าใจได้ง่ายก็จะใช้ a, b, c และ d แทนจำนวนเอกสารที่แตกต่างกันดังรายการต่อไปนี้



รูปที่ 2 : ตัวอย่างการกระจายที่แตกต่างกันของเอกสารที่ประกอบด้วย 6 เทอมในเอกสารที่รวบรวมทั้งหมด

โดย a แทนจำนวนเอกสารที่อยู่ในคลาสบวกที่ไม่ได้ประกอบด้วยคำนี้

b แทนจำนวนเอกสารที่อยู่ในคลาสบวกที่ประกอบด้วยคำนี้

c แทนจำนวนเอกสารที่อยู่ในคลาสลบที่ประกอบด้วยคำนี้

d แทนจำนวนเอกสารที่อยู่ในคลาสลบที่ไม่ได้ประกอบด้วยคำนี้

ดังนั้น N แทนจำนวนเอกสารทั้งหมดที่รวบรวมโดยเป็นผลรวมของ a, b, c และ d

ในที่นี้สมมุติว่าทั้ง 5 คำนี้มีค่าความถี่ของคำ (tf) เท่ากัน โดยที่สามคำแรก t_1, t_2, t_3 มีค่า idf เท่ากันซึ่งจะแทนด้วยค่า idf1 ส่วนสองคำสุดท้าย t_4 และ t_5 ก็จะมีค่า idf เท่ากันซึ่งแทนด้วยค่า idf2 ซึ่งจากรูปจะเห็นได้ว่าค่า idf1 > idf2 ดังนั้นแสดงว่า 3 คำแรกก็จะมีค่าน้ำหนักของคำมากกว่าเนื่องจากมีความถี่ของเอกสารในคลังเอกสารน้อยกว่า 2 คำสุดท้าย จะเห็นได้ชัดว่าทั้ง 5

คำนั้นจะสนับสนุนการจำแนกประเภทความพึงพอใจได้แตกต่างกันซึ่งจะให้ความสำคัญอย่างมากกับเทอมหรือคำที่ทำให้มีจำนวนคลาสบวกมากกว่าคลาสลบซึ่งจะใช้ในการคัดแยกตัวอย่างที่เป็นคลาสบวกออกจากตัวอย่างที่เป็นคลาสลบ ถ้าพิจารณาจากรูปที่ 2 จะพบว่า t_1 จะสนับสนุนมากกว่า t_2 และ t_3 ในการคัดแยกตัวอย่างที่เป็นคลาสบวกออกจากตัวอย่างที่เป็นคลาสลบในทำนองเดียวกัน t_4 จะสนับสนุนการคัดแยกได้มากกว่า t_5 เช่นกัน

อย่างไรก็ตามถ้ามีการกำหนดค่าน้ำหนักคำด้วย idf จะเห็นได้ว่าไม่สามารถเห็นความแตกต่างระหว่าง 3 คำแรกและเช่นเดียวกันก็ไม่สามารถแยกความแตกต่างระหว่าง 2 คำสุดท้ายได้เช่นกัน ถ้าพิจารณาจากรูปที่ 2 จะเห็นได้ว่า t_1 จะสนับสนุนพลังในการจำแนกข้อความบทวิจารณ์ที่เป็นคลาสบวกออกจากข้อความบทวิจารณ์ที่เป็นคลาสลบได้ดีกว่า t_2 และ t_3 ดังนั้น t_1

ควรจะมีค่าน้ำหนักมากกว่า t_2 และ t_3 อย่างไรก็ตามการกำหนดค่าน้ำหนักด้วย idf ทั้ง t_1, t_2 และ t_3 จะให้พลังในการจำแนกประเภทความพึงพอใจได้เท่า ๆ กัน จึงเห็นได้ว่ากรณีนี้การกำหนดค่าน้ำหนักด้วย idf จะล้มเหลวในการกำหนดค่าที่เหมาะสมให้กับคำเมื่อพิจารณาจากการสนับสนุนที่แตกต่างกันในงานการจำแนกประเภทข้อความบทวิจารณ์

ถ้าพิจารณา t_1 และ t_4 จะเห็นได้ว่าการปรับค่าน้ำหนักด้วย idf จะทำให้ $idf1(t_1) > idf2(t_4)$ ซึ่งก็คือ t_1 มีพลังในการจำแนกได้ดีกว่า t_4 แต่จากรูปที่ 2 ถ้าพิจารณาแล้วจะเห็นได้ชัดเจนว่า t_4 ควรมีพลังในการจำแนกได้ดีกว่า t_1 ดังนั้น t_4 ควรมีค่าน้ำหนักมากกว่า t_1 ซึ่งในทำนองเดียวกันค่า idf ของ t_4 และ t_5 เท่ากัน แสดงว่าทั้ง 2 คำจะมีพลังในการจำแนกข้อความบทวิจารณ์คลาสบวกออกจากคลาสลบได้เท่า ๆ แต่ในความเป็นจริงแล้ว t_4 ควรจะมีพลังในการจำแนกคลาสบวกออกจากคลาสลบได้มากกว่า t_5 ดังนั้น t_4 ควรจะต้องมีค่าน้ำหนักมากกว่า t_5 เช่นกัน

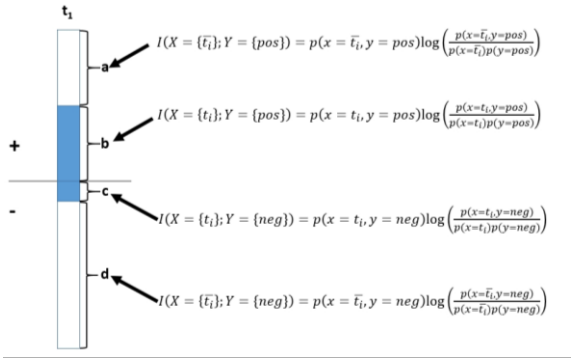
จากที่กล่าวมาข้างต้นเป็นความล้มเหลวของการกำหนดค่าน้ำหนักคำแบบ idf ที่ไม่สามารถแก้ปัญหาที่กล่าวมาข้างต้นได้จึงทำให้การกำหนดค่าน้ำหนักคำด้วยวิธีการ tf.idf ไม่ทำให้ประสิทธิภาพในงานการจำแนกประเภทความพึงพอใจสูงขึ้นได้ ดังนั้นในงานวิจัยนี้จึงนำเสนอวิธีการกำหนดค่าน้ำหนักคำด้วยคลาสมิวชวลอินฟอร์เมชันที่จะแก้จุดด้อยจากปัญหาข้างต้นของวิธีการ idf และช่วยเพิ่มประสิทธิภาพการจำแนกประเภทความพึงพอใจ รวมถึงตัวชี้วัดต่าง ๆ ที่สำคัญ ให้สูงขึ้นกว่าวิธีการ idf ซึ่งจะได้อธิบายรายละเอียดให้หัวข้อถัดไป

3.2) แนวคิดของวิธีการที่นำเสนอ

จากปัญหาข้างต้นที่เกิดขึ้นกับวิธีการกำหนดค่าน้ำหนักคำด้วย idf ซึ่งเป็นการปรับค่าน้ำหนักคำแบบไม่มีผู้สอน เนื่องจากวิธีนี้ให้ความสำคัญกับค่าความถี่ของคำที่เกิดขึ้นเท่านั้น โดยไม่ได้พิจารณาว่าข้อความบทวิจารณ์มาจากคลาสบวกหรือคลาสลบ ในงานวิจัยนี้จะแก้ปัญหาของวิธีการ idf ด้วยการปรับค่าน้ำหนักคำ tf.idf ด้วยคลาสมิวชวลอินฟอร์เมชัน (Class Mutual Information: cmi) โดยวิธีการ cmi จัดว่าเป็นการปรับค่าน้ำหนักคำแบบมีผู้สอน ซึ่งน้ำหนักคำที่ปรับจะขึ้นอยู่กับพลังอำนาจในการสนับสนุนการจำแนกคลาสบวกออกจากคลาสลบของคำนั้น

ในการคำนวณ cmi นั้นมีหลักการตรงไปตรงมา ซึ่งจะให้ความสำคัญกับการคำนวณหาข่าวสารร่วมของ $\boldsymbol{t_i}$ และคลาสบวก (คำนวณบริเวณส่วนของ b ในรูปที่ 2) ซึ่งจะเห็นได้ว่าถ้าส่วนที่แรงเงาส่วน b มีค่ามากก็จะทำให้มีพลังอำนาจมากในการสนับสนุนการจำแนกเอกสารข้อความคลาสบวกออกจากคลาสลบ ซึ่งถ้าพิจารณาในส่วนของ 3 คำแรก คือ t_1, t_2 และ t_3 จะเห็นได้ว่าค่า idf เท่ากันทุกคำซึ่งไม่สะท้อนความเป็นจริงและล้มเหลวในการกำหนดค่าที่เหมาะสมให้กับคำ แต่เมื่อพิจารณาจากส่วนแรงเงา b ซึ่งเป็นข่าวสารร่วมระหว่างคำ $\boldsymbol{t_i}$ ที่ปรากฏอยู่ในเอกสารข้อความและเอกสารข้อความที่อยู่ในคลาสบวกจะพบว่าพลังอำนาจในการสนับสนุนการของคำ t_1 มากกว่า t_2 และ t_2 มากกว่า t_3 นั่นคือ $cmi(t_1) > cmi(t_2) > cmi(t_3)$ ซึ่งสามารถดูตัวอย่างวิธีการคำนวณส่วนต่างของคำได้จากรูปที่ 3

ในทำนองเดียวกันเมื่อพิจารณา t_1 และ t_4 จะเห็นได้ว่าการปรับค่าน้ำหนักด้วย cmi จะทำให้ $cmi(t_4) > cmi(t_1)$ ซึ่งก็คือ t_4 จะมีพลังในการจำแนกเอกสารข้อความได้ดีกว่า t_1 ดังนั้น t_4 ก็จะมีค่าน้ำหนักมากกว่า t_1 นอกจากนี้ถ้าพิจารณาจากค่า cmi ของ t_4 และ t_5 พบว่าการกำหนดค่าน้ำหนักของคำ t_4 มากกว่า t_5 ซึ่งสะท้อนถึงความเป็นจริงที่จำนวนเอกสารข้อความในคลาสบวกของเทอม t_4 มากกว่า รวมไปถึงการสนับสนุนในการคัดแยกตัวอย่างของคลาสบวกออกจากคลาสลบของเทอม t_4 และ t_5 จะมากกว่า t_1, t_2 และ t_3 ซึ่งการกำหนดค่าน้ำหนักคำของ t_4 และ t_5 ควรจะต้องมากกว่า t_1, t_2 และ t_3 ด้วยเช่นกัน โดยการคำนวณคลาสมิวชวลอินฟอร์เมชันของแต่ละส่วนสามารถพิจารณาได้จากรูปที่ 3



รูปที่ 3 : ตัวอย่างการคำนวณคลาสมิวชวลอินฟอร์เมชัน
สำหรับพื้นที่แต่ละส่วนของคำ

จากรูปที่ 3 สามารถเขียนสมการในการคำนวณค่าน้ำหนักคำ t_i ด้วยวิธีการ idf ใหม่ได้ดังนี้

$$idf_i = log\left(\frac{a+b+c+d}{b+c}\right) \quad (14)$$

กำหนดให้ t_i แทนคำหรือเทอมที่ปรากฏอยู่ในเอกสาร, $\overline{t_i}$ แทนคำหรือเทอมที่ไม่ปรากฏอยู่ในเอกสาร และกำหนดให้ pos แทนเอกสารที่อยู่ในคลาสบวก และ neg แทนเอกสารที่อยู่ในคลาสลบ ดังนั้นจากสมการ (2) การคำนวณหาค่าข่าวสารร่วมของคลาสและเอกสารที่ประกอบด้วยคำ t_i สามารถเขียนสมการได้เป็น

$$I(X;Y) = \sum_{x\epsilon\Psi}\sum_{y\epsilon\Omega} p(x,y)\log\frac{p(x,y)}{p(x)\cdot p(y)} \quad (15)$$

โดยที่ $\Psi = \{t_i, \overline{t_i}\}, \Omega = \{pos, neg\}$

เนื่องจากแนวคิดในงานวิจัยนี้จะพิจารณากำหนดค่าน้ำหนักคำโดยให้ความสนใจเฉพาะขนาดพื้นที่ส่วนแรงเงาของ b ที่จะสามารถสนับสนุนอำนาจในการจำแนกเอกสารข้อความที่เป็นคลาสบวกออกจากคลาสลบ ได้ดีกว่าส่วนอื่น ๆ ซึ่งถ้าปริมาณเอกสารในส่วนของ b มีค่าปริมาณข่าวสารมาก ๆ ย่อมส่งผลดีต่อการกำหนดค่าน้ำหนักของคำในการเพิ่มอำนาจในการจำแนกเอกสารข้อความบทวิจารณ์คลาสบวกออกจากคลาสลบ ดังนั้นการคำนวณหาข่าวสารของจำนวนเอกสารข้อความบทวิจารณ์ที่ประกอบด้วยคำ t_i และเอกสารเหล่านั้นอยู่ในคลาสบวก (pos) สามารถคำนวณได้จากสมการต่อไปนี้

$$I(X = \{t_i\}; Y = \{pos\}) = Mlog(S) \quad (16)$$

โดยที่ $M = p(x = t_i, y = pos)$, $S = \frac{p(x=t_i,y=pos)}{p(x=t_i)p(y=pos)}$

จากการพิจารณารูปที่ 2 และ 3 และ จากสมการ (16) ข้างต้นแทนจะมีการแทนค่า $p(t_i, pos)$ ด้วยค่า b/N แทนค่า $p(t_i)$ และ $p(pos)$ ด้วยค่า (b+c)/N และ (a+b)/N ตามลำดับ ดังนั้นการแทนค่าจะได้เป็น

$$I(X = \{t_i\}; Y = \{pos\}) = \left(\frac{b}{N}\right) log\left(\frac{\frac{b}{N}}{\left(\frac{b+c}{N}\right)\left(\frac{a+b}{N}\right)}\right) \quad (17)$$

เพื่อความสะดวกในการนำสมการที่ (17) ไปประยุกต์ใช้และเปรียบเทียบกับวิธีการอื่น ๆ ในการคำนวณหาข่าวสารร่วมระหว่างเอกสารที่ประกอบด้วยคำ t_i และเอกสารที่อยู่ในคลาส C_q จะมีการสร้างตัวแปร cmi_{iq} ที่แสดงความสัมพันธ์กับคลาสมิวชวลอินฟอร์เมชัน สามารถเขียนแสดงความสัมพันธ์ได้ดังนี้ คือ

$$cmi_{iq} = I\left(X = \{t_i\}; Y = C_q\right) \quad (18)$$

โดยที่ $C_q = \{pos\}$

สำหรับวิธีการปรับค่าน้ำหนักคำที่นำเสนอนี้สามารถคำนวณหาค่าน้ำหนัก w_{ij} ของคำ t_i และเอกสาร d_j ได้ดังนี้

$$w_{ij} = tf_{ij} \cdot idf_i \cdot cmi_{iq} \quad (19)$$

ซึ่งการปรับค่าน้ำหนักคำด้วยวิธีการที่นำเสนอนี้จะนำค่าน้ำหนักคำ tf.idf ไปคูณกับ cmi_{iq} ตามสมการที่ (19) เพื่อทำการปรับค่าน้ำหนักคำให้เหมาะสมกับศักยภาพของคำต่าง ๆ ที่สนับสนุนพลังอำนาจในการจำแนกเอกสารข้อความคลาสบวกออกจากคลาสลบ ซึ่งในหัวข้อถัดไปจะได้อธิบายถึงขั้นตอนของการนำวิธีการปรับค่าน้ำหนักคำที่นำเสนอไปประยุกต์ใช้ในการจำแนกประเภทความพึงพอใจข้อความบทวิจารณ์ที่มีกระบวนการขั้นตอนต่าง ๆ ที่จะได้อธิบายต่อไป

3.3) การปรับค่าน้ำหนักคำด้วยคลาสมิวชวลอินฟอร์เมชัน

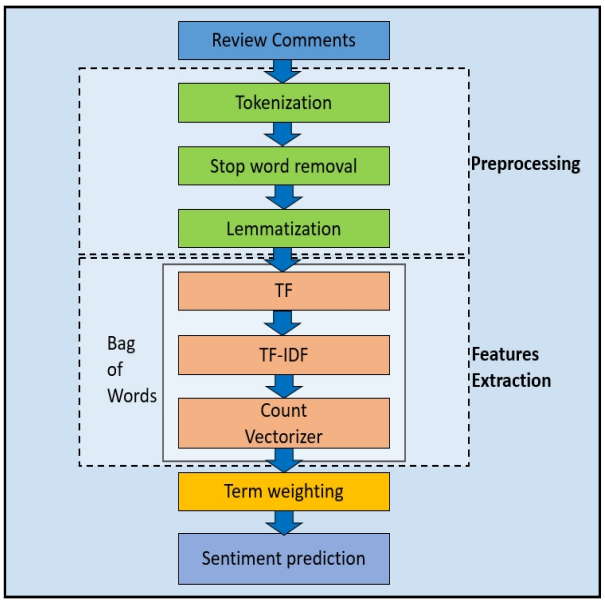
การจำแนกประเภทความพึงพอใจของข้อความบทวิจารณ์นั้นเป็นกระบวนการหนึ่งในการพิจารณาบทวิจารณ์ เช่น บทวิจารณ์ผลิตภัณฑ์/บริการ แบบสำรวจออนไลน์ เป็นต้น ว่ามีทัศนคติเป็นเชิงบวก หรือ เชิงลบ ซึ่งจะมีผลดีต่อผู้ผลิตสินค้าหรือ ผู้ให้บริการที่จะเข้าใจความชื่นชอบของผู้บริโภคที่มีต่อสินค้าและบริการและสามารถรวบรวมความคิดเห็นของผู้บริโภคมาปรับปรุงการให้บริการ [24] ต่าง ๆ ได้ ขั้นตอนการประมวลผลการจำแนกประเภทความพึงพอใจของความคิดเห็นแสดงดังรูปที่ 4 มีขั้นตอนดังนี้

3.3.1) การเตรียมข้อมูลก่อนการประมวลผล (Preprocessing)
ทำการแตกประโยคออกเป็นคำย่อย จากนั้นจะเป็นการลบคำหยุด คำเว้นวรรค จุด รวมไปถึง ‘a’, ‘an’, ‘the’ เป็นต้น แล้วจากนั้นจึงไปหารากศัพท์ของคำ และในขั้นตอนสุดท้ายจะเป็นกระบวนการแปลงคำด้วยรายการคำศัพท์ในภาษาเพื่อลดคำที่ผันอยู่ในรูปแบบต่าง ๆ

3.3.2) การสกัดคุณลักษณะของคำ (Feature Extraction)
ขั้นตอนนี้จะทำการแปลงคำให้อยู่ในรูปตัวเลขโดยมองแต่ละแถว

เป็นเอกสารและคอลัมน์เป็นคำ ขั้นตอนนี้เริ่มต้นด้วยการคำนวณหาค่าน้ำหนักคำด้วยความถี่ของคำ (tf) และคำนวณหาความถี่เอกสารผกผันและสร้างโทเค็นของการรวบรวมเอกสารข้อความและสร้างคำศัพท์สำหรับคำที่สร้างขึ้นให้สามารถที่จะเข้ารหัสเอกสารใหม่ได้โดยใช้คำศัพท์นั้น

3.3.3) การจำแนกประเภทความพึงพอใจ (Sentiment Classification) เป็นการจำแนกประเภทเอกสารข้อความบทวิจารณ์ของผลิตภัณฑ์และบริการทำได้โดยใช้แนวทางการเรียนรู้ของเครื่อง เช่น เบย์อย่างง่าย (Naïve Bayes) และ ป่าของต้นไม้ตัดสินใจแบบสุ่ม (Random Forest) เป็นต้น ในการสร้างโมเดลการจำแนกจากชุดข้อมูลที่ใช้สอน เพื่อนำมาใช้ในการจำแนกข้อมูลทดสอบเพื่อการทำนาย



รูปที่ 4 : ขั้นตอนการจำแนกประเภทความพึงพอใจเอกสารข้อความบทวิจารณ์

ขั้นตอนการจำแนกประเภทเอกสารข้อความบทวิจารณ์จะเริ่มต้นด้วยการอ่านข้อความบทวิจารณ์เข้ามาเพื่อทำขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล แล้วเข้าสู่กระบวนการการสกัดคุณลักษณะของคำ หลังจากนั้นก็จะคำนวณหาค่า tf.idf สำหรับนำไปใช้คำนวณร่วมกับวิธีการที่นำเสนอ โดยวิธีการ cmi_{iq} จะมีการคำนวณโดยพิจารณาจากค่าของความถี่ของคำ tf_{ij} ที่แปลงให้อยู่ในรูปแบบมาตรฐานที่เรียกว่า ถุงของคำ (Bag of words) โดยแถวแทนเอกสารทั้งหมดในคลังและคอลัมน์แทนค่าน้ำหนักของคำดังรูปที่ 5 โดยแต่ละแถวนั้นก็จะแทนด้วยเวกเตอร์ของเอกสาร ที่แสดงในรูปแบบ $d_j = (tf_{1j}, tf_{2j}, tf_{3j}, \dots, tf_{ij})$ ซึ่งก่อนที่จะมีการคำนวณหาค่าคลาสมิวชวลอินฟอร์เมชันของเทอม t_i ของแต่ละคลาส เพื่อความสะดวกในการคำนวณหาค่าความน่าจะเป็นในการคำนวณนั้นจะต้องมีการแปลงตารางที่อยู่ในรูปที่ 5 ไปเป็นตารางที่แสดงค่าความถี่ของคำในรูปแบบไบนารี่ก่อน ดังรูปที่ 6

จากนั้นก็จะทำการคำนวณหาค่าน้ำหนักของคำด้วยคลาสมิวชวลอินฟอร์เมชัน หรือ คำนวณหาข่าวสารร่วมระหว่างเทอม t_i ที่สัมพันธ์ร่วมกับคลาสที่ q สามารถคำนวณด้วยสมการที่ (17) เมื่อคำนวณ cmi_{iq} ของแต่ละคลาสแล้ว ก็จะมีการคำนวณค่าน้ำหนักเทอม t_i สำหรับแต่ละข้อความบทวิจารณ์ที่อยู่ในคลาส q ด้วยการใช้สมการที่ (19)

Bag of Words

	tf_1	tf_2	tf_3	. . .	tf_i	class
d_1	0	0.4	0.33		0	pos
d_2	0.8	0.5	0		0.3	pos
d_3	0.4	0.1	0.1		0.2	neg
	. . .	. . .	. . .	. . .	. . .	. . .
d_N	0.7	0	0		0.6	pos

รูปที่ 5 : ถุงของคำที่แสดงค่าความถี่ของคำในรูปแบบมาตรฐาน

Binary Bag of Words

	tf_1	tf_2	tf_3	. . .	tf_i	class
d_1	0	1	1		0	pos
d_2	1	1	0		1	pos
d_3	1	1	1		1	neg
	. . .	. . .	. . .	. . .	. . .	. . .
d_N	1	0	0		1	pos

รูปที่ 6 : ถุงของคำในรูปแบบทวิภาค

ในงานวิจัยนี้ได้นำเสนอการเพิ่มประสิทธิภาพการปรับค่าน้ำหนักของคำ tf.idf ในการแก้ข้อด้อยที่เกิดขึ้นของวิธีการ idf เนื่องจากวิธีการนี้พิจารณาความถี่ของคำเป็นสำคัญแต่เพิกเฉยต่อศักยภาพอำนาจในการจำแนกตัวอย่างคลาสบวกออกจากคลาสลบของเทอมนั้น ดังนั้นในงานวิจัยนี้จะนำเสนอวิธีการที่จะช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทความพึงพอใจข้อความบทวิจารณ์ด้วยการคำนวณ tf.idf ร่วมกับคลาสมิวชวลอินฟอร์เมชัน นอกจากนี้ในหัวข้อถัดไปจะได้มีการเปรียบเทียบประสิทธิภาพ

วิธีการที่นำเสนอกับวิธีการ tf.idf และยังทำการเปรียบเทียบประสิทธิภาพกับวิธีการปรับค่าน้ำหนักเทอมด้วยวิธีการการแจกแจงเทอม และวิธีการอื่น ๆ ซึ่งผลการทดลองแสดงให้เห็นถึงประสิทธิภาพของวิธีการที่นำเสนอเมื่อพิจารณาทั้ง ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าความระลึก และ ค่าความถ่วงดุล

4) ผลการทดลอง

ในส่วนนี้จะแสดงผลการทดลองระหว่างวิธีที่นำเสนอและเทคนิคอื่นที่ได้รับความนิยมในการปรับค่าน้ำหนักคำร่วมกับข้อมูลบทวิจารณ์ที่นิยมนำมาใช้ในการทดลองและขนาดของข้อมูลก็ไม่ใหญ่มากที่จะเกินข้อจำกัดของเครื่องคอมพิวเตอร์ที่ใช้ในการทดลอง โดยในส่วนนี้จะมีการเปรียบเทียบประสิทธิภาพวิธีการปรับค่าน้ำหนักคำกับตัวชี้วัดต่าง ๆ ที่นิยมใช้กันในงานด้านการหมวดหมู่ประเภทเอกสาร ซึ่งตัวชี้วัดที่ใช้ประกอบด้วย ค่าความถูกต้องการทำนาย (Accuracy), ค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าความถ่วงดุล (F1-Measure) โดยในการทดลองงานครั้งนี้จะแบ่งชุดข้อมูลใช้สอนและชุดข้อมูลใช้ทดสอบออกเป็น 5 ชุด (5 Fold Cross Validation) และทำการทดสอบบนตัวจำแนกประเภทข้อมูลที่นิยมนำมาใช้กันในงานวิจัย ซึ่ง ได้แก่ เบย์อย่างง่าย (Naïve Bayes), ป่าของต้นไม้ตัดสินใจแบบสุ่ม (Random forest) และ เพื่อนบ้านที่ใกล้ที่สุด K อันดับ (K-NN) เพื่อใช้เป็นตัววัดประสิทธิภาพของตัวชี้วัดต่าง ๆ ของเทคนิคการกำหนดค่าน้ำหนักคำที่นำมาเปรียบเทียบประสิทธิภาพ โดยวิธีการจำแนกประเภทด้วยเพื่อนบ้านที่ใกล้ที่สุด K อันดับ ในการทดลองนี้จะใช้ K = 3

4.1) ข้อมูลบทวิจารณ์สินค้าที่ใช้ในการทดลอง

เพื่อแสดงให้เห็นถึงประสิทธิผลของวิธีการที่นำเสนอสำหรับการจำแนกประเภทความพึงพอใจจะขออธิบายรายละเอียดถึงชุดข้อมูลที่ใช้ในการวิเคราะห์ความพึงพอใจจากหลาย ๆ เว็บไซต์ที่นิยมใช้กันในการศึกษางานวิจัย ข้อมูลแรกคือ Amazon เป็นเอกสารข้อความที่สกัดมาจากบทวิจารณ์สินค้าต่าง ๆ ของลูกค้าจากเว็บไซต์ amazon.com ประกอบด้วยข้อความ 1000 บทวิจารณ์ ซึ่งประกอบด้วยข้อความที่เป็นความรู้สึกด้านบวกต่อสินค้า 500 บทความคิดเห็นและความรู้สึกที่เป็นด้านลบต่อสินค้า 500 บทความคิดเห็นเช่นกัน ในขณะที่ข้อมูล Yelp เป็นเอกสารข้อความที่สกัดมาจากบทความคิดเห็นร้านอาหารของลูกค้าจากเว็บไซต์ yelp.com ประกอบด้วยข้อความบทความคิดเห็น 1000 บทความคิดเห็น ซึ่งประกอบด้วยความรู้สึกที่เป็นด้านบวกต่อร้านอาหาร 500 บทความคิดเห็นและความรู้สึกที่เป็นด้านลบ 500 บทความคิดเห็น ส่วนชุดข้อมูลถัดไป คือ IMDb ก็จะสกัดมาจากบทความคิดเห็นภาพยนต์ของลูกค้าจากเว็บไซต์ imdb.com ที่ประกอบด้วยข้อความบทความคิดเห็น 1000 บทความคิดเห็น โดยแบ่งเป็นความรู้สึกที่เป็นด้านบวก 500 บทความคิดเห็นและความรู้สึกที่เป็นด้านลบ 500 บทความคิดเห็นเช่นกัน ทั้งข้อมูล Amazon, Yelp และ IMDb เป็นข้อมูลที่แต่ละเอกสารประกอบด้วยข้อความบทความคิดเห็นและตามด้วยชนิดของคลาสของความรู้สึก โดยทั้ง 3 ข้อมูลนี้เป็นข้อมูลที่อ้างอิงมาจากงานวิจัย [25] ซึ่งนักวิจัยสามารถที่จะดาวน์โหลดข้อมูลเหล่านี้ได้ได้จากเว็บไซต์ https://archive.ics.uci.edu/ml/m-achine-learning-databases/00331/

ตารางที่ 1 : สรุปรายละเอียดชุดข้อมูลที่ใช้ในการทดลอง

ลำดับ	ชุดข้อมูล	คำอธิบาย
1	Amazon	บทวิจารณ์สำหรับผลิตภัณฑ์ที่ขายบนเว็บไซต์ amazon.com ในหมวดโทรศัพท์มือถือและอุปกรณ์เสริม
2	Yelp	บทวิจารณ์จากชุดข้อมูล Yelp challenge2 ซึ่งได้รวบรวมมาจากบทความคิดเห็นร้านอาหารบนเว็บไซต์ yelp.com
3	IMDb	ข้อมูลความคิดเห็นบทวิจารณ์ภาพยนตร์ซึ่งเป็นรายการบทวิจารณ์ภาพยนตร์ที่โพสต์บนเว็บไซต์ imdb.com
4	Eco-Pre	ข้อมูลบทความคิดเห็นผลิตภัณฑ์ EcoDot สำหรับ Gen3 ของ amazon จากเว็บไซต์ amazon.in
5	Magazine	ข้อมูลบทวิจารณ์ความคิดเห็นนิตยสารของสมาชิกโดยได้รวบรวมมาจากเว็บไซต์ amazon.com

ในขณะที่ข้อมูลบทความคิดเห็นสินค้า Eco-Pre เป็นข้อมูลความคิดเห็นผลิตภัณฑ์ EcoDot ซึ่งเป็นรุ่น Gen3 ของ amazon ที่มีการรวบรวมความคิดเห็นบทวิจารณ์จากผู้ใช้งานมาจากเว็บไซต์ amazon.in โดยข้อมูลบทความคิดเห็นนี้จะมีทั้งสิ้น 4,084 เอกสารข้อความและแต่ละข้อความก็จะมีการระบุถึงความรู้สึกที่มีต่อสินค้าซึ่งเป็นความรู้สึกด้านบวกจำนวน 3,066 เอกสารข้อความ ความรู้สึกด้านลบจำนวน 482 เอกสารข้อความและความรู้สึกเป็นกลางจำนวน 536 เอกสารข้อความ โดยข้อมูล

Eco-Pre สามารถที่จะทำการดาวน์โหลดได้ข้อมูลนี้จากเว็บไซต์ https://www.kaggle.com/datasets/pradeeshprabhakar/preprocessed-dataset-sentiment-analysis และสำหรับข้อมูลบทความคิดเห็นนิตยสาร Magazine เป็นข้อมูลบทวิจารณ์นิตยสารของสมาชิกโดยได้รวบรวมมาจากเว็บไซต์ amazon.com โดยข้อมูลประกอบด้วยข้อความบทวิจารณ์ 1,571 ข้อความซึ่งเป็นความรู้สึกด้านบวกจำนวน 1,454 ข้อความและความรู้สึกด้านลบจำนวน 117 ข้อความ โดยข้อมูลบทวิจารณ์ความคิดเห็นนิตยสารนี้สามารถดาวน์โหลดได้จากเว็บไซต์ https://nijianmo.github.io/amazon/index.html โดยข้อมูลทั้งหมดที่ใช้ในการทดลองในงานวิจัยครั้งนี้สามารถพิจารณาดูรายละเอียดได้จากตารางที่ 1 และ ตารางที่ 2

ตารางที่ 2 : คุณลักษณะของข้อมูลที่ใช้ในการจำแนกประเภทความพึงพอใจ

ลำดับที่	ชุดข้อมูล	#คอลัมน์	#แถว	คลาส
1	Amazon	422	1,000	บวก/ลบ
2	Yelp	464	1,000	บวก/ลบ
3	IMDb	748	1,000	บวก/ลบ
4	Eco-Pre	1,330	4,084	บวก/ลบ/กลาง
5	Magazine	1,213	1,571	บวก/ลบ

ในงานวิจัยนี้จะแสดงให้เห็นถึงประสิทธิภาพที่ดีกว่าของวิธีการที่นำเสนอ cmi ซึ่งสามารถแก้ปัญหาข้อจำกัดต่าง ๆ เมื่อเปรียบเทียบกับวิธีการ tf.idf โดยในการเปรียบเทียบจะพิจารณาค่าเฉลี่ยของตัวชี้วัดทั้งหมด 5 ครั้ง ทั้งในแง่ของค่าความถูกต้องในการทำนายเอกสารข้อความ (Accuracy: Acc) ค่าความแม่นยำ (Precision: Pre) เพื่อแสดงถึงสัดส่วนของการทำนายคลาสนั้นที่ถูกต้องจากการทำนายคลาสนั้นทั้งหมด ค่าความระลึก (Recall: Rec) ที่แสดงถึงสัดส่วนของการทำนายคลาสนั้นที่ถูกต้องจากคลาสจริงนั้นทั้งหมด และค่าความถ่วงดุล (F1) นอกจากนี้วิธีการที่นำเสนอยังได้เปรียบเทียบประสิทธิภาพกับเทคนิคการปรับค่าน้ำหนักคำด้วยวิธีการอื่น ๆ ได้แก่ การแจกแจงเทอม (icsd, csd และ sd), วิธีการปรับค่าน้ำหนักด้วยมิวชวลอินฟอร์เมชัน (MI), วิธีการปรับค่าน้ำหนักคำด้วยเทคนิค rtf-sisf [18] และวิธีการ tf.rf [21] โดยรายละเอียดของแต่ละวิธีการปรับค่าน้ำหนักคำที่นำมาเปรียบเทียบประสิทธิภาพในงานวิจัยนี้แสดงรายละเอียดดังในตารางที่ 3

ตารางที่ 3 : สรุปวิธีการปรับค่าน้ำหนักคำที่ใช้ในการเปรียบเทียบผลการทดลอง

วิธีการ	สัญลักษณ์ที่ใช้	คำอธิบาย
การปรับค่าน้ำหนักคำแบบไม่มีผู้สอน	tf.idf	ใช้สมการ (5)
	rtf-sisf	ใช้สมการ (6)
การปรับค่าน้ำหนักคำแบบมีผู้สอน	tf.idf.icsd	ใช้สมการ (10)
	tf.idf.csd	ใช้สมการ (11)
	tf.idf.sd	ใช้สมการ (12)
	tf.rf	ใช้สมการ (13)
	tf.idf.MI	ใช้สมการ (15)
	tf.idf.cmi	วิธีการที่นำเสนอ

3.2) ผลการทดลองเมื่อเปรียบเทียบประสิทธิภาพบนตัวชี้วัด

ดังที่ได้อธิบายไปแล้วข้างต้นว่าในงานวิจัยนี้จะแบ่งชุดข้อมูลใช้สอนและชุดข้อมูลทดสอบออกเป็น 5 ชุด (5 Fold Cross Validation) โดยได้ทำการทดสอบประสิทธิภาพบนตัวจำแนกประเภทข้อมูล เบย์อย่างง่าย, ป่าของต้นไม้ตัดสินใจแบบสุ่ม และเพื่อนบ้านที่ใกล้ที่สุด K อันดับ โดยค่าความถูกต้องการทำนาย (Acc) เป็นการคำนวณจากค่าเฉลี่ยจากการทำนาย 5 ครั้ง พร้อมทั้งหาค่าส่วนเบี่ยงเบนมาตรฐาน และสำหรับตัวชี้วัดอื่น ๆ ก็จะมีการคำนวณหาค่าเฉลี่ยด้วยเช่นกัน

ตารางที่ 4 ได้แสดงการเปรียบเทียบประสิทธิภาพของตัวชี้วัดต่าง ๆ บนชุดข้อมูล Amazon ซึ่งพบว่าวิธีการที่นำเสนอ cmi สามารถปรับค่าน้ำหนักคำได้ดีกว่าวิธีการ tf.idf, rtf-sisf, MI และการแจกแจงเทอมทั้ง 3 แบบ ในทุกตัวจำแนกประเภทข้อมูล และให้ค่าตัวชี้วัด Acc, Pre, Rec และ F1 สูงที่สุดทุกค่าเมื่อเปรียบเทียบกับวิธีการอื่น ๆ บนตัวจำแนกประเภทเบย์อย่างง่าย และป่าของต้นไม้ตัดสินใจแบบสุ่ม ในขณะที่บนตัวจำแนกประเภทเพื่อนบ้านที่ใกล้ที่สุด K อันดับ วิธีการ tf.rf ให้ผลลัพธ์ของตัวชี้วัดที่สูงกว่า จึงกล่าวได้ว่าบนข้อมูลบทวิจารณ์ Amazon วิธีการปรับค่าน้ำหนักคำที่นำเสนอมีประสิทธิภาพที่ดีกว่าวิธีการอื่น ๆ บนตัวจำแนกประเภทข้อมูลทั้ง เบย์อย่างง่าย และป่าของต้นไม้ตัดสินใจแบบสุ่ม ในขณะที่บนตัวจำแนกประเภทข้อมูล เพื่อนบ้านที่ใกล้ที่สุด K อันดับ วิธีการที่นำเสนอให้ผลที่ดีกว่าวิธีการการปรับค่าน้ำหนักแบบไม่มีผู้สอน rtf-sisf, tf.idf และการปรับค่าน้ำหนักแบบมีผู้สอน ได้แก่ การแจกแจงเทอมทั้ง 3 แบบ และวิธีการ MI แสดงว่าวิธีการ cmi สามารถปรับค่าน้ำหนักคำได้มีประสิทธิภาพที่ดีกว่าวิธีการเหล่านี้

ตารางที่ 4 : ผลการทดลองเปรียบเทียบประสิทธิภาพตัวชี้วัดของวิธีการปรับค่าน้ำหนักบนชุดข้อมูล Amazon

วิธีการปรับค่าน้ำหนัก	เบย์อย่างง่าย				ป่าของต้นไม้ตัดสินใจแบบสุ่ม				เพื่อนบ้านที่ใกล้ที่สุด K อันดับ			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
rtf-sisf	0.689±0.02	0.705	0.693	0.683	0.753±0.02	0.756	0.757	0.752	0.685±0.04	0.692	0.688	0.681
tf×idf	0.713±0.02	0.730	0.714	0.708	0.786±0.03	0.788	0.791	0.785	0.650±0.08	0.686	0.648	0.624
tf×idf×icsd	0.749±0.02	0.760	0.750	0.746	0.779±0.02	0.781	0.784	0.778	0.715±0.03	0.716	0.716	0.713
tf×idf×sqrt(icsd)	0.745±0.02	0.760	0.746	0.741	0.779±0.03	0.781	0.784	0.782	0.740±0.04	0.747	0.742	0.737
tf×idf/csd	0.627±0.03	0.637	0.630	0.621	0.810±0.04	0.813	0.813	0.809	0.511±0.03	0.52	0.521	0.508
tf×idf/sqrt(csd)	0.617±0.03	0.630	0.621	0.610	0.775±0.03	0.777	0.780	0.774	0.534±0.02	0.545	0.543	0.531
tf×idf/sd	0.693±0.01	0.702	0.695	0.689	0.786±0.02	0.788	0.790	0.785	0.643±0.02	0.649	0.648	0.641
tf×idf/sqrt(sd)	0.723±0.03	0.731	0.726	0.720	0.775±0.02	0.776	0.779	0.775	0.683±0.04	0.689	0.69	0.683
tf×idf×cmi	**0.825±0.03**	**0.830**	**0.826**	**0.824**	**0.867±0.02**	**0.866**	**0.868**	**0.866**	0.753±0.03	0.755	0.752	0.750
tf×idf×sqrt(Miik)	0.799±0.03	0.808	0.800	0.797	0.833±0.02	0.834	0.836	0.832	0.710±0.04	0.718	0.71	0.706
tf×idf×MI	0.728±0.02	0.743	0.729	0.724	0.776±0.02	0.78	0.782	0.776	0.661±0.05	0.663	0.659	0.657
tf×idf×sqrt(MI)	0.725±0.02	0.740	0.726	0.721	0.782±0.02	0.784	0.787	0.782	0.666±0.05	0.668	0.664	0.662
tf×rf	0.795±0.02	0.799	0.798	0.794	0.795±0.04	0.8	0.8	0.794	**0.770±0.04**	**0.787**	**0.772**	**0.766**

ตารางที่ 5 : ผลการทดลองเปรียบเทียบประสิทธิภาพตัวชี้วัดของวิธีการปรับค่าน้ำหนักบนชุดข้อมูล Yelp

วิธีการปรับค่าน้ำหนัก	เบย์อย่างง่าย				ป่าของต้นไม้ตัดสินใจแบบสุ่ม				เพื่อนบ้านที่ใกล้ที่สุด K อันดับ			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
rtf-sisf	0.706±0.02	0.729	0.707	0.699	0.750±0.01	0.666	0.664	0.663	0.664±0.03	0.752	0.75	0.749
tf×idf	0.713±0.02	0.73	0.714	0.707	0.757±0.00	0.66	0.625	0.605	0.629±0.02	0.762	0.757	0.756
tf×idf×icsd	0.749±0.02	0.762	0.749	0.745	0.751±0.01	0.726	0.724	0.723	0.725±0.0	0.759	0.752	0.749
tf×idf×sqrt(icsd)	0.745±0.02	0.759	0.745	0.741	0.763±0.02	0.735	0.729	0.729	**0.731±0.01**	0.768	0.763	0.761
tf×idf/csd	0.647±0.03	0.664	0.648	0.638	0.775±0.01	0.521	0.517	0.511	0.519±0.04	0.778	0.774	0.773
tf×idf/sqrt(csd)	0.645±0.03	0.662	0.646	0.636	0.751±0.01	0.546	0.544	0.538	0.544±0.02	0.753	0.75	0.75
tf×idf/sd	0.700±0.02	0.717	0.701	0.694	0.761±0.01	0.653	0.648	0.644	0.647±0.01	0.766	0.761	0.759
tf×idf/sqrt(sd)	0.729±0.02	0.744	0.73	0.725	0.764±0.02	0.667	0.655	0.65	0.657±0.02	0.77	0.764	0.762
tf×idf×cmi	**0.825±0.03**	**0.829**	**0.825**	**0.824**	**0.858±0.01**	0.725	0.72	0.72	0.722±0.02	**0.867**	**0.858**	**0.857**
tf×idf×sqrt(Miik)	0.799±0.03	0.808	0.799	0.797	0.823±0.02	0.691	0.685	0.684	0.688±0.02	0.831	0.823	0.822
tf×idf×MI	0.728±0.02	0.742	0.729	0.724	0.764±0.02	0.631	0.63	0.629	0.632±0.03	0.768	0.764	0.763
tf×idf×sqrt(MI)	0.725±0.02	0.74	0.725	0.72	0.755±0.01	0.631	0.624	0.621	0.627±0.03	0.76	0.755	0.753
tf×rf	0.802±0.03	0.81	0.803	0.801	0.792±0.01	**0.809**	**0.792**	**0.789**	0.716±0.05	0.719	0.714	0.714

เมื่อพิจารณาการการจำแนกประเภทความพึงพอใจที่ทดสอบกับข้อมูล Yelp ดังตารางที่ 5 พบว่าวิธีการ cmi ให้ค่า Acc สูงที่สุดในตัวจำแนกประเภทข้อมูล เบย์อย่างง่าย และป่าของต้นไม้ตัดสินใจแบบสุ่ม และให้ค่าตัวชี้วัด Pre, Rec และ F1 สูงกว่าวิธีการ tf.idf, rtf-sisf, MI, การแจกแจงเทอมทั้ง 3 แบบและ tf.rf บนตัวจำแนกประเภท เบย์อย่างง่าย และ เพื่อนบ้านที่ใกล้ที่สุด K อันดับ ในขณะที่การทดสอบบนป่าของต้นไม้ตัดสินใจแบบสุ่มนั้นวิธีการ tf.rf ให้ค่าที่ดีกว่าบนตัวชี้วัด Pre, Rec และ F

ในขณะที่วิธีการ icsd ซึ่งเป็นการแจกแจงเทอมรูปแบบหนึ่งให้ค่าความถูกต้องในการทำนายสูงที่สุดบนเพื่อนบ้านที่ใกล้ที่สุด K อันดับ เมื่อพิจารณาประสิทธิภาพโดยรวมบนข้อมูลบทวิจารณ์ร้านอาหาร Yelp การปรับค่าน้ำหนักคำด้วยวิธีการ cmi ให้ผลลัพธ์ที่ดีกว่า rtf-sisf และ tf.idf อย่างชัดเจน เนื่องจากวิธีการเหล่านี้เป็นการปรับค่าน้ำหนักเทอมแบบไม่มีผู้สอนและพิจารณาความถี่เป็นสำคัญ ซึ่งย่อมให้ผลที่ไม่ดีกว่าวิธีการแบบมีผู้สอน นอกจากนี้วิธีการ cmi ยังให้ผลการทดลองที่ดีกว่าการแจกแจงเทอม (csd และ sd) และวิธีการ MI อย่างเห็นได้ชัดและบางครั้งให้ผลที่ดีกว่า tf.rf ในกรณีส่วนใหญ่

ตารางที่ 6 : ผลการทดลองเปรียบเทียบประสิทธิภาพตัวชี้วัดของวิธีการปรับค่าน้ำหนักบนชุดข้อมูล IMDb

วิธีการปรับค่าน้ำหนัก	เบย์อย่างง่าย				ป่าของต้นไม้ตัดสินใจแบบสุ่ม				เพื่อนบ้านที่ใกล้ที่สุด K อันดับ			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
rtf-sisf	0.674±0.02	0.683	0.679	0.673	0.731±0.02	0.739	0.737	0.731	0.618±0.04	0.651	0.625	0.603
tf×idf	0.683±0.04	0.687	0.686	0.682	0.786±0.03	0.788	0.791	0.785	0.520±0.05	0.626	0.535	0.424
tf×idf×icsd	0.731±0.03	0.734	0.734	0.731	0.779±0.02	0.781	0.784	0.778	0.699±0.02	0.702	0.702	0.698
tf×idf×sqrt(icsd)	0.722±0.03	0.724	0.725	0.721	0.779±0.03	0.781	0.784	0.738	0.623±0.03	0.644	0.63	0.614
tf×idf/csd	0.494±0.07	0.494	0.494	0.491	0.750±0.02	0.753	0.754	0.75	0.433±0.03	0.421	0.442	0.403
tf×idf/sqrt(csd)	0.535±0.03	0.538	0.538	0.533	0.723±0.01	0.727	0.727	0.723	0.461±0.05	0.449	0.469	0.422
tf×idf/sd	0.670±0.03	0.674	0.673	0.669	0.754±0.02	0.757	0.758	0.754	0.547±0.04	0.575	0.559	0.518
tf×idf/sqrt(sd)	0.686±0.03	0.689	0.689	0.685	0.719±0.02	0.724	0.724	0.719	0.554±0.03	0.578	0.564	0.536
tf×idf×cmi	**0.852±0.03**	**0.851**	**0.853**	**0.851**	**0.867±0.02**	**0.866**	**0.868**	**0.866**	**0.753±0.03**	**0.754**	**0.751**	**0.75**
tf×idf×sqrt(Miik)	0.826±0.03	0.826	0.827	0.826	0.833±0.02	0.834	0.836	0.832	0.71±0.04	0.717	0.71	0.706
tf×idf×MI	0.687±0.04	0.689	0.69	0.686	0.776±0.02	0.78	0.782	0.718	0.66±0.05	0.662	0.659	0.657
tf×idf×sqrt(MI)	0.682±0.04	0.684	0.685	0.681	0.782±0.02	0.784	0.787	0.782	0.665±0.05	0.668	0.663	0.661
tf×rf	0.789±0.03	0.79	0.791	0.788	0.757±0.05	0.771	0.764	0.755	0.670±0.04	0.737	0.679	0.649

ตารางที่ 7 : ผลการทดลองเปรียบเทียบประสิทธิภาพตัวชี้วัดของวิธีการปรับค่าน้ำหนักบนชุดข้อมูล Magazine

วิธีการปรับค่าน้ำหนัก	เบย์อย่างง่าย				ป่าของต้นไม้ตัดสินใจแบบสุ่ม				เพื่อนบ้านที่ใกล้ที่สุด K อันดับ			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
rtf-sisf	0.627±0.01	0.514	0.549	0.455	0.938±0.01	0.877	0.618	0.666	0.943±0.01	0.895	0.877	0.7
tf×idf	0.625±0.01	0.514	0.548	0.455	0.940±0.02	0.927	0.612	0.663	0.938±0.02	0.903	0.596	0.637
tf×idf×icsd	0.700±0.01	0.531	0.593	0.501	0.941±0.01	0.944	0.612	0.663	0.942±0.02	0.844	0.666	0.713
tf×idf×sqrt(icsd)	0.637±0.01	0.515	0.549	0.46	0.941±0.02	0.944	0.618	0.671	0.939±0.02	0.852	0.621	0.669
tf×idf/csd	0.603±0.01	0.506	0.52	0.438	0.945±0.02	0.944	0.642	0.7	0.840±0.01	0.513	0.511	0.509
tf×idf/sqrt(csd)	0.610±0.01	0.509	0.532	0.444	0.940±0.02	0.891	0.617	0.668	0.908±0.01	0.616	0.559	0.572
tf×idf/sd	0.619±0.01	0.516	0.555	0.454	0.941±0.01	0.944	0.612	0.663	0.938±0.01	0.918	0.595	0.639
tf×idf/sqrt(sd)	0.635±0.01	0.511	0.538	0.455	0.942±0.02	0.941	0.623	0.676	0.932±0.01	0.783	0.617	0.648
tf×idf×cmi	0.729±0.01	0.542	0.623	0.523	0.971±0.01	**0.978**	**0.825**	**0.882**	**0.965±0.01**	**0.974**	**0.773**	**0.839**
tf×idf×sqrt(Miik)	0.682±0.01	0.526	0.582	0.489	0.958±0.02	0.978	0.73	0.797	0.943±0.02	0.96	0.621	0.669
tf×idf×MI	0.680±0.02	0.523	0.575	0.486	0.940±0.01	0.940	0.606	0.656	0.936±0.01	0.815	0.603	0.645
tf×idf×sqrt(MI)	0.630±0.01	0.515	0.551	0.457	0.942±0.02	0.944	0.623	0.677	0.936±0.02	0.906	0.582	0.619
tf×rf	**0.904±0.02**	**0.662**	**0.69**	**0.67**	**0.986±0.00**	0.798	0.707	0.736	0.783±0.03	0.558	0.636	0.555

อย่างไรก็ดีเมื่อพิจารณาเมื่อพิจารณาดูประสิทธิภาพของวิธีการต่าง ๆ ในการปรับค่าน้ำหนักคำบนข้อมูลบทวิจารณ์ IMDb ดังในตารางที่ 6 จะพบว่าวิธีการที่นำเสนอให้ผลลัพธ์ที่ดีที่สุดในทุกตัวชี้วัดสำหรับการวัดประสิทธิภาพการสืบค้นเอกสารข้อความบทวิจารณ์และมีประสิทธิภาพสูงกว่าทุกวิธีที่มีการเปรียบเทียบและดีกว่าอย่างน่าประทับใจ โดยเฉพาะอย่างยิ่งเมื่อเปรียบเทียบกับวิธีการ rtf-sisf, tf.idf, การแจกแจงเทอมทั้ง 3 วิธี, MI และ tf.rf ซึ่งเป็นเครื่องบ่งชี้ได้ว่าหลักการที่นำเสนอนี้สามารถปรับปรุงการปรับค่าน้ำหนักคำ tf.idf ให้มีพลังอำนาจในการจำแนกเอกสารข้อความบทวิจารณ์ของคลาสบวกออกจากเอกสารข้อความคลาสลบได้อย่างมีประสิทธิภาพ

ในตารางที่ 7 เป็นการพิจารณาผลการทดลองบนชุดข้อมูลบทวิจารณ์ Magazine ซึ่งพบว่าโดยรวมแล้ววิธีการที่นำเสนอก็ยังคงให้ประสิทธิภาพของตัวชี้วัดดีกว่าวิธีการอื่น ๆ บนตัวจำแนกประเภทเอกสารข้อความ ป่าของต้นไม้ตัดสินใจแบบสุ่ม และ เพื่อนบ้านที่ใกล้ที่สุด K อันดับ ในขณะที่วิธี tf.rf ให้ผลของประสิทธิภาพของตัวชี้วัดดีกว่าบนตัวจำแนกประเภทเอกสารข้อความเบย์อย่างง่ายและให้ Acc ดีที่สุดบนป่าของต้นไม้ตัดสินใจแบบสุ่ม

ตารางที่ 8 : ผลการทดลองเปรียบเทียบประสิทธิภาพตัวชี้วัดของวิธีการปรับค่าน้ำหนักบนชุดข้อมูล Eco-Pre

วิธีการปรับค่าน้ำหนัก	เบย์อย่างง่าย				ป่าของต้นไม้ตัดสินใจแบบสุ่ม				เพื่อนบ้านที่ใกล้ที่สุด K อันดับ			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
rtf-sisf	0.502±0.03	0.456	0.52	0.437	0.713±0.00	0.769	0.726	0.726	0.854±0.00	0.571	0.59	0.55
tf×idf	0.529±0.03	0.476	0.537	0.461	0.462±0.01	0.797	0.695	0.724	0.858±0.00	0.548	0.48	0.379
tf×idf×icsd	0.649±0.02	0.552	0.618	0.555	0.803±0.01	0.804	0.697	0.727	0.859±0.01	0.656	0.664	0.657
tf×idf×sqrt(icsd)	0.573±0.03	0.501	0.563	0.494	0.771±0.01	0.79	0.687	0.716	0.853±0.00	0.645	0.644	0.621
tf×idf/csd	0.485±0.03	0.446	0.507	0.424	0.535±0.01	0.817	0.746	0.76	0.877±0.01	0.392	0.462	0.379
tf×idf/sqrt(csd)	0.488±0.03	0.447	0.508	0.425	0.480±0.02	0.797	0.721	0.738	0.862±0.00	0.441	0.464	0.359
tf×idf/sd	0.517±0.03	0.468	0.529	0.451	0.620±0.01	0.804	0.693	0.723	0.858±0.00	0.529	0.481	0.441
tf×idf/sqrt(sd)	0.550±0.03	0.483	0.544	0.474	0.645±0.01	0.801	0.693	0.723	0.857±0.00	0.587	0.577	0.492
tf×idf×cmi	**0.832±0.01**	**0.758**	**0.753**	**0.738**	0.815±0.01	**0.881**	**0.828**	**0.842**	**0.917±0.00**	0.708	0.676	0.661
tf×idf×sqrt(Miik)	0.690±0.01	0.627	0.642	0.605	0.761±0.01	0.858	0.77	0.797	0.893±0.00	0.689	0.638	0.606
tf×idf×MI	0.610±0.02	0.517	0.579	0.517	0.746±0.01	0.798	0.699	0.726	0.859±0.00	0.541	0.53	0.531
tf×idf×sqrt(MI)	0.557±0.03	0.489	0.552	0.48	0.698±0.02	0.798	0.699	0.728	0.857±0.01	0.517	0.527	0.507
tf×rf	0.744±0.014	0.624	0.64	0.621	**0.906±0.01**	0.829	0.818	0.82	0.893±0.013	**0.825**	**0.82**	**0.812**

จากการทดลองพบว่าวิธีการแจกแจงเทอม icsd นั้นจะให้ผลที่ดีกว่า rtf-sisf, tf.idf, csd และ sd ซึ่งก็สอดคล้องกับงานวิจัยก่อนหน้านี้ที่ได้ศึกษากันมา [9], [10], [19], [20] โดยปกติแล้วหลักการแจกแจงเทอมจะพยายามเพิ่มค่าน้ำหนักของคำที่มาจากคลาสเดียวกันและลดค่าน้ำหนักของคำที่อยู่คนละคลาสกันโดยใช้หลักการทางสถิติโดยการคำนวณหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคำสำหรับแต่ละเอกสาร แต่อย่างไรก็ดีเมื่อเปรียบเทียบวิธีการ icsd กับ cmi แล้วจะพบว่าวิธีการ cmi ให้ผลการทดลองที่ดีกว่า icsd อย่างเห็นได้ชัดในทุกตัวชี้วัด ซึ่งเป็นสิ่งที่ยืนยันได้ว่าวิธีการปรับค่าน้ำหนักคำ rtf-sisf, tf.idf, icsd, csd และ sd ไม่เหมาะที่จะนำมาประยุกต์ใช้ในการจำแนกประเภทความพึงพอใจของเอกสารบทวิจารณ์ โดยวิธีการ cmi และ tf.rf จะมีความเหมาะสมกว่าตามลำดับ

เมื่อพิจารณาตารางที่ 8 เป็นผลการทดลองการเปรียบเทียบตัวชี้วัดบนการทดสอบกับข้อมูลบทวิจารณ์ Eco-Pre ซึ่งเป็นข้อมูลที่ประกอบด้วย 3 คลาส คือ เชิงบวก เชิงลบ และเป็นกลาง โดยหลักการแล้วการจำแนกประเภทความพึงพอใจงานวิจัยโดยส่วนใหญ่จะพิจารณาเฉพาะข้อมูลที่มี 2 คลาสเท่านั้น คือ เชิงบวกและเชิงลบ เนื่องจากคลาสเป็นกลางไม่ได้สะท้อนถึงความรู้สึกที่แท้จริงของผู้บริโภคที่มีต่อสินค้าในการปรับปรุงและพัฒนาผลิตภัณฑ์และบริการ แต่ในการทดลองนี้จะนำข้อมูลนี้เพื่อมาเปรียบเทียบเพื่อดูประสิทธิภาพของตัวชี้วัดต่าง ๆ เมื่อข้อมูลประกอบด้วย 3 คลาส ว่าส่งผลอย่างไรต่อวิธีการกำหนดค่าน้ำหนักคำของวิธีการต่าง ๆ ทั้งในเรื่องของประสิทธิภาพของตัวชี้วัดการสืบค้นเอกสารข้อความและในเรื่องของเวลาที่ใช้ในการคำนวณ ซึ่งวิธีการปรับค่าน้ำหนักคำแบบมีผู้สอน cmi และ tf.rf ให้ค่าตัวชี้วัดสูงกว่าและมีประสิทธิภาพดีกว่าวิธีการอื่น ๆ ซึ่งเมื่อเปรียบเทียบกับวิธีการ rtf-sisf, tf.idf, การแจกแจงเทอมทั้ง 3 วิธี และ MI ซึ่งเป็นสิ่งที่ยืนยันได้ว่าการพิจารณาค่าความถี่ของคำและความถี่เอกสารผกผัน (rtf-sisf และ tf.idf) จะมีประสิทธิภาพไม่เพียงพอในการจำแนกประเภทเอกสารข้อความบทวิจารณ์ เนื่องจากข้อความบทวิจารณ์นั้นจะมีประโยคที่ยาวไม่มากและคำที่มีพลังอำนาจในการแยกแยะคลาสเชิงบวกและเชิงลบก็ไม่ได้เกิดขึ้นบ่อยและไม่ค่อยเกิดขึ้นซ้ำ ๆ ในประโยคเหมือนกับในเอกสารอื่น ๆ เช่น หนังสือ หรือ หนังสือพิมพ์

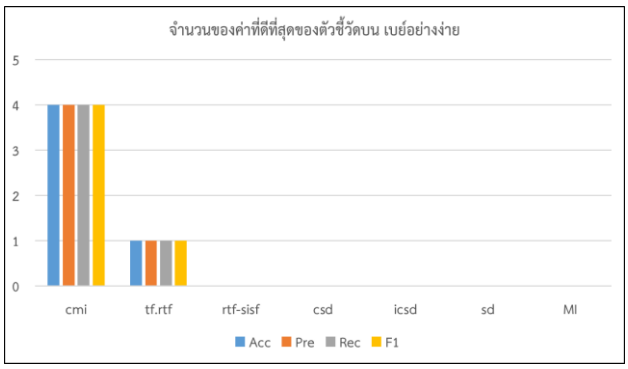
ในขณะที่วิธีการ tf.rf เมื่อพิจารณาจากการทดลองบนข้อมูล Eco-Pre พบว่าให้ผลลัพธ์ที่ดีที่รองจากวิธีการที่นำเสนอเนื่องจากให้ค่าความถูกต้องจากการทำนาย Acc สูงที่สุดบน ป่าของต้นไม้ตัดสินใจแบบสุ่ม และบนตัวจำแนกประเภทเอกสารข้อความเพื่อนบ้านที่ใกล้ที่สุด K อันดับ ก็ให้ค่าตัวชี้วัด Pre, Rec และ F1 ดีที่สุด ในขณะที่วิธีการ cmi ให้ค่า Acc ดีที่สุดบน เบย์อย่างง่าย และ เพื่อนบ้านที่ใกล้ที่สุด K อันดับ และสำหรับค่าตัวชี้วัด Pre, Rec และ F1 ก็ให้ผลลัพธ์ที่ดีที่สุดบนตัวจำแนกประเภทเอกสารข้อความบน เบย์อย่างง่าย และ ป่าของต้นไม้ตัดสินใจแบบสุ่ม

จากการทดลองบนชุดข้อมูล Eco-Pre ในเรื่องของเวลาในการคำนวณสามารถพิจารณาได้จากตารางที่ 9 ซึ่งพบว่าวิธีการแจกแจงเทอมจะใช้เวลาในการคำนวณนานที่สุด พบว่าวิธีการแจกแจงเทอมทั้ง 3 แบบใช้เวลาคำนวณนานกว่าวิธีการที่นำเสนอ โดยใช้เวลานานกว่าประมาณ 800 เท่าสำหรับวิธีการ icsd ในขณะที่ csd และ sd จะใช้เวลานานกว่า 1500 เท่า ทั้งนี้เวลาในการ

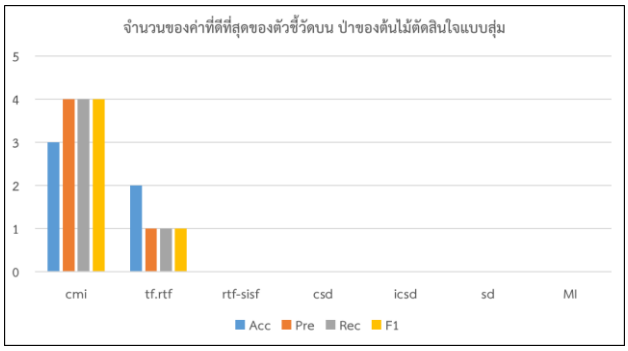
คำนวณนั้นก็จะเพิ่มขึ้นอย่างทวีคูณตามจำนวนเอกสาร จำนวนคำ และจำนวนคลาสของเอกสาร ดังนั้นวิธีการปรับค่าน้ำหนักด้วยการแจกแจงเทอมจึงไม่เหมาะที่จะนำไปใช้กับข้อมูลเอกสารขนาดใหญ่และยิ่งมีจำนวนคลาสมากก็ยิ่งช้าเป็นทวีคูณ

ตารางที่ 9 : การเปรียบเทียบเวลาในการคำนวณการปรับค่าน้ำหนัก

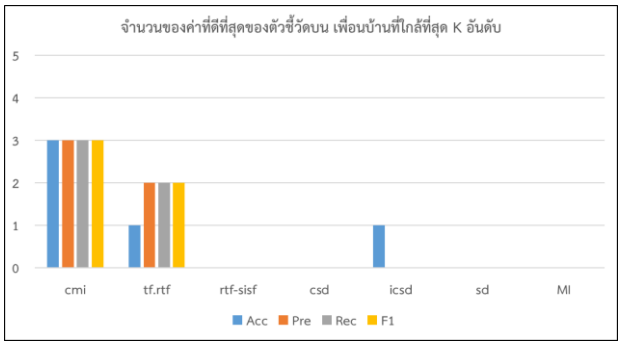
วิธีการปรับค่าน้ำหนัก	เวลาที่ใช้ประมวลผล (วินาที)
csd	26,020
sd	26,020
icsd	14,241
cmi	17
tf.idf, tf.rt, rtf-sisf	< 17



รูปที่ 7 : จำนวนของค่าที่ดีที่สุดของตัวชี้วัดประสิทธิภาพการสืบค้นเอกสารข้อความบนเบย์อย่างง่าย



รูปที่ 8 : จำนวนของค่าที่ดีที่สุดของตัวชี้วัดประสิทธิภาพการสืบค้นเอกสารข้อความบนป่าของต้นไม้ตัดสินใจแบบสุ่ม



รูปที่ 9 : จำนวนของค่าที่ดีที่สุดของตัวชี้วัดประสิทธิภาพการสืบค้นเอกสารข้อความบนเพื่อนบ้านที่ใกล้ที่สุด K อันดับ

รูปที่ 7–9 ได้สรุปถึงประสิทธิภาพตัวชี้วัดของการสืบค้นเอกสาร โดยแสดงค่าที่สูงที่สุดของแต่ละวิธีที่ได้รับบนตัวจำแนกประเภทเอกสารข้อความ เบย์อย่างง่าย, ป่าของต้นไม้ตัดสินใจแบบสุ่ม และ เพื่อนบ้านที่ใกล้ที่สุด K อันดับ ที่ได้มีการทดสอบกับชุดข้อมูลบทวิจารณ์ทั้ง 5 ชุด เมื่อพิจารณาจากกราฟที่แสดงนั้นจะเห็นได้ว่าการปรับค่าน้ำหนัก ด้วยคลาสมิวชวลอินฟอร์เมชัน cmi จะมีประสิทธิภาพที่ดีกว่าวิธีการปรับค่าน้ำหนักวิธีการอื่น ๆ โดยเฉพาะอย่างยิ่งเมื่อเปรียบเทียบกับวิธีการปรับค่าน้ำหนักคำแบบดั้งเดิมที่นิยมใช้กัน tf.idf และ rtf-sisf ที่พิจารณาใช้ความถี่ของคำและความถี่เอกสารผกผัน รวมถึงการแจกแจงเทอมทั้ง 3 แบบ (icsd, csd, และ sd) ที่ใช้กับหลักการทางสถิติ ซึ่งเป็นสิ่งที่ยืนยันถึงประสิทธิภาพของวิธีการปรับค่าน้ำหนักคำที่นำเสนอนี้ได้เป็นอย่างดี

5) สรุปผลการทดลอง

งานวิจัยนี้มีแนวคิดโดยใช้มิวชวลอินฟอร์เมชันในการหาความสัมพันธ์ร่วมระหว่างคำและคลาสของประเภทความพึงพอใจ การปรับค่าน้ำหนักคำ tf.idf ด้วยวิธีการที่นำเสนอนี้ช่วยเพิ่มประสิทธิภาพในการสืบค้นเอกสารข้อความบทวิจารณ์ อย่างไรก็ตามวิธีการนี้จะต้องมีการปรับค่าความถี่คำให้เป็นแบบไบนารีก่อนเพื่อคำนวณค่าความน่าจะเป็นจึงทำให้ต้องเสียเวลาในขั้นตอนนี้ รวมถึงอาจจะทำให้ข่าวสารบางอย่างหายไป ซึ่งงานวิจัยในอนาคตจะคำนวณค่ามิวชวลอินฟอร์เมชันโดยตรงจากค่าความถี่ของคำด้วยเทคนิคอื่น ๆ เช่น ฟัซซี่เซต (Fuzzy sets) หรือ การให้เหตุผลบนหลักการกรณีศึกษา (Case based reasoning) เป็นต้น

เทคนิคที่นำเสนอนี้ให้ความสำคัญกับจำนวนเอกสารที่อยู่ในคลาสบวก ถ้าคำใดให้ค่าข่าวสารร่วมกับคลาสบวกมากก็จะถูก

ปรับค่าน้ำหนักคำเพื่อเพิ่มอำนาจการจำแนกให้มากขึ้น ซึ่งต่างจากวิธี tf.rf ที่พิจารณาสัดส่วนจำนวนของเอกสารคลาสบวกต่อคลาสลบ ซึ่งอาจจะประสบปัญหาในกรณีที่คำต่าง ๆ มีสัดส่วนของคลาสบวกต่อคลาสลบเท่ากัน ซึ่งก็จะมีการปรับค่าน้ำหนักคำให้เท่ากัน ในขณะที่วิธีการ cmi อาจจะประสบปัญหาในกรณีที่คำต่าง ๆ มีจำนวนเอกสารในคลาสบวกเท่ากันแต่ในคลาสลบแตกต่างกันก็จะมีการปรับค่าน้ำหนักคำเหล่านั้นเท่ากัน ซึ่งถ้ามีการพิจารณาคลาสลบด้วยจะสามารถช่วยปรับค่าน้ำหนักคำต่าง ๆ ให้เหมาะสมได้

การทดลองการจำแนกประเภทความพึงพอใจด้วยเบย์อย่างง่ายพบว่าวิธีการที่นำเสนอให้ค่าความถูกต้องเฉลี่ยและค่าเฉลี่ยของตัวชี้วัดอื่น ๆ สูงกว่าวิธีการต่าง ๆ แสดงว่าวิธีการที่นำเสนอสามารถทำให้คำในถุงของคำมีความเป็นอิสระกันมากว่าวิธีการอื่น ๆ ในขณะที่การวัดประสิทธิภาพบนป่าของต้นไม้ตัดสินใจแบบสุ่ม ก็ให้ผลใกล้เคียงกับเบย์อย่างง่าย เนื่องจากตัวจำแนกประเภทชนิดนี้มีการพิจารณาความสำคัญของคำเช่นเดียวกับเทคนิคที่นำเสนอ และเมื่อพิจารณาจากเพื่อนบ้านที่ใกล้ที่สุด K อันดับ พบว่าวิธีการ tf.rf และ icsd ให้ผลลัพธ์ซึ่งบางครั้งสูงกว่าวิธีการที่นำเสนอ เนื่องจากวิธีการ tf.rf นี้ช่วยปรับให้เอกสารที่อยู่ในคลาสเดียวกันอยู่ใกล้กัน ในขณะที่วิธีการ icsd จะพยายามทำให้กลุ่มของคำที่มาจากคลาสเดียวกันมีความกระชับและอยู่ใกล้กัน ดังนั้นจากผลการทดลองสามารถกล่าวได้ว่า วิธีการแจกแจงคำทั้ง 3 แบบ รวมถึง tf.idf และ rtf-sisf จึงไม่เหมาะที่จะนำมาใช้ในการจำแนกประเภทความพึงพอใจเนื่องจากแต่ละบทวิจารณ์เป็นประโยคไม่ยาว โดยวิธีการที่นำเสนอนี้มีความเหมาะสมกว่าทั้งในแง่ประสิทธิภาพและเวลาที่ใช้ จึงเหมาะกับการวิเคราะห์ความพึงพอใจแบบประมวลผลทันทีเพื่อการปรับปรุงการให้บริการและคุณภาพของสินค้าให้ตรงกับความต้องการของลูกค้าและทันสถานการณ์มากที่สุด

REFERENCES

[1] O. Gokalp, E. Tasci, and A. Ugur, “A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification,” *Expert Syst. Appl.*, vol. 146, May 2020, Art. no. 113176.

[2] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.,* vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[3] D. M. E.-D. M. Hussein, “A survey on sentiment analysis challenges,” *J. King Saud Univ. – Eng. Sci.*, vol. 30, no. 4, pp. 330–338, Oct. 2018.

[4] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, “Sentiment classification: The contribution of ensemble learning,” *Decis. Support Syst.*, vol. 57, pp. 77–93, Jan. 2014.

[5] J. Chen, J. Yu, S. Zhao, and Y. Zhang, “User’s review habits enhanced hierarchical neural network for document-level sentiment classification,” *Neural Process. Lett.*, vol. 53, pp. 2095–2111, Apr. 2021.

[6] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. A. Larson, “POS-RS: A random subspace method for sentiment classification based on part-of-speech analysis,” *Inf. Process. Manage.*, vol. 51, no. 4, pp. 458–479, Jul. 2015.

[7] C. Yang, X. Chen, L. Liu, and P. Sweetser, “Leveraging semantic features for recommendation: Sentence-level emotion analysis,” *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102543.

[8] R. K. Yadav, L. Jiao, O.-C. Granmo, and M. Goodwin, “Human-level interpretable learning for aspect-based sentiment analysis,” in *Proc. 35th AAAI Conf. Artif. Intell. (AAAI-21)*, Palo Alto, CA, USA, Feb. 2021, pp. 14203–14212.

[9] V. Lertnattee and T. Theeramunkong, “Effect of term distributions on centroid-based text categorization,” *Inf. Sci.*, vol. 158, pp. 89–115, Jan. 2004.

[10] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, “Improving seeded k-means clustering with deviation- and entropy-based term weightings,” *IEICE Trans. Inf. Syst.*, vol. E103–D, no. 4, pp. 748–758, Apr. 2020.

[11] H. Liu, X. Chen, and X. Liu, “A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis,” *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[12] I. Almalis, E. Kouloumpris, and I. Vlahavas, “Sector-level sentiment analysis with deep learning,” *Knowl.-Based Syst.*, vol. 258, 2022, Art. no. 109954.

[13] Y. Dang, Y. Zhang, and H. Chen, “A lexicon-enhanced method for sentiment classification: An experiment on online product reviews,” *IEEE Intell. Syst.*, vol. 25, no. 4, pp. 46–53, 2010.

[14] S. Foithong, O. Pinngern, and B. Attachoo, “Feature subset selection wrapper based on mutual information and rough sets,” *Expert Syst. Appl.*, vol. 39, no. 1, pp. 574–584, 2012.

[15] A. K. Paul and P. C. Shill, "Sentiment mining from Bangla data using mutual information," in *Proc. 2nd Int. Conf. Elect., Comput. & Telecommun. Eng. (ICECTE)*, Rajshahi, Bangladesh, Dec. 2016, pp. 1–4.

[16] X. -Y. Jiang and J. Shui, "An improved mutual information-based feature selection algorithm for text classification," in *Proc. 5th Int. Conf. Intell. Human-Mach. Syst. and Cybernetics*, Hangzhou, China, Aug. 2013, pp. 126–129.

[17] A. Bagheri, M. Saraee, and F. de Jong, "Sentiment classification in Persian: Introducing a mutual information-based method for feature selection," in *Proc. 21st Iranian Conf. Elect. Eng. (ICEE)*, Mashhad, Iran, May 2013, pp. 1–6.

[18] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "The impact of term-weighting schemes and similarity measures on extractive multi-document text summarization," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114510.

[19] V. Lertnattee and T. Theeramunkong, "Multidimensional text classification for drug information," *IEEE Trans. Inf. Technol. Biomed.,* vol. 8, no. 3, pp. 306–312, Sep. 2004, doi: 10.1109/TITB.2004.832542.

[20] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Document clustering using k-means with term weighting as similarity-based constraints," *Symmetry*, vol. 12, no. 6, p. 967, 2020.

[21] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 31, no. 4, pp. 721–735, Apr. 2009.

[22] Z. Feng, H. Zhou, Z. Zhu, and K. Mao, " Tailored text augmentation for sentiment analysis," *Expert Syst. Appl*., vol. 205, Nov. 2022, Art. no. 117605.

[23] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Inf. Process. Manage*., vol. 58, no. 5, Sep. 2021, Art. no. 102656.

[24] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Inf. Fusion*, vol. 44, pp. 65–77, 2018.

[25] D. Kotzias, M. Denil, N. D. Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining*, Sydney, Australia, Aug. 2015, pp. 597–606.