

การวิเคราะห์ความรู้สึกแบบอัตโนมัติจากข้อความแสดงความคิดเห็น

Automatic Feeling Analysis from Opinion Text

เจตรินทร์ วงศ์ศิลป์(Chetarin Wongsin)¹ ณัฐกิตติ ศรีกาญจน์เพียรศ(Nattakit Srikanjanaperd)²

และจันทิมา พลพินิจ(Jantima Polpinij)³

ห้องปฏิบัติการวิจัยเชิงปัญญา (Intellect Laboratory) คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

¹chetarin.1995@gmail.com, ²armong.ton@icloud.com, ³jantima.p@msu.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอกระบวนการของการวิเคราะห์ความรู้สึก เพื่อวิเคราะห์ความรู้สึกของนักท่องเที่ยว ภายหลังจากที่ได้เยี่ยมชมสถานที่ท่องเที่ยวในภาคตะวันออกเฉียงเหนือ วัตถุประสงค์ของงานวิจัยฉบับนี้ เพื่อวิเคราะห์ความรู้สึกของนักท่องเที่ยวที่มีการเขียนในรูปแบบของข้อความบนโซเชียลมีเดียเป็น 2 กลุ่มความรู้สึก คือความรู้สึกที่ดี และความรู้สึกที่ไม่ดี ต่อสถานที่ท่องเที่ยวที่ได้ไปเยี่ยมชม โดยงานวิจัยฉบับนี้ได้พัฒนาโมเดลเพื่อการวิเคราะห์ความรู้สึกด้วยตัวจัดกลุ่มเอกสารแบบนาอิวเบย์ หลังจากทดสอบด้วยการวัดค่าเอฟ-เมตริกซ์ได้แสดงให้เห็นว่า วิธีที่นำเสนอสามารถให้มีประสิทธิภาพมากขึ้นสำหรับการวิเคราะห์ความรู้สึกที่อยู่ในเนื้อหาข้อความ เนื่องจากให้ความถูกต้องของการวิเคราะห์ความรู้สึกที่น่าพอใจ

คำสำคัญ: การวิเคราะห์ความรู้สึก การจัดกลุ่มข้อความ นาอิวเบย์

Abstract

This paper presents a method of sentiment analysis to analyse the visitors' feeling after visiting the Tourist attraction in the Northeastern, Thailand. The objective of this work is to analyse the visitors' feeling that are represented by text on a social media into two classes: positive and negative feeling. This work implements the model of sentiment analysis based on the Naïve Bayes text classification. After testing by F-measure, it could be demonstrated that the proposed method can provide more effectiveness for sentiment analysis

based on text content. This is because the proposed method returns the satisfactory accuracy.

Keyword: Sentiment Analysis, Text Classification, Naïve Bayes.

1. บทนำ

ในปัจจุบันการวิเคราะห์ความรู้สึกของผู้คน (Sentiment Analysis) [1-5] ที่มีต่อสินค้าและบริการ ซึ่งเป็นงานวิจัยสาขาหนึ่งด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ที่มีกระบวนการมุ่งเน้นการวิเคราะห์และตรวจสอบความรู้สึก (Opinion) ของผู้คน จากข้อความ (Text) ที่คนเหล่านั้นเขียนหรือโพสต์เอาไว้เพื่อบ่งบอกความรู้สึกของตนเองที่มีต่อบางสิ่งบางอย่าง เช่น ความรู้สึกดี (Positive หรือ Good) หรือความรู้สึกที่ไม่ดีหรือไม่ชอบ (Negative หรือ Bad) [1-5]

ปัจจุบันเทคนิคการวิเคราะห์ความรู้สึกของผู้คน กำลังได้รับความสนใจและมีการประยุกต์อย่างแพร่หลาย รวมถึงการประยุกต์ไปสู่ระบบ e-Tourism เป็นเครื่องมือที่ใช้ในการวิเคราะห์ความรู้สึกที่แท้จริงของนักท่องเที่ยว เนื่องจากข้อความนักท่องเที่ยวเขียนไว้นั้น จะบ่งบอกความรู้สึกที่แท้จริงของคนเหล่านั้นว่า “ชอบ” หรือ “ไม่ชอบ” ต่อการให้บริการด้านการท่องเที่ยว หรือสถานที่ท่องเที่ยวที่นักท่องเที่ยวเหล่านั้นได้ไปเยี่ยมชม ถ้าผู้ให้บริการด้านการท่องเที่ยวให้ความสนใจกับข้อความเหล่านี้ สามารถนำคำติชมเหล่านี้ไปเป็นองค์ประกอบสำคัญต่อการปรับปรุงการให้บริการด้านการท่องเที่ยวของตน หรือสร้างฐานการท่องเที่ยวของตน รวมถึงการขยายฐานการท่องเที่ยวไปยังนักท่องเที่ยวคนอื่นๆ ให้กว้างขึ้น เพราะข้อความ

ต่างๆ ที่นักท่องเที่ยวยกย่องไว้ มักจะส่งผลต่อความต้องการในการท่องเที่ยวของนักท่องเที่ยวคนอื่นๆ ด้วย

สำหรับงานวิจัยฉบับนี้นำเสนอการวิเคราะห์ความรู้สึกของนักท่องเที่ยวต่อสถานที่ท่องเที่ยวในประเทศไทย ที่มีการบันทึกข้อความเอาไว้บน Social Media โดยงานวิจัยนี้ที่นำเสนอจะวิเคราะห์ข้อความภาษาไทยว่า นักท่องเที่ยว “ชอบ” หรือ “ไม่ชอบ” สถานที่ท่องเที่ยวที่ได้ไปเยี่ยมชม

2. การวิเคราะห์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึก (Sentiment Analysis) [1-5] คือเป็นงานวิจัยที่อยู่ในกลุ่มของการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ที่มีกระบวนการมุ่งเน้นการวิเคราะห์และตรวจสอบความรู้สึก (Opinion) ของผู้คน จากข้อความ (Text) ที่คนเหล่านั้นเขียนหรือโพสต์เอาไว้เพื่อแบ่งบอกความรู้สึกของตนเองที่มีต่อบางสิ่งบางอย่าง เช่น ความรู้สึกดี (Positive หรือ Good) หรือความรู้สึกที่ไม่ดีหรือไม่ชอบ (Negative หรือ Bad)

ส่วนใหญ่งานวิจัยด้านการวิเคราะห์ความรู้สึกนั้น มีการประยุกต์ใช้กับข้อมูลที่เป็นการแสดงความคิดเห็นที่มีต่อสินค้าและบริการ (Product Review) [1-5] จุดประสงค์ก็เพื่อการวิเคราะห์ความรู้สึกของลูกค้าที่มีต่อสินค้าและบริการของบริษัทหรือองค์กรต่างๆ เนื่องจากการวิเคราะห์ความพึงพอใจจากข้อมูล Rating บางครั้งจะไม่สามารถระบุถึงปัญหาหรือความรู้สึกของลูกค้าที่มีต่อสินค้าและบริการที่แท้จริงได้ เช่น การที่ลูกค้าซื้อกล้องดิจิทัลไป 1 ตัว ลูกค้าอาจจะให้คะแนนเฉลี่ยเกี่ยวกับกล้องดิจิทัลยี่ห้ออื่นๆ ไว้ที่ 4 จากคะแนนเต็ม 5 แต่หัวข้อที่ต่างๆ ที่สอบถามไปยังลูกค้าอาจจะไม่ครอบคลุมในทุกๆ กรณี ที่เป็นความต้องการหรือความคาดหวังต่อสินค้าและบริการของลูกค้า ดังนั้นลูกค้าก็อาจจะไปเขียนแสดงความรู้สึกเกี่ยวกับคอมพิวเตอร์ยี่ห้ออื่นๆ ไว้ในส่วนอื่นๆ เช่น ใน Blog, Twitter หรือ Facebook ของตน [6, 7] โดยความรู้สึกเหล่านั้นอาจจะเป็นมุมมองที่เจ้าของสินค้าและบริการมองข้ามหรือคาด

ไม่ถึง ซึ่งหากนำความรู้สึกเหล่านั้นมาพิจารณา ก็อาจจะสามารถส่งผลต่อการปรับปรุงสินค้าและบริการของตนเองได้ [5]

ปัจจุบัน เทคโนโลยีด้านการวิเคราะห์ความรู้สึก เริ่มเข้ามามีบทบาทเป็นอย่างมากในหลายๆ องค์กร [1-5] ทั้งธุรกิจที่เกี่ยวข้องกับการสินค้าและบริการ การศึกษา และการให้บริการด้านการแพทย์ โดยเทคโนโลยีด้านการวิเคราะห์ความรู้สึกได้ถูกรวมเอาไว้ในระบบ commercial website หรือ Customer Relationship Management (CRM) ของแต่ละบริษัทหรือองค์กร เพื่อให้ง่ายต่อการวิเคราะห์ความรู้สึกของลูกค้าหรือผู้ใช้บริการ จนนำไปสู่การแก้ปัญหาอย่างรวดเร็ว [5]

3. กระบวนการวิจัย (Research Methodology)

ในส่วนนี้ จะแสดงภาพรวมของการดำเนินการวิจัย (ดังแสดงในภาพที่ 1) รวมถึงการแสดงรายละเอียดในแต่ละขั้นตอนของงานวิจัยที่นำเสนอ

จากภาพที่ 1 จะเห็นว่า ระบบการวิเคราะห์ความรู้สึก นักท่องเที่ยวแบบอัตโนมัติ นั้น จะแบ่งออกเป็น 2 ส่วนหลักๆ คือ (1) ส่วนของการสร้างโมเดล และ (2) ส่วนของการนำโมเดลไปใช้

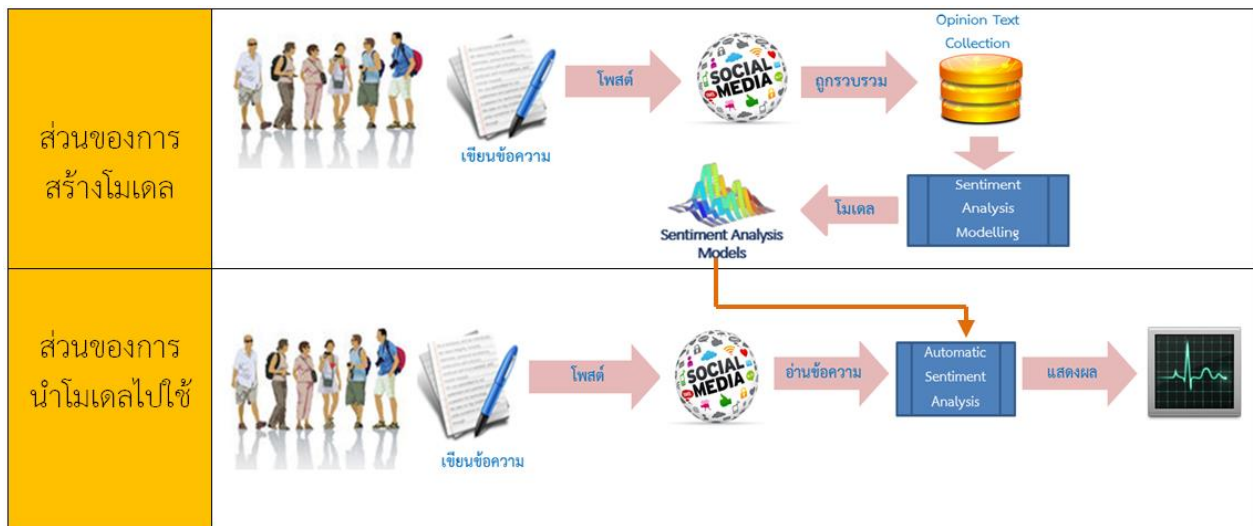
3.1 การสร้างโมเดลเพื่อการวิเคราะห์ความรู้สึก (Sentiment Classifier Modelling)

ในการสร้างโมเดลเพื่อวิเคราะห์ความรู้สึกนั้น จะประกอบด้วยขั้นตอนย่อยๆ ดังต่อไปนี้

ขั้นตอนที่ 1: การเตรียมเอกสาร (Document Pre-processing)

ในการเตรียมเอกสารจะเริ่มจาก การตัดคำ (Tokenization) [8, 9] คือ กระบวนการแยกข้อความออกเป็น “คำ” เนื่องจาก คำ จัดว่าเป็นหน่วยที่เล็กที่สุดในภาษา ที่สื่อความหมายได้

สำหรับภาษาไทย โดยทั่วไป จะใช้การตัดคำแบบพจนานุกรม (Dictionary approach) [8] ที่อยู่บนพื้นฐานของอัลกอริทึมแบบ เทียบความยาวของคำที่มากที่สุด (Longest matching Algorithm) [10] เช่น ประโยค “ฉันกินข้าว” สามารถตัดคำได้เป็น ฉัน/กิน/ข้าว



ภาพที่ 1: ภาพรวมของกระบวนการที่นำเสนอ

ต่อมาคือการตัดคำหยุด (หรือ Stop words) [8, 9] คือ กระบวนการตัดคำหรือสัญลักษณ์ที่พบบ่อยมากในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อการจัดกลุ่มเอกสาร ดังนั้นเมื่อทำการตัดออกแล้วไม่ทำให้ใจความในเอกสารนั้นๆ เปลี่ยนไป

การตัดคำหยุด มีความจำเป็นอย่างมากในงานด้านการจัดกลุ่มเอกสารแบบอัตโนมัติ เพราะจะช่วยลดระยะเวลาในการประมวลผลลงได้เป็นอย่างมาก เนื่องจากระบบฯ จะไม่เสียเวลาในการประมวลผลคำเหล่านี้ หากเป็นภาษาไทยคำหยุดอาจจะหมายถึง กลุ่มคำสันธาน คำเชื่อมประโยค คำนำหน้าชื่อบุคคล เป็นต้น

ขั้นตอนที่ 2: การนำเสนอเอกสาร (Document Representation)

ในการนำเสนอเอกสาร (Document Representation) [9] คือการแสดงความสัมพันธ์ระหว่าง “คำ” และ “เอกสาร” ที่มีอยู่ในรูปแบบของเวกเตอร์ ส่วนการให้น้ำหนักคำ [9] เป็นการกำหนดค่าน้ำหนัก “คำ” หรือ “วลี” เพื่อการเรียนรู้และทดสอบโครงข่ายประสาทเทียม โดยเป็นหลักการให้ความสำคัญกับคำหรือวลีในด้านของค่าสถิติศาสตร์ เช่น การให้ค่าสถิติความถี่ (frequency) การเกิดขึ้นของคำ จากหลักไวยากรณ์ภาษาในการ

ใช้คำหรือวลีเดิมซ้ำๆ เพื่อเป็นการเน้นย้ำถึงความสำคัญของคำหรือวลีนั้นในเอกสาร ส่วนการนำเสนอเอกสาร เป็นการจัดรูปแบบเอกสารที่แสดงความสัมพันธ์ระหว่างเอกสารและคำที่พบในเอกสารนั้น ในรูปแบบของเมตริกซ์แบบ 2 มิติที่เรียกว่า

Vector Space Model (VSM) หรือ Bag of Words (BOW) [9]

การแทนเอกสารด้วยรูปแบบ BOW เป็นการกำหนด “คำ” ในเอกสารด้วย W_j ดังนั้นเอกสารลำดับที่ j ใดๆ สามารถเขียนแทนได้ด้วย $d_j = (w_{11}, w_{12}, w_{13}, \dots, w_{1j})$ โดยสามารถแสดงได้ดังภาพที่ 2

จากนั้นจะทำการให้น้ำหนัก “คำ” ที่เป็นตัวแทนเอกสาร เพื่อแสดงความสอดคล้องระหว่าง “คำ” นั้นๆ และเอกสาร ในงานวิจัยฉบับนี้ จะใช้วิธีการให้น้ำหนักที่เรียกว่า *tf-idf* (Term Frequency – Inverted Document Frequency) [9, 11]

$$\text{Document } (d_j) = \begin{bmatrix} w_{11} & \dots & w_{1j} \\ \dots & \dots & \dots \\ w_{i1} & \dots & w_{ij} \end{bmatrix}$$

ภาพที่ 2: แสดง BOW

tf-idf เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์เพื่อใช้ในการจัดกลุ่มของเอกสารให้ตรงกับหมวดหมู่ที่กำหนดไว้ โดย *tf* เป็นการหาความถี่ของคำที่ปรากฏในเอกสาร และ *idf* เป็นการหาส่วนกลับของเอกสารหรือที่เรียกว่าระบบน้ำหนักความถี่เอกสารผกผัน โดยสามารถหาได้จากสมการ

$$idf = 1 + \log(N/df) \quad (1)$$

โดยที่ N คือจำนวนเอกสารทั้งหมดในกลุ่ม ขณะที่ df คือจำนวนเอกสารที่มีคำๆ นั้นปรากฏอยู่ ดังนั้น

$$tf-idf = tf \times idf \quad (2)$$

สมการ *tf-idf* เป็นวิธีการหาตัวแทนเวกเตอร์เพื่อนำไปค้นคืนสารสนเทศที่เป็นกลุ่มของเอกสาร ซึ่งวิธีนี้เป็นกรให้น้ำหนักอย่างง่ายแต่ก็ได้รับการยอมรับว่ามีประสิทธิภาพที่น่าพอใจกับการจัดกลุ่มเอกสาร [11]

ตัวอย่างเอกสารที่ผ่านการให้น้ำหนักคำในแต่ละเอกสารสามารถแสดงได้ดังตารางที่ 1

ตารางที่ 1: ตัวอย่างเอกสารที่ผ่านการให้น้ำหนักคำ

คำ	D1	D2	D3	D4
บรรยากาษ	0.602	0	0	0
หนองบัวแดง	0.602	0	0	0
ผาแต้ม	0	0.301	0	0
สวยมก	0.301	0.301	0	0.301
สามพันโบก	0	0	0.602	0
ชยะ	0	0	0.602	0
เยอะมก	0	0	0.602	0
สกปรก	0	0	0.301	0.301

ขั้นตอนที่ 3: การสร้างโมเดลเพื่อการวิเคราะห์ความรู้สึก

ด้วย Naïve Bayes

นาอ์ฟเบย์ (Naïve Bayes) [12] เป็นอัลกอริทึมที่ถูกนำมาใช้อย่างแพร่หลายในงานจำแนก เอกสาร และให้ผลที่ดี สร้างตัวจำแนกแบบอัตโนมัติด้วยนาอ์ฟเบย์

เริ่มจากแต่ละ อินสแตนซ์ (Instance) x ซึ่งจัดอยู่ในรูปเวกเตอร์ โดยเวกเตอร์นั้นจะแสดงคุณลักษณะที่ถูกคัดเลือกเช่น $\langle a_1, a_2 \dots a_n \rangle$ โดยที่ค่าเป้าหมายที่ต้องการของแต่ละอินสแตนซ์ จะเป็นค่าใดๆ ภายในเซต V เมื่อ V มีสมาชิกเป็นค่าเป้าหมายที่ต้องการ ในที่นี้หมายถึงจำนวนกลุ่มของข้อมูลนั่นเอง

นาอ์ฟเบย์ประยุกต์ใช้งานในการจำแนกประเภทของข้อความ (Text Classification) และพบว่าสามารถใช้งานได้ไม่ต่างจากวิธีการจำแนกวิธีอื่นๆ เป็นเหตุให้ผู้วิจัยเลือกใช้วิธีนี้ในงานวิจัยนี้ เนื่องจากให้ประสิทธิภาพการทำงานที่ดี และวิธีการทำงานไม่ซับซ้อนเหมือนวิธีการอื่นๆ

การกำหนดความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม V_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{a_1, a_2, \dots, a_n\}$ หรือใช้สัญลักษณ์ว่า $P(a_1, a_2, \dots, a_n)$ คือ

$$P(a_1, a_2, \dots, a_n | v_j) = \prod_{i=1}^n P(a_i | v_j) \quad (3)$$

โดยที่ \prod หมายถึงผลคูณของค่า $P(a_i | v_j)$ เมื่อ i และ j มีค่าเท่ากับ $1, 2, 3, \dots, n$

วิธีการเรียนรู้เบย์อย่างง่ายไปใช้วิธีดังต่อไปนี้คือ

- (1) หาค่าความน่าจะเป็นของค่าที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, \dots, a_n | v_j)$ จากสมการมาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ $P(v_j)$ ได้เท่ากับ V_{NB}
- (2) นำค่าที่ได้มาเปรียบเทียบกับกลุ่มที่มีความน่าจะเป็นสูงสุดคือกลุ่มที่ข้อมูลนั้นอยู่ และจะถูกจัดเข้าไปเขียนเป็นสมการได้คือ

$$V_{NB} = \arg \max P(v_j) \times \prod_{i=1}^n P(a_i | v_j) \quad (4)$$

บางส่วนของตัวอย่างโมเดลที่ใช้ในการการวิเคราะห์ความรู้สึกจากข้อความแบบ 2 กลุ่ม คือความรู้สึกที่ดี (Positive) และความรู้สึกที่ไม่ดี (Negative) สมมติว่าสามารถแสดงได้ดังตารางที่ 2

ตารางที่ 2: ตัวอย่างโมเดลที่ได้จาก Naïve Bayes

Class = "Positive"	
$P(\text{Class} = \text{Positive})$	0.5
$P(\text{บรรยากาษ}) \times \text{tf-idf}_{\text{บรรยากาษ}}$	0.602
$P(\text{หนองบัวแดง}) \times \text{tf-idf}_{\text{หนองบัวแดง}}$	0.602
$P(\text{ผาแต้ม}) \times \text{tf-idf}_{\text{ผาแต้ม}}$	0.1505
$P(\text{สวยมก}) \times \text{tf-idf}_{\text{สวยมก}}$	0.602
Class = "Negative"	
$P(\text{Class} = \text{Negative})$	0.5
$P(\text{ผาแต้ม}) \times \text{tf-idf}_{\text{ผาแต้ม}}$	0.1505
$P(\text{สามพันโบก}) \times \text{tf-idf}_{\text{สามพันโบก}}$	0.1505
$P(\text{ชยะ}) \times \text{tf-idf}_{\text{ชยะ}}$	0.1505
$P(\text{เยอะมก}) \times \text{tf-idf}_{\text{เยอะมก}}$	0.602
$P(\text{สกปรก}) \times \text{tf-idf}_{\text{สกปรก}}$	0.602

3.2 การนำโมเดลเพื่อการวิเคราะห์ความรู้สึกไปใช้งาน (Sentiment Model Usage)

เป็นขั้นตอนของการนำเอาโมเดลเพื่อการวิเคราะห์ความรู้สึกจากข้อความมาใช้งาน โดยเมื่อมีการอ่าน "ข้อความ" เข้ามา โมเดลดังกล่าวจะวิเคราะห์ให้ว่า ความรู้สึกของ

นักท่องเที่ยวยังสถานที่ท่องเที่ยวอื่นๆ เป็น Positive หรือ Negative

สมมติให้มีชุดข้อความที่นักท่องเที่ยวแสดงความคิดเห็น มีข้อความดังนี้

D_{NEW} : ผมคิดว่าผาแต้มสวยมากนะ

ขั้นตอนการจัดกลุ่มข้อความเริ่มจาก การตัดคำและตัดคำหยุดในข้อความ เมื่อผ่านการตัดคำและตัดคำหยุดในข้อความแล้ว จะได้คำสำคัญดังนี้

ตารางที่ 3: แสดงการหาคำสำคัญจากข้อความ

Document	ข้อความภายหลังการตัดคำ	ภายหลังการตัดคำหยุด
ข้อมูลใหม่	ผม/ คิดว่า/ ผาแต้ม/ สวยมาก/ นะ	ผาแต้ม / สวยมาก

เมื่อได้ ‘คำสำคัญ’ จากข้อความมาแล้ว จากนั้นจะใช้ข้อมูลจากตารางที่ 1 และ 2 ในการวิเคราะห์เพื่อพิจารณาการจัดกลุ่มเอกสารภายใต้สมการที่ (4)

จากสมการที่ (4) สามารถคำนวณหาค่าความน่าจะเป็นเพื่อประเมิน Class ที่ละ Class ตามลำดับ ในที่นี้ $P(\text{Class}) = 0.5$ ทั้ง Class ที่เป็น Positive และ Negative จะได้ว่า

ตรวจสอบใน Class = “Positive”

$$\begin{aligned} V_{NEW} &= P(+) \times P(\text{สวยมาก}|+) \times P(\text{ผาแต้ม}|+) \\ &= (0.5) \times (0.602) \times (0.1505) = 0.0453005 \end{aligned}$$

ตรวจสอบใน Class = “Negative”

$$\begin{aligned} V_{NEW} &= P(-) \times P(\text{สวยมาก}|+) \times P(\text{ผาแต้ม}|+) \\ &= (0.5) \times (0) \times (0.1505) = 0 \end{aligned}$$

จากผลลัพธ์ข้างต้นเห็นได้ชัดเจนว่า เอกสาร D_{NEW} นั้นมีความน่าจะเป็นอยู่ที่ Class = “Positive” มากกว่า Class = “Negative” ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ Positive หรือความรู้สึกที่ดีนั่นเอง

4. ผลการทดสอบโมเดล (Model Evaluation)

ในส่วนนี้จะกล่าวถึงการนำตัวจัดกลุ่มเอกสารที่ได้จากขั้นตอนการดำเนินงาน มาทำการทดลองจัดกลุ่มข้อความแสดงความรู้สึกของนักท่องเที่ยว โดยนักวิจัยได้พัฒนาโปรแกรมด้วยภาษา Java

4.1 ชุดข้อมูล (Dataset)

ข้อมูลเอกสารที่ใช้ในงานวิจัย ฉบับนี้ รวบรวมมาจาก Social Media เช่น Facebook จำนวน 480 เอกสาร โดยจัดเก็บในรูปแบบ text file

4.2 ผลการทดสอบ (Experimental Results)

การประเมินประสิทธิภาพของการค้นคืนสารสนเทศ ได้ทำการทดสอบเพื่อวัดประสิทธิภาพ ด้วยค่าความแม่นยำ (Recall: R) [9] ค่าการค้นคืน (Precision: P) [9] และค่าการวัดประสิทธิภาพพื้นฐานในการจัดกลุ่มเอกสาร F -measure [9] เพื่อประเมินค่าความแม่นยำในการค้นคืนข้อมูล

เมื่อทำการทดสอบ โมเดลการจัดกลุ่มเอกสาร เพื่อประยุกต์ใช้ในการวิเคราะห์ความรู้สึก ด้วยเอกสารทดสอบที่แตกต่างเอกสารที่ใช้ในการสร้างโมเดล จำนวน 120 เอกสาร ด้วยค่าความระลึก ค่าความแม่นยำ และค่าทดสอบเอฟ โดยพิจารณาจากจำนวนเอกสาร ที่ละกลุ่มด้วยจำนวนที่แตกต่างกัน คือ 60, 80, และ 100 ตามลำดับ สามารถแสดงผลของการทดสอบได้ในตารางที่ 3

จากการทดสอบโมเดลการจัดกลุ่มเอกสารที่ประยุกต์ใช้ในการวิเคราะห์ความรู้สึกที่แสดงในตารางที่ 4 จะเห็นได้ว่า ให้ผลการทดสอบที่น่าพึงพอใจ เพราะมีเมื่อทดสอบด้วยค่าทดสอบเอฟ มีค่าเฉลี่ยที่ 0.883

ตารางที่ 4: ผลการทดสอบโมเดลการวิเคราะห์ความรู้สึก

โมเดล	R	P	F-measure
โมเดลที่สร้างจาก 30 เอกสาร/กลุ่ม	0.86	0.88	0.869
โมเดลที่สร้างจาก 40 เอกสาร/กลุ่ม	0.87	0.89	0.879
โมเดลที่สร้างจาก 50 เอกสาร/กลุ่ม	0.89	0.91	0.899
ค่าเฉลี่ย	0.873	0.89	0.883

อย่างไรก็ตาม แม้ว่าโมเดลการจัดกลุ่มเอกสารที่ประยุกต์ใช้ในการวิเคราะห์ความรู้สึกจะให้ผลลัพธ์ที่ดี แต่ก็ยังมีข้อผิดพลาด เนื่องจากการสร้างโมเดลการจัดกลุ่มเอกสารจะอาศัย “คำ” เป็นคุณลักษณะสำคัญ โดยการดำเนินการเพื่อให้ได้ “คำ” มานั้น จะอาศัยพจนานุกรมเป็นหลัก

เนื่องจากการรวบรวมเอกสารจาก Social Media อาจจะมีการใช้ภาษาที่ผิดเพี้ยนไปจากปกติ ดังนั้น “คำ” บางคำที่พบในเอกสาร แต่ไม่มีในพจนานุกรม แม้ว่าจะมีความสำคัญ แต่ก็

อาจจะไม่ถูกคัดเลือกมาเป็น “คำ” ที่ใช้คุณลักษณะสำคัญในการสร้างโมเดล และเมื่อนำโมเดลดังกล่าวไปใช้ในการทดสอบ หากเอกสารที่ใช้ทดสอบมี “คำ” ดังกล่าว ก็ไม่อาจจะวิเคราะห์ความรู้สึกได้

5. สรุปงานวิจัย (Conclusion)

งานวิจัยฉบับนี้ เป็นการวิเคราะห์ข้อความที่แสดงความรู้สึกแบบอัตโนมัติของนักท่องเที่ยวที่มีต่อสถานที่ท่องเที่ยวในประเทศไทย ที่มีการแสดงออกไว้ในข้อความแสดงความรู้สึกผ่าน Social Media ต่างๆ โดยในงานวิจัยนี้จะวิเคราะห์ความรู้สึกแบบ 2 กลุ่ม คือ ความรู้สึกที่ดีต่อสถานที่ท่องเที่ยว (Positive) และความรู้สึกที่ไม่ดีต่อสถานที่ท่องเที่ยว (Negative) ซึ่งอัลกอริทึม Naïve Bayes ถูกใช้เป็นอัลกอริทึมหลักในการสร้างโมเดลสำหรับวิเคราะห์ความรู้สึก

สำหรับขั้นตอนการวัดประสิทธิภาพของการจัดกลุ่มข้อความความรู้สึกของนักท่องเที่ยว ในงานวิจัยฉบับนี้ใช้ค่าความแม่นยำ ค่าความระลึก และค่า F -measure เมื่อทำการทดสอบวัดประสิทธิภาพของการสร้างโมเดลในการค้นคืนสารสนเทศเชิงความหมาย พบว่าค่าความแม่นยำ ค่าความระลึก และค่า F -measure ได้เท่ากับ 0.883 แม้จากผลการทดสอบจะให้ผลลัพธ์ที่น่าพอใจ แต่ก็ยังพบความผิดพลาด อันเนื่องมาจากการรวบรวมเอกสารจาก Social Media นั้น มีเอกสารหลายฉบับที่ใช้ภาษาที่ผิดเพี้ยนไปจากปกติ เช่น คำแสดง คำอุทาน แปลกๆ

เมื่องานวิจัยนี้ใช้การตัดคำแบบพจนานุกรมเพื่อกำหนด “คำ” ที่จะใช้เป็นคุณลักษณะสำหรับการสร้างโมเดล ดังนั้น “คำ” บางคำที่พบในเอกสาร แต่ไม่มีในพจนานุกรม แม้อาจมีความสำคัญ แต่ก็อาจจะไม่ถูกคัดเลือกมาเป็น “คำ” ที่ใช้คุณลักษณะสำคัญในการสร้างโมเดลส่งผลให้โมเดลที่สร้างมาเพื่อวิเคราะห์ความรู้สึกมีความผิดพลาด ส่งผลทำให้ความแม่นยำของการวิเคราะห์ลดต่ำลงได้

เอกสารอ้างอิง

[1] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis”, Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) PP: 1–135.

[2] J.S. Modha, H. G. S. Pandi, S.J. Modha, “Automatic Sentiment Analysis for Unstructured Data”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, 2013.

[3] S. Modak, A.C. Mondal, “A Study on Sentiment Analysis”, International Journal of Advanced Research in Computer Science & Technology, Vol. 2, Issue 2, 2014.

[4] H. Gomes, M. de Castro Neto, R. Henriques, R., “Text Mining: Sentiment analysis on news classification”, The 8th Iberian Conference on Information Systems and Technologies (CISTI), 2013.

[5] J. Polpinij and A.K. Ghose, “An Ontology-Based Sentiment Classification Methodology for Online Consumer Reviews”, Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 518 - 524, 2008.

[6] K. A.C.E.S Lima and L.N. de Castro, “Automatic sentiment analysis of Twitter messages”, Fourth International Conference on Computational Aspects of Social Networks, 2012.

[7] P L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, “Target-dependent Twitter sentiment classification”, Proceeding HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[8] R. Dale, H. Moisl, H. Somers (Editor), Handbook of Natural Language Processing, Marcel Dekker, Inc. NY., 2000.

[9] R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1997.

[10] S. Meknavin, P. Charoenpornasawat, and B. Kijisirikul, “Feature-Based Thai Word Segmentation” Proc. of NLPRS 97, pp. 289–296, 1997.

[11] N. Pascal Soucy, Guy W. Mineau, “Beyond TFIDF Weighting for Text Categorization in the Vector Space Model”, IJCAI'05 Proceedings of the 19th international joint conference on Artificial intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA 2005.

[12] E. Frank and R.R. Bouckaert, “Naive bayes for text classification with unbalanced classes”, Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases, 2006.