

โมเดลสำหรับจำแนกความรู้สึกของความคิดเห็นโดยใช้เทคนิคต้นไม้ตัดสินใจ กรณีศึกษาเว็บไซต์จองโรงแรม

Model for the Classification of Feelings of Reviews Using Techniques Decision Trees Case Studies Hotel Reservation Web Site

มุกตา ฆมานเหม^{1*} สิทธิพงศ์ ต่อ๊ะ¹ สารภี จุลแก้ว² และสุภาวดี มากอัน³
Mukta manhem^{1*}, Sittipong Dolah¹, Sarapee Chunkaew²
and Supawadee Mak-on³

บทคัดย่อ

การจำแนกความรู้สึกของการวิจารณ์เป็นการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล ซึ่งมีหลากหลายธุรกิจนำไปใช้งานอย่างกว้างขวาง โดยงานวิจัยนี้เลือกศึกษากลุ่มการท่องเที่ยวที่เจาะจงกลุ่มโรงแรมที่พักจากเว็บไซต์ Booking.com และ Traveloka.com ที่มีการแยกความรู้สึกในเบื้องต้นเพื่อใช้แทนผู้เชี่ยวชาญในการสร้างป้ายกำกับข้อมูลร่วมกับผู้วิจัยตรวจสอบ และจำแนกความรู้สึกด้วยต้นไม้ตัดสินใจ โดยเปรียบเทียบเครื่องมือตัดคำ ได้แก่ MM NEWMM และ LONGEST ร่วมกับชุดคัดแยกคำ Stopword ของ pythainlp และชุดคำของผู้วิจัย ให้ผลประสิทธิภาพค่าความถูกต้องสูงสุดในเครื่องมือตัดคำ NEWMM ที่ค่าความถูกต้อง 0.90 และค่าความแม่นยำร่วมกับชุดคัดแยกคำของผู้วิจัยที่ 0.90 ทั้งความรู้สึกชอบ และไม่ชอบ ชุดคัดแยกคำ Stopword ให้ค่าความถูกต้องเพียง 0.81 และค่าความแม่นยำที่ 0.81 เช่นกัน ดังนั้นแบบสำหรับจำแนกความรู้สึกการวิจารณ์ที่พักโรงแรมด้วยเทคนิคต้นไม้ตัดสินใจ มีประสิทธิภาพดีในชุดตัดคำ NEWMM ร่วมกับชุดคัดแยกคำของผู้วิจัย

คำสำคัญ : วิจารณ์โรงแรมที่พัก การจำแนกความรู้สึก ต้นไม้ตัดสินใจ การทำเหมืองข้อมูล

¹ หลักสูตรเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา

² หลักสูตรวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา

³ สาขาวิชาเทคโนโลยีคอมพิวเตอร์ วิทยาลัยรัตนภูมิ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

* Corresponding author, E-mail: Muktamanhem051@gmail.com

Abstract

Sentiment analysis with text data mining is widely used for business success. The researcher's study is to identify the customer review sentiment from Booking.com and Traveloka.com websites that both websites have indicated the customer reviews and we cross-check in the prepare data process. We used more tokenize engine including included MM, NEWMM and LONGEST with those of cut word suite pythainlp's stopword and our custom stopword. The results show that the NEWMM engine has accuracy at 0.90 and 0.90 for precision both positive and negative reviews. However, the NEWMM engine with pythainlp's stopword has accuracy at only 0.81 and precision at only 0.81. Concluding, our sentiment model with the decision tree technique can found better performance with the NEWMM engine and research's custom stopword.

Keywords: hotel reviews, sentiment analysis, decision tree, data mining

บทนำ

ในปัจจุบันการท่องเที่ยวถือว่าเป็นอุตสาหกรรมสำคัญหนึ่งที่สร้างรายได้มหาศาลให้กับประเทศไทย ซึ่งเป็นไปตามแผนพัฒนาการท่องเที่ยวแห่งชาติ ฉบับที่ 2 (พ.ศ. 2560 – 2564) จึงจะทำให้การท่องเที่ยวพัฒนามากขึ้น ซึ่งนักท่องเที่ยวส่วนใหญ่ทั้งที่เป็นคนไทยและเป็นชาวต่างชาติย่อมมีความต้องการ การบริการที่ดีและคุณภาพของที่พักที่คุ้มค่ากับเงินที่จ่ายไป ช่องทางหนึ่งที่นักท่องเที่ยวจะทราบถึงข้อมูลของที่พักเหล่านั้นโดยการหาข้อมูลในเว็บไซต์ทางสังคมที่ให้บริการจองโรงแรม ซึ่งเว็บไซต์จองโรงแรมส่วนใหญ่นำเสนอ ความคิดเห็น จากนักท่องเที่ยวที่เคยใช้บริการโรงแรมนั้น ๆ ในรูปแบบข้อความหากโรงแรมได้มีผู้แสดงความคิดเห็นมากจะปรากฏข้อความหลายร้อยข้อความ ทำให้ผู้ใช้ต้องเสียเวลาอ่านข้อคิดเห็นทั้งหมด แล้วทำการสรุปข้อดีข้อเสียในด้านต่าง ๆ สำหรับโรงแรมที่ตนสนใจ และความคิดเห็นในบางครั้งอาจไม่ตรงกับช่องความรู้สึกของหน้าเว็บกำหนด จึงอาจเสนอตลาดเคลื่อน โดยในบางเว็บไซต์ของโรงแรมรวมความคิดเห็นที่ไม่ถูกจำแนกไว้ความรู้สึกไว้ ทำให้ผู้ประกอบการโรงแรมต้องมีการนำไปสรุปจึงจะได้ผลข้อแสดงความคิดเห็น

จากปัญหาและที่มาข้างต้นผู้วิจัยได้นำเสนอโมเดลจำแนกความรู้สึกของความคิดเห็นของนักท่องเที่ยวบนเว็บไซต์จองโรงแรม จำนวน 2,000 ชุด แยกข้อความรู้สึกชอบ และไม่ชอบอย่างละเท่ากัน โดยนำข้อความคิดเห็นที่อยู่ในช่วงของการท่องเที่ยวเทศกาลเดือนพฤศจิกายน 2562 ในจังหวัดภูเก็ต เพื่อสร้างโมเดลจำแนกความรู้สึกของข้อความด้วยเทคนิคต้นไม้ตัดสินใจ เพื่อหาความหมายเชิงบวกหรือเชิงลบของประโยคว่าไปทางด้านทิศทาง สำหรับผู้ใช้ประกอบการตัดสินใจนอกจากนี้ผู้พัฒนานำไปสู่การพัฒนาชุดแปลผลความรู้สึกออนไลน์ให้กับ

วัตถุประสงค์การวิจัย

เพื่อพัฒนาโมเดลในการจำแนกความรู้สึกของความคิดเห็นภาษาไทยด้วยภาษา Python และเครื่องมือ Pythainlp

ขอบเขตของการศึกษาวิจัย

1. การทำเหมืองข้อมูลด้วยการจำแนกแบบต้นไม้การตัดสินใจ

โมเดลต้นไม้การตัดสินใจทำการแบ่งข้อมูลตามเงื่อนไขซึ่งระบุในโหนดต่าง ๆ ของโครงสร้างต้นไม้และจะสิ้นสุดการแบ่งข้อมูลที่โหนดใบ เพื่อที่จะให้คำตอบ ตามเงื่อนไขที่เข้ามาเป็นข้อมูล ที่ถือว่าการตัดสินใจสามารถแบ่งผลลัพธ์ ตามลำดับเงื่อนไขเป็นปัจจัยที่ส่งผลต่อการจำแนก ที่กำหนดค่า Entropy และคำนวณหา Information gain เพื่อเป็นการวัดว่าหากจำแนกข้อมูลลักษณะค่าที่กำหนดได้

กรณีตัวอย่าง ผลลัพธ์ที่ได้จากการจำแนกข้อมูล ควรได้ผลลัพธ์ ที่เป็นเอกลักษณ์ เช่นถ้าใช้การจำแนก ทั้งหมด 7 ตัวในการจำแนกประเภทข้อมูล ผลลัพธ์จากการจำแนกจะต้องตอบเหมือนกันอย่างน้อย 5 ตัวหรือ 70% ขึ้นไป จึงสามารถเลือกผลลัพธ์นั้นเพื่อเป็นคำตอบได้

2. งานวิจัยที่เกี่ยวข้อง

จิราภา เลหาหะวรนนท์ และคณะ (2558) ด้วยคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง มีหลักสูตรการเรียนตามแขนงวิชาโดยมีการแบ่งเป็น 4 แขนงวิชา (วิศวกรรมซอฟต์แวร์, เทคโนโลยีเครือข่ายและระบบ, การพัฒนาสื่อประสมและเกม และ อัจฉริยะทางธุรกิจ) นักศึกษาส่วนใหญ่ไม่ทราบความถนัดของตนเอง จึงเป็นที่มาของการพัฒนา “ระบบแนะนำแขนงวิชา” เพื่อเป็นเครื่องมือช่วยตัดสินใจในการเลือกแขนงวิชา โดยงานวิจัยนี้ใช้ปัจจัยผลการเรียนและผลการวัดความสามารถด้านต่าง ๆ นำมาสร้างแบบจำลองพยากรณ์ด้วยเทคนิคการทำเหมืองข้อมูล 5 เทคนิค พบว่าการพยากรณ์มีความแม่นยำอยู่ที่ 72.92% มีความแม่นยำถึง 86.67% ในแขนงวิศวกรรมซอฟต์แวร์

อุษนิษา เถาว์โท และคณะ (2561) ศึกษาการจำแนกคำวิจารณ์ เพื่อแนะนำภาพยนตร์ ด้วยปัจจัย 4 ประเด็นได้แก่ นักแสดง เสียง ภาพหรือกราฟิก และ ภาพรวม ให้ผลการสร้างโมเดลต้นไม้ตัดสินใจ เช่นเดียวกับในงานวิจัยนี้ โดยกลุ่มตัวอย่างทำการสุ่มเลือกข้อความรายประโยค ซึ่งผสมระหว่างประโยคความเดียว ประโยคความซ้อนและประโยคความรวม จำนวน 400 ข้อความ แต่การพยากรณ์ในงานวิจัยนี้เน้นที่การแยกประเภทการวิจารณ์ ซึ่งให้ผลการจำแนกอธิบายในรูปแบบของ Confusion matrix ให้ค่าถูกต้องสูงสุดสำหรับการวิจารณ์เรื่องเสียง ในแง่บวก

นลินี พุกษาชาติ และ ธนพล เจนสุทธิเวชกุล (2560) การพัฒนาอย่างรวดเร็วของเทคโนโลยีของเครือข่ายสังคมออนไลน์ด้วยแพลตฟอร์มที่ได้รับความนิยมสำหรับบุคคลทั่วไป ด้วยอุตสาหกรรมโรงแรมและการแสดงความคิดเห็นของนักท่องเที่ยวถือเป็นข้อมูลสำคัญ สำหรับผู้ประกอบการ ที่ต้องนำมาวิเคราะห์เพื่อปรับปรุงการบริการและคุณภาพของโรงแรม โดยงานวิจัยฉบับนี้นำเสนอกระบวนการจำแนกความคิดเห็นออนไลน์โดยใช้เทคนิค Word2vec ร่วมกับขั้นตอนวิธีเอนอีฟเบย์เพื่อจำแนกความคิดเห็น ซึ่งผู้วิจัยได้รวบรวมข้อมูลความคิดเห็นภาษาอังกฤษจำนวน 90,1906 ความคิดเห็นจากเว็บไซต์ agoda.com และ Tripadvisor.com เพื่อวิเคราะห์คุณสมบัติ 10 ด้าน ได้แก่ ที่ตั้ง (Location) การบริการ (service) อาหาร (Food) ห้องพัก (Room) ความสะอาด (Clean) ราคา (Price) สิ่งอำนวยความสะดวก (Facility) การเดินทางสถานที่ที่เี่ยวรอบโรงแรม (Attraction) และความปลอดภัย (Location) โดยพบว่าสถานที่ตั้งของโรงแรมมีผลต่อการตัดสินใจเลือกโรงแรมของนักท่องเที่ยวมากที่สุด

วิธีการดำเนินการวิจัย

1. การเตรียมข้อมูล การเก็บรวบรวมข้อมูลคำรีวิวโรงแรมต่าง ๆ จากเว็บไซต์จองโรงแรมโดยใช้เกณฑ์การคัดเลือกจากโรงแรมที่กำลังเป็นที่ต้องการมากที่สุดในช่วงวันที่ 21-25 กุมภาพันธ์ 2562 โดยเก็บข้อมูลจากโรงแรมทั้งหมด 40 โรงแรมจากโรงแรมทั้งหมดที่ตั้งอยู่ในจังหวัดภูเก็ต ซึ่งข้อมูลทั้งหมดที่เก็บมามีจำนวน 2,000 ประโยค โดยจะเก็บเฉลี่ยตามคำรีวิวในแต่ละโรงแรมว่ามีมากหรือน้อย

ตัวอย่างภาพประกอบ

★ **รีวิวยอดเยี่ยม** รีวิวเมื่อ: 19 พฤศจิกายน 2019

แสนวิเศษ

😊 · พนักงานต้อนรับดีมาก ๆ ยิ้มแย้มแจ่มใส ให้คำแนะนำดีมาก ๆ ที่พักสะอาด ที่จอดรถสะดวก นื่องๆ พนักงานน่ารักทุกคนเลยคะ

😞 · อาหารเข้าน้อยไปนิดคะ เปรียบเทียบกับราคา

ภาพที่ 1 ตัวอย่างคำรีวิวโรงแรมในเว็บไซต์

2. การตัดภาษาไทยด้วย pythainlp

ในโปรแกรม Python มีชุดคำสั่งสำหรับการจัดคำภาษาไทยที่ได้รับการพัฒนาที่มีชื่อว่า Thai Natural Language Processing in Python (PyThaiNLP) และมีชุดเครื่องมือตัดคำที่สำคัญดังนี้

- การตัดคำแบบ mm Dictionary-based Thai Word Segmentation using maximal matching algorithm and Thai Character Cluster (TCC).
- การตัดคำแบบ newmm Dictionary-based Thai Word Segmentation using maximal matching algorithm and Thai Character Cluster (TCC). The code is based on the notebooks created by Korakot Chaovavanich.
- การตัดคำแบบ longest Dictionary-based longest-matching Thai word segmentation. Implementation based on the code from Patorn Utenpattanan.

ซึ่งชุดตัดคำที่เมื่อนอกเหนือจากนี้ ต้องทำการติดตั้งเพิ่มเติม สำหรับงานวิจัยนี้เลือกพิจารณาชุดตัดคำพื้นฐานสำหรับนำมาเปรียบเทียบประสิทธิภาพในการจำแนกความรู้สึก โดยอธิบายการพัฒนาและประสิทธิภาพการจำแนกความรู้สึกร่วมกับเทคนิคต้นไม้ตัดสินใจในหัวข้อถัดไป ข

3. การพัฒนาโปรแกรม

การโปรแกรมสำหรับงานวิจัยนี้ ใช้ภาษาหลักคือ Python ในรุ่น 3.7.4 เพื่อรองรับ pythainlp 1.5 ในการประมวลผลและการตัดคำ ร่วมกับ Natural Language Toolkit (NLTK) และ scikit-learn เพื่อใช้เครื่องมือการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งมีขั้นตอนในการออกแบบพัฒนาโปรแกรม

ตารางที่ 1 ชุดคำสั่งภาษา Python สำหรับสร้างและวัดประสิทธิภาพ Model การจำแนกข้อมูล

ชุดคำสั่งการสร้างและวัดประสิทธิภาพ Model การจำแนกข้อมูล

```
# การอ่านคำวิจารณ์สู่ชุดข้อมูล
data_pos = [(line.strip(), 'pos') for line in codecs.open("pos-hotel.txt", 'r', 'utf-8')]
data_neg = [(line.strip(), 'neg') for line in codecs.open("neg-hotel.txt", 'r', 'utf-8')]
features_data = [(split_words(sentence), sentiment) for (sentence, sentiment) in
data_pos + data_neg]
# คำนวณหา Feature และการแปลงข้อมูลสู่ตาราง และ Label ค่าความรู้สึก
x_word = get_words_in_sentences(features_data)
word_features = get_word_features(x_word)
x_features = apply_features(extract_features, features_data)
x_data = [x_features[i][0] for i in range(len(x_features))]
y_data = [x_features[i](Thinkaboutwealth, 2563) for i in range(len(x_features))]
```

ข้อมูลที่ถูกสร้างขึ้นก่อนนำไปสร้างต้นแบบต้องทำการแปลงให้อยู่ในค่า “0” หรือ “1” เพื่อให้ตรงกับค่าของคำคุณลักษณะที่กำหนดไว้ ซึ่งได้จำนวนมากกว่า 2,000 คำ ตามค่า Feature ที่ปรากฏในขั้นตอนผลการวิจัย

4. การทดสอบ K-fold (10 -fold)

เทคนิค “K-Fold Cross Validation” ถูกระบุให้เป็น Golden Standard สำหรับการสร้างและประเมินประสิทธิภาพ หรือเรียกว่า k-fold ด้วยหลักการเทคนิคการทำ Resampling คือการแบ่งข้อมูลเป็น k ส่วนเพื่อสร้างและประเมินประสิทธิภาพโมเดล (train + test) คำนวณค่าเฉลี่ยค่าความถูกต้อง (accuracy) แบ่งข้อมูลเป็น 5 folds โดยการแบ่งข้อมูลต้องเป็นไปอย่าง random ที่มีการกระจายตัวใกล้เคียงกันช่วยลดความ bias ในการวิจัยส่วนใหญ่นิยมใช้กันในทางปฏิบัติมีสองค่าคือ k=5 หรือ k=10

ตารางที่ 2 ชุดคำสั่งภาษา Python เพื่อวัดประสิทธิภาพ Model การจำแนกข้อมูล แบบ 10-fold validation

ชุดคำสั่งการประมวลผลประสิทธิภาพแบบ 10-fold validation

```
# การทำ 10-fold validation
k_fold = KFold(n_splits=10, random_state=1992, shuffle=True)
word_features = None
accuracy_scores = [ ]
for train_set, test_set in k_fold.split(features_data):
    x_train = [x_data_feat[i] for i in train_set]
    y_train = [y_data[i] for i in train_set]
    x_test = [x_data_feat[i] for i in test_set]
    y_test = [y_data[i] for i in test_set]
    classifier = DecisionTreeClassifier(criterion='entropy', random_state=0)
    classifier.fit(x_train, y_train)
    y_pred = classifier.predict(x_test)
```

ผลการวิจัย

การทดสอบเพื่อเปรียบเทียบข้อมูลด้วยเครื่องมือตัดคำภาษาไทยของฟังก์ชัน pythainlp ได้แก่ MM NEWMM และ LONGEST ที่มีการกำหนดคุณลักษณะคำ (features) พร้อมชุดตัดคำ ไว้ 3 แบบ ได้แก่ ตัดคำด้วยชุดคำสร้างเอง ตัดคำด้วย Stop word ของ pythainlp เทียบกับการไม่ตัดคำ เพื่อเสนอผลลัพธ์ของค่าความถูกต้อง (accuracy) และจำนวนคุณลักษณะคำ ของแต่ละโมเดล

1. เครื่องมือตัดคำ MM

เครื่องมือชุดที่ 1 ที่ได้รับการพัฒนาในรุ่นเริ่มต้น โดยมีชุดพจนานุกรมสำหรับเทียบคำไทยในการตัดคำ ซึ่งมีประสิทธิภาพในการจำแนกความรู้สึกได้ถูกต้องแยกตามกรณีที่มีการตัดแยกคำในกลุ่ม Stop word พร้อมจำนวนคำคุณลักษณะหลังได้รับการตัดคำตาม ตารางที่ 3

ตารางที่ 3 ค่าความถูกต้อง และจำนวนคำคุณลักษณะด้วยเครื่องมือ MM

เครื่องมือตัดคำ	ค่าความถูกต้อง (accuracy)	คำคุณลักษณะ(features)
cutword	0.85	2382
stopword	0.82	2174
not stopword	0.84	2448

ตามตารางที่ 3 เครื่องมือตัดคำแบบไม่ตัดแยกคำได้ค่าความถูกต้องที่ 0.84 ซึ่งหากใช้ชุดตัดคำของผู้วิจัย cutword มีค่าความถูกต้องสูงขึ้นที่ 0.85 แต่การตัดแยกคำด้วยชุด Stop word ของ pythainlp มีค่าความถูกต้องลดลงที่ 0.82 ในเชิงลึกความแม่นยำของการจำแนกข้อมูล ในตารางที่ 4 ที่อธิบายแยกตามประเภทความรู้สึก

ตารางที่ 4 ค่าความแม่นยำการทำนายความรู้สึกด้วยเครื่องมือ MM

เครื่องมือตัดคำ		precision recall		precision recall	
Cutword	ความรู้สึกไม่ชอบ	0.84	0.86	ความรู้สึกชอบ	0.86
Stopword	ความรู้สึกไม่ชอบ	0.81	0.83	ความรู้สึกชอบ	0.83
not stopwords	ความรู้สึกไม่ชอบ	0.83	0.87	ความรู้สึกชอบ	0.86

สำหรับตารางที่ 4 ด้วยชุดเครื่องมือ MM ให้ค่าความแม่นยำกรณีไม่มีการคัดแยกคำทิ้ง ที่ 0.83 (ความรู้สึกไม่ชอบ) และ 0.86 (ความรู้สึกชอบ) และกรณีใช้ชุดคัดแยกคำของผู้วิจัยให้ค่าความแม่นยำที่ 0.84 (ความรู้สึกไม่ชอบ) และ 0.86 (ความรู้สึกชอบ) แต่กรณีใช้ชุดคัดแยกคำของ pythianlp ให้ค่าความแม่นยำเพียง 0.81 (ความรู้สึกไม่ชอบ) และ 0.83 (ความรู้สึกชอบ) เท่านั้น

2. เครื่องมือตัดคำ NEWMM

เครื่องมือชุดใหม่ NEWMM ได้พัฒนาพจนานุกรมใหม่ สำหรับใช้เพื่อการตัดคำภาษาไทยซึ่งได้รับการเพิ่มคำ ให้ค่าความถูกต้องตารางที่ 5 และค่าความแม่นยำตามตารางที่ 6

ตารางที่ 5 ค่าความถูกต้อง และจำนวนคำคุณลักษณะด้วยเครื่องมือ NEWMM

เครื่องมือตัดคำ	ค่าความถูกต้อง (accuracy)	คำคุณลักษณะ(features)
Cutword	0.90	1764
Stopword	0.81	1541
ไม่ใช่ stopwords	0.90	1833

ตามตารางที่ 5 เครื่องมือตัดคำแบบไม่คัดแยกคำได้ค่าความถูกต้องที่ 0.90 ซึ่งหากใช้ชุดตัดคำของผู้วิจัย cutword มีค่าความถูกต้องเท่ากับ 0.90 แต่การคัดแยกคำด้วยชุด Stop word ของ pythianlp มีค่าความถูกต้องลดลงที่ 0.81 และค่าความแม่นยำของการจำแนกข้อมูล ในตารางที่ 6 ที่อธิบายแยกตามประเภทความรู้สึก

ตารางที่ 6 ค่าความแม่นยำการทำนายความรู้สึกด้วยเครื่องมือ NEWMM

เครื่องมือตัดคำ		precision recall		precision recall	
cutword	ความรู้สึกไม่ชอบ	0.90	0.90	ความรู้สึกชอบ	0.90
stopword	ความรู้สึกไม่ชอบ	0.81	0.80	ความรู้สึกชอบ	0.81
ไม่ใช่ stopwords	ความรู้สึกไม่ชอบ	0.89	0.90	ความรู้สึกชอบ	0.90

สำหรับตารางที่ 6 ด้วยชุดเครื่องมือ NEWMM ให้ค่าความแม่นยำกรณีไม่มีการคัดแยกคำทิ้ง ที่ 0.89 (ความรู้สึกไม่ชอบ) และ 0.90 (ความรู้สึกชอบ) และกรณีใช้ชุดคัดแยกคำของผู้วิจัยให้ค่าความแม่นยำที่ 0.90 (ความรู้สึกไม่ชอบ) และ 0.90 (ความรู้สึกชอบ) แต่กรณีใช้ชุดคัดแยกคำของ pythianlp ให้ค่าความแม่นยำเพียง 0.81 (ความรู้สึกไม่ชอบ) และ 0.81 (ความรู้สึกชอบ) ซึ่งน้อยกว่ากรณีอื่น ๆ เช่นเดียวกับชุดเครื่องมือ MM

3. เครื่องมือตัดคำ LONGEST

เครื่องมือตัดคำนี้ เน้นที่การตัดคำที่มีความยาวเป็นหลัก และถือเป็นเครื่องมือที่มีลักษณะแตกต่างจาก 2 ชุดแรก ซึ่งการให้ค่าความถูกต้องตามตารางที่ 7 และตารางที่ 8 สำหรับค่าความแม่นยำของชุดเครื่องมือตัดคำนี้

ตารางที่ 7 ค่าความถูกต้อง และจำนวนคำคุณลักษณะด้วยเครื่องมือ LONGEST

เครื่องมือตัดคำ	ค่าความถูกต้อง (accuracy)	คำคุณลักษณะ(features)
Cutword	0.86	1883
Stopword	0.81	1573
ไม่ใช่ stopwords	0.85	1875

ตามตารางที่ 7 เครื่องมือตัดคำแบบไม่คัดแยกคำได้ค่าความถูกต้องที่ 0.85 ซึ่งหากใช้ชุดตัดคำของผู้วิจัย cutword มีค่าความถูกต้องที่สูงกว่าที่ 0.86 แต่การคัดแยกคำด้วยชุด Stop word ของ pythainlp มีค่าความถูกต้องลดลงที่ 0.81 และค่าความแม่นยำของการจำแนกข้อมูล ในตารางที่ 8 ที่อธิบายแยกตามประเภทความรู้สึก

ตารางที่ 8 ค่าความแม่นยำการทำนายความรู้สึกด้วยเครื่องมือ LONGEST

เครื่องมือตัดคำ		precision	recall		precision	recall
cutword	ความรู้สึกไม่ชอบ	0.86	0.87	ความรู้สึกชอบ	0.86	0.86
stopword	ความรู้สึกไม่ชอบ	0.82	0.80	ความรู้สึกชอบ	0.80	0.83
ไม่ใช่ stopwords	ความรู้สึกไม่ชอบ	0.85	0.86	ความรู้สึกชอบ	0.86	0.84

สำหรับตารางที่ 8 ด้วยชุดเครื่องมือ LONGEST ให้ค่าความแม่นยำกรณีไม่มีการคัดแยกคำทั้งที่ 0.85 (ความรู้สึกไม่ชอบ) และ 0.86 (ความรู้สึกชอบ) และกรณีใช้ชุดคัดแยกคำของผู้วิจัยให้ค่าความแม่นยำที่ 0.82 (ความรู้สึกไม่ชอบ) และ 0.80 (ความรู้สึกชอบ) แต่กรณีใช้ชุดคัดแยกคำของ pythainlp ให้ค่าความแม่นยำเพียง 0.86 (ความรู้สึกไม่ชอบ) และ 0.86 (ความรู้สึกชอบ) ซึ่งน้อยกว่ากรณีอื่น ๆ เช่นเดียวกับชุดเครื่องมือ MM และ NEWMM

อภิปรายผล

การทดสอบสร้างต้นแบบการจำแนกความรู้สึก การวิจารณ์โรงแรมด้วยเทคนิคต้นไม้ตัดสินใจ ผู้วิจัยทำการเปรียบเทียบประเด็นหลักของสมมติฐานด้วยเครื่องมือตัดคำภาษาไทย ได้แก่ MM NEWMM และ LONGEST โดยในแต่ละชุดมีการคัดแยกคำเพื่อเป็นการลดปริมาณคำที่ไม่จำเป็นตามกระบวนการสร้างต้นแบบให้มีประสิทธิภาพและคงไว้ซึ่งความเร็วของการในการประมวลผล มีชุดคำสำหรับคัดแยกคำทั้ง 2 แบบ แบบที่ 1 ใช้ชุดคำ Stop word ของ pythainlp และแบบที่ 2 ผู้วิจัยได้สร้างชุดคำขึ้นเองโดยมีจำนวนน้อยกว่า เพื่อให้คำที่แสดงความรู้สึกยังคงปรากฏอยู่ เพราะภาษาไทยคำปิดท้าย หรือคำสัน มักมีผลต่อการแสดงความรู้สึกในภาษาพูด ด้วยการวิจารณ์จะถูกเขียนในลักษณะของภาษาพูดมากกว่า ภาษาเขียนที่ใช้เป็นการบรรยายที่เป็นทางการ สำหรับผลการประเมิน Model ที่พบว่าให้ค่าความถูกต้องที่สุดในชุดเครื่องมือตัดคำ คือ MM (กรณี ใช้ชุดคัดแยกคำของผู้วิจัย ค่าความถูกต้อง 0.85 ค่าความแม่นยำรู้สึกชอบ 0.86 และค่าความแม่นยำรู้สึกไม่ชอบ 0.84) NEWMM (กรณี ใช้ชุดคัดแยกคำของผู้วิจัย ค่าความถูกต้อง 0.90 ค่าความแม่นยำรู้สึกชอบ 0.90 และค่าความแม่นยำรู้สึกไม่ชอบ 0.90) และ LONGEST (กรณี ใช้ชุดคัดแยกคำของผู้วิจัย ค่าความถูกต้อง 0.86 ค่าความแม่นยำรู้สึกชอบ 0.86 และค่าความแม่นยำรู้สึกไม่ชอบ 0.86) เมื่อเทียบกับการไม่ใช้ชุดคัดแยกคำซึ่งควรมีประสิทธิภาพสูงกว่า พบว่า ชุดคัดแยกคำของผู้วิจัยปรากฏประสิทธิภาพที่ดีกว่าอย่างชัดเจน ซึ่งในมุมมองคำที่พบในข้อมูลที่นำมาสร้างต้นแบบ ซึ่งผู้วิจัยนำไปสร้างกลุ่มเมฆคำเพื่อแสดงคำหลังถูกตัดคำแล้ว ตามภาพที่ 2 ถึง ภาพที่



ภาพที่ 2 กลุ่มเมฆคำการนิวิจารณ์ด้วยความรู้สึกไม่ชอบแบบไม่คัดแยกคำ



ภาพที่ 3 กลุ่มเมฆคำการนิวิจารณ์ด้วยความรู้สึกชอบแบบไม่คัดแยกคำ



ภาพที่ 4 กลุ่มเมฆคำกรณิวิจารณ์ด้วยความรู้สึกไม่ชอบแบบคัดแยกคำด้วยชุดคำปิดท้ายของงานวิจัย



ภาพที่ 5 กลุ่มเมฆคำกรณิวิจารณ์ด้วยความรู้สึกไม่ชอบแบบคัดแยกคำด้วยชุดคำปิดท้ายของงานวิจัย

สำหรับ ภาพที่ 2 และ ภาพที่ 4 คำที่มีในการวิจารณ์ด้วยความรู้สึกไม่ชอบ เมื่อตัดคำว่า “และ” “แต่” “การ” จากภาพที่ 2 ด้วยชุดคัดแยกคำของผู้วิจัย ทำให้มีคำว่า “ควร” “เป็น” ซึ่งเสริมความรู้สึกของการวิจารณ์ได้ชัดเจนขึ้น และในคำวิจารณ์ด้วยความรู้สึกชอบ ภาพที่ 3 คำที่หายไปเมื่อใช้ชุดคัดแยกคำ ได้แก่ “แต่” “ที่” แล้วปรากฏคำ “สวย” ชัดเจนขึ้น

สรุปผลและข้อเสนอแนะ

การจำแนกความรู้สึกมีการวิจัยในหลายธุรกิจ และงานวิจัยของ นลินี พุกษาศาติ และ ธนพล เจนสุทธิเวชกุล (2560) เลือกใช้เทคนิค Word2vec ร่วมกับขั้นตอนวิธีเอนีฟเบย์ ที่ให้ความแม่นยำ และความถูกต้องในชุดคำวิจารณ์ภาษาอังกฤษ แต่ในงานวิจัยนี้เลือกใช้คำภาษาไทย และสร้างต้นแบบที่มีขนาดใหญ่ด้วยคำคุณลักษณะจำนวน 1764 คำ สำหรับเครื่องมือตัดคำ NEWMM ให้ค่าความถูกต้องที่ 0.90 เทียบกับการไม่ใช้ชุดตัดแยกคำที่มีคำคุณลักษณะที่ 1833 คำ สำหรับเครื่องมือตัดคำ MM แบบเดิมก่อนพัฒนา มีค่าความถูกต้องสูงสุดเพียง 0.85 และ LONGEST ให้ค่าความถูกต้องสูงสุดเพียง 0.86 สำหรับต้นแบบที่ใช้ร่วมกับเทคนิคต้นไม้ตัดสินใจ และปรากฏประสิทธิภาพสูงสุดสรุปได้คือ NEWMM

การพัฒนาต่อยอดเพื่อให้ประสิทธิภาพในการประมวลผลที่มีจำนวนมากและความรวดเร็ว ด้วยการลดจำนวนคำคุณลักษณะที่มีหลักการทางสถิติ เพื่อใช้เลือกคำ และกระบวนการ Feature Selection ที่นิยมสำหรับการเพิ่มความเร็วในการจำแนก หรือ การใช้เทคนิคอื่น ๆ ในการจำแนกความรู้สึกของการทำเหมืองข้อมูล เพื่อเพิ่มประสิทธิภาพในโอกาสต่อไป

กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา ในส่วนของต้นสังกัดที่ส่งเสริม และผลักดันการทำวิจัยให้เกิดขึ้น และสนับสนุนการเผยแพร่งานวิจัยในเวทีระดับชาติต่อเนื่อง และสม่ำเสมอตลอดมา กับทีมผู้วิจัยในโอกาสนี้

รายการอ้างอิง

- จิราภา เลหาหะวรนนท์, รชต. ลิ้มสุทธิวันภูมิ และบัณฑิต ฐานะโสภณ. (2558). การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือก แขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. **วารสารเทคโนโลยีสารสนเทศลาดกระบัง**. ฉบับที่ 4 ปี 2558 : 1-9
- นลินี พุกษาศาติ และ ธนพล เจนสุทธิเวชกุล. (2560). **การจำแนกความคิดเห็นออนไลน์ของนักท่องเที่ยว โดยใช้ Word Vector และการเรียนรู้ของเครื่อง**. ภาควิชาการเทคโนโลยีสารสนเทศ มหาวิทยาลัย เทคโนโลยีพระจอมเกล้าพระนครเหนือ
- อุษนิษา เถาว์โท, ฉัฐเมศร์ พูลทรัพย์, เกียรติกุล เจียรนัยธนะกิจ และ บัณฑิต พัสยา. (2561). การทดสอบโมเดลของระบบแนะนำภาพยนตร์โดยการทำเหมืองข้อความจากคำวิจารณ์ของผู้ชม. **การประชุมวิชาการปัญญาภิวัฒน์ระดับชาติ ครั้งที่ 8**. 1-9
- Wannaphong. (2560). **PyThaiNLP โมดูล NLP ภาษาไทยใน Python** (Online). <https://python3.wannaphong.com/2017/05/pythainlp-nlp-python.html>, กรกฎาคม 15, 2563
- Puidui. (2552). **SYSTEM ANALYSIS TECHNIQUES AND DESIGN FOR BUSINESS** (Online). http://sasdkmitl09.blogspot.com/2009/07/blog-post_23.html, 18 กรกฎาคม 2563
- Kasidis Satangmongkol. (2562). **อธิบาย K-Fold Cross Validation พร้อมโค้ดตัวอย่างใน R** (Online). <https://datarockie.com/2019/11/15/k-fold-cross-validation>, 20 กรกฎาคม 2563