การจำแนกความคิดเห็นของคนไทยเกี่ยวกับสื่อออนไลน์โดยใช้การทำเหมืองข้อความ

Classifying Thai opinions on online media using text mining

นุชนาฎ ปิ่นเมือง[1,*], จารี ทองคำ[2]

Noochanat Pinmuang[1], Jaree Thongkam[2]

Received: 27 March 2017 ; Accepted: 12 October 2017

บทคัดย่อ

เหมืองข้อความเป็นกระบวนการวิเคราะห์ข้อมูลตัวอักษรเพื่อสกัดข้อมูลที่เป็นประโยชน์จากแหล่งข้อมูล ปัจจุบันเทคนิคในการจำแนกเหมืองข้อความมีหลายวิธี เพย์ งานวิจันนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคการจำแนก จาก 5 เทคนิคที่มีประสิทธิภา คือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine (SVM) เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ (Decision tree) และเทคนิค C4.5 ในการเก็บรวบรวมข้อมูลความคิดเห็นต่อการใช้บริการพร้อมเพย์บนสื่อออนไลน์จำนวนทั้งหมด 1,570 ข้อความ ในกระบวนการคัดเลือกคำบ่งชี้เพื่อใช้ในการแยกคุณลักษณะได้เลือกใช้คำวิเศษณ์เพื่อทำการแยกคุณลักษณะเชิงบวกและเชิงลบ คณะผู้วิจัยได้ใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพการจำแนกของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) เมื่อทำการทดสอบและวัดประสิทธิภาพของโมเดลพบว่า เทคนิค Naïve Bayes ให้ผลดีที่สุดในการจำแนกข้อความคิดเห็น โดยให้ค่าความถูกต้อง 93.88% ค่าความแม่นยำ 94.02% และค่าความระลึก 93.54%

คำสำคัญ เหมืองข้อความ การจัดกลุ่มความคิดเห็น พร้อมเพย์

Abstract

Text mining is one of the most effective data analysis processes using alphabetic methods. Currently, text mining techniques are classified a variety of ways. This research aims to find the most effectiveness of 5 techniques that were Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor, Decision tree และ C4.5. The data collected were all made by the viewers in a total of 1,570 messages. The categorization process divided the data into 2 groups: positive character and negative character. Interestingly, the process only indicated to that adverbs can be selected as a core division to produce positive and negative characters. 10-fold cross validation was applied to segment the data into training and testing sets. Moreover, accuracy, precision and recall were used as the criteria for selecting the most effective model. It was concluded that the Naïve Bayes technique produced the greatest accuracy in categorizing the messages with an accuracy score of 93.88%, precision of 94.02% and recall of 93.54%.

Keywords: Text mining, Grouping the opinion, Classification the opinion, Prompt pay

[1] *นิสิต, สาขาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, มหาสารคาม, 44150.*

[2] *ผู้ช่วยศาสตราจารย์, อาจารย์ที่ปรึกษา หน่วยวิจัยสารสนเทศประยุกต์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, มหาสารคาม, 44150.*

[1] *Student, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Maha Sarakham, Thailand, 44150.*

[2] *Assistant Professor, Applied informatics Research Unit, Faculty of Informatics, Mahasarakham University, Maha Sarakham, Thailand, 44150.*

** Corresponding author: Tel: +66 086 8596761 Email address: anakinnooch1983@gmail.com*

บทนำ

พร้อมเพย์ เป็นบริการเพิ่มจากการโอนเงินแบบเดิมซึ่งถือว่าเป็นอีกทางเลือกใหม่ให้ประชาชน ผู้ประกอบธุรกิจและหน่วยงานต่างๆ ใช้ในการโอนเงินและรับเงิน ทำให้ประชาชนสามารถเลือกใช้งานได้อีกทั้งทำให้ผู้ใช้มีความสะดวกมากขึ้น[1] อย่างไรก็ตามระบบพร้อมเพย์ก็ยังถูกวิจารณ์และมีการแสดงความคิดเห็นทั้งในแง่ความคิดเห็นเชิงบวกและเชิงลบบนสื่อออนไลน์หลายๆ สื่อด้วยกัน โดยเฉพาะบนเฟสบุ๊คซึ่งเป็นสื่อออนไลน์และคำวิจารณ์อยู่ในรูปแบบของข้อความคิดเห็นซึ่งมีลักษณะเป็นภาษาธรรมชาติหรือภาษาพูดนั้นเอง ถือว่าเป็นลักษณะภาษาที่ไม่เป็นทางการและไม่มีรูปแบบที่แน่นอน ดังนั้นกระบวนการวิเคราะห์ข้อความจะมีรูปแบบที่แตกต่างกันออกไปตามลักษณะของความซับซ้อนทางความหมายของคำและประโยค

เหมืองข้อความเป็นกระบวนการเพื่อสกัดเอาความรู้จากภาษาธรรมชาติที่มีลักษณะของข้อมูลแบบไม่มีโครงสร้าง อีกทั้งไม่มีการกำหนดรูปแบบไว้ล่วงหน้า[2] ผลของข้อมูลที่ได้จึงไม่มีความแน่นอนของเหตุผลนั้นๆ ที่แฝงอยู่ในกลุ่มของข้อความซึ่งมีลักษณะเป็นภาษาธรรมชาติ การทำเหมืองข้อมูลจำนวนมากนั้นจะเป็นลักษณะของการนำเสนอรูปแบบสำหรับการทำเหมืองข้อความที่ใช้ประโยชน์ในเอกสารและใช้ความรู้ที่ได้จากการสกัดข้อความเหล่านั้นมาเพื่อปรับปรุงรูปแบบและปัญหาต่างๆ ของสิ่งนั้น[3] การค้นพบรูปแบบยังคงเป็นปัญหาเนื่องจากส่วนใหญ่แล้ววิธีการทำเหมืองข้อความมักจะประสบปัญหาของหลักภาษาและคำพ้องซึ่งมีลักษณะที่ซับซ้อนกันออกไป โดยได้มีการจัดสมมติฐานที่เกี่ยวกับกลุ่มคำหรือวลี ซึ่งวิธีดังกล่าวน่าจะได้ผลการทำงานที่ดีกว่า ในการวิเคราะห์ข้อความที่ไม่มีโครงสร้างแบบอัตโนมัติจึงเป็นเรื่องที่ท้าทายการนำเทคนิคการทำเหมืองข้อความมาประยุกต์ในการทำเหมืองแสดงความคิดเห็นซึ่งเป็นอีกกระบวนการวิเคราะห์เหมืองข้อความ[4] โดยการนำเอาข้อคิดเห็นมาทำการวิเคราะห์เพื่อให้ทราบถึงความพึงพอใจที่มีต่อสิ่งนั้นๆ และในการสกัดคำตามคุณลักษณะที่แตกต่างกันอาจได้มาซึ่งประโยชน์เพื่อการพิจารณาวิเคราะห์ที่หลากหลายและแม่นยำมากขึ้น ซึ่งมีนักวิจัยหลายท่านได้นำเอาเทคนิคการจำแนกในเหมืองข้อมูลและการเรียนรู้ของเครื่องมาทำการสกัดความรู้เพื่อใช้เป็นต้นแบบการจำแนก เช่น ประพัฒน์ พรมน้ำอ่างและคณะ[5] ได้นำเสนอการจำแนกกลุ่มข้อความแสดงความคิดเห็นที่มีต่อสินค้าโดยการใช้เทคนิคเหมืองข้อมูล เพื่อหาความแม่นยำของรูปแบบการจำแนกข้อความแสดงความคิดเห็นเพื่อนำไปพัฒนารูปแบบการจำแนกข้อความ ซึ่งได้ทำการเปรียบเทียบ 4 เทคนิค คือเทคนิค SVM เทคนิคต้นไม้ตัดสินใจ เทคนิค K-Nearest Neighbor และเทคนิค Naïve Bayes ผลการทดลองพบว่าเทคนิค SVM ได้ให้ค่าความถูกต้องมากที่สุดที่ 86.26% ที่สุด

พัชรนิกานต์ พงษ์ธนู และคณะ[6] ได้นำเสนอรูปแบบการวิเคราะห์เหมืองข้อความจากการเก็บข้อมูลการแสดงความพึงพอใจของลูกค้าจากข้อความคำแนะนำบนเว็บไซต์เพื่อหาแนวทางในการปรับปรุงการบริการของเว็บไซต์ผู้ให้บริการโรงแรมให้มีประสิทธิภาพมากขึ้น โดยงานวิจัยนี้ได้ใช้วิธีการสกัดคำความคิดเห็นด้านดีและด้านไม่ดีเพื่อสรุปการให้บริการของเว็บไซต์ และเพื่อค้นหาว่าจำนวนของคำบ่งชี้มีผลต่อการให้ค่าความถูกต้องหรือไม่ ซึ่งได้เปรียบเทียบผลจากการสร้างโมเดลที่ใช้ในการวิเคราะห์ด้วยเทคนิควิธีต้นไม้ตัดสินใจโดยการใช้อัลกอริทึม ID3 และเทคนิค Naïve Bayes โดยผลของการทดลองนั้นเทคนิคต้นไม้ตัดสินใจให้ค่าเฉลี่ยของมากที่สุดที่ 95.50%

อดิเทพ ไชยสารและรัฐสิทธิ์ สุขะหุต[7] ได้นำเสนอผลการเปรียบเทียบการประมาณอารมณ์จากความคิดเห็นภาษาไทยโดยใช้วิธีการจำแนกอารมณ์จากข้อความแสดงความคิดเห็นจากเว็บไซต์บริการข่าวและบริการวิจารณ์สินค้า โดยได้ใช้เทคนิคการจำแนก 3 เทคนิค ด้วยกันคือ เทคนิค Naïve Bayes เทคนิค SVM และเทคนิคต้นไม้ตัดสินใจ ซึ่งได้มีผู้เชี่ยวชาญจำแนกความคิดเห็นตามกลุ่มอารมณ์ 6 กลุ่มอารมณ์ ซึ่งผลการทดสอบการความถูกต้องในการประมาณอารมณ์โดยรวมพบว่าเทคนิค SVM สามารถประมาณอารมณ์ได้ถูกต้องที่สุดที่ 69.15%

Peiman Barnaghi และ John G. Breslin[8] ได้นำเสนอกระบวนการสร้างโมเดลในการวิเคราะห์ความคิดเห็นเพื่อค้นหาภาพสะท้อนความเชื่อมั่นของประชาชนต่อเหตุการณ์ต่างๆ เพื่อนำมาประยุกต์ใช้เป็นพื้นฐานสำหรับการคาดคะเนเหตุการณ์ในอนาคต ผ่านการใช้งานบน Twitter โดยการใช้ข้อคิดเห็นเกี่ยวกับการจัดงาน FIFA World Cup 2014 โดยทำการพิจารณาข้อคิดเห็นที่เป็นเชิงบวกและเชิงลบ จากนั้นได้ทำการเปรียบเทียบการสร้างโมเดลจากการใช้เทคนิค Bayesian Logistic Regression (BLR) และเทคนิค Naïve Bayes โดยผลการทดสอบพบว่าเทคนิค BLR ให้ผลที่ดีกว่าการใช้เทคนิค Naïve Bayes

Mondher Bouazizi และ Tomoaki Ohtsuki[9] ได้นำเสนอกระบวนการวิเคราะห์ข้อความคิดเห็นของภาษาบนสื่อออนไลน์และเปรียบเทียบโมเดลในการจำแนกข้อความคิดเห็นที่ดีที่สุด เพื่อนำมาวิเคราะห์การจำแนกหัวข้อและทัศนคติที่ผู้ใช้งาน Twitter สนใจ และได้ใช้เทคนิคการจำแนก 2 เทคนิค ด้วยกันคือ เทคนิค Naïve Bayes และเทคนิค SVM เมื่อทำการ

เปรียบเทียบผลการทดลองของทั้งสองเทคนิควิธีด้วยข้อมูล 2 ลักษณะ คือ ข้อมูลแสดงความคิดเห็นและข้อมูลแสดงความคิดที่ผ่านกระบวนการวิเคราะห์ความเชื่อมั่นแล้ว ผลที่ได้จากการทดลองนั้น คือ เทคนิค SVM ให้ผลของการจำแนกดีที่สุด

จากงานวิจัยที่ได้มีการศึกษานั้น พบว่างานวิจัยส่วนใหญ่มีการจำแนกความคิดเห็นออกเป็น 2 กลุ่ม คือ ความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ ดังนั้นในงานวิจัยนี้จึงได้ทำเปรียบเทียบเทคนิคที่ใช้ในสร้างแบบจำลองเพื่อจำแนกความคิดเห็น คือ เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิค C4.5 จากนั้นใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

โดยแบบจำลองนี้สามารถนำไปจำแนกกลุ่มความสนใจของผู้ใช้ได้อัตโนมัติทำให้เกิดความรวดเร็วต่อการจำแนกกลุ่มของผู้ใช้พร้อมเพย์ซึ่งจะมีประโยชน์ต่อธนาคารเพื่อใช้ในการปรับปรุงและเพิ่มประสิทธิภาพในงานบริการ

วัตถุประสงค์

เพื่อทำเปรียบเทียบเทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิค C4.5 ในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็น จากนั้นใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

ทฤษฎีที่เกี่ยวข้อง

ในงานวิจัยนี้เทคนิคที่นำมาใช้ในการสร้างแบบจำลองได้แก่[10] เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิคต้นไม้ตัดสินใจ C4.5

1. เทคนิค Naïve Bayes

เป็นเทคนิควิธีการจำแนกประเภทที่อาศัยหลักการของทฤษฎีความน่าจะเป็นตามกฎของเบย์ เพื่อหาว่าสมมติฐานใดน่าจะมีความถูกต้องมากที่สุด ซึ่งจะสามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลชุดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลนั้นๆ โดยทฤษฎีของเบย์สามารถคำนวณได้ดังสมการที่ 1

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

โดย X แทนข้อมูลการแจกแจงความน่าจะเป็น
P(H) คือ ความน่าจะเป็นก่อนหน้าของ
HP(X) คือ ความน่าจะเป็นก่อนหน้าของ X
P(H|X) คือ ความน่าจะเป็นของ H เมื่อรู้ X
P(X|H) คือ ความน่าจะเป็นของ X เมื่อรู้ H

2. เทคนิค Support Vector Machine

คือ ขั้นตอนวิธีการที่มีความรวดเร็วและเป็นอัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด แนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) ในลักษณะเชิงเส้น จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกัน โดยจะทำการสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้น เส้นใดเป็นเส้นที่ดีที่สุด

3. เทคนิค K-Nearest Neighbor

เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบาง ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด ซึ่งขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดนั้นเป็นเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูล โดยการจัดข้อมูลที่อยู่ใกล้กันให้เป็นกลุ่มเดียวกันซึ่งเทคนิคนี้จะทำให้ตัดสินใจได้ว่า คลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวน K ซึ่งถ้าหากเงื่อนไขของการตัดสินใจมีความซับซ้อน วิธีนี้สามารถสร้างโมเดลที่มีประสิทธิภาพได้

4. เทคนิคต้นไม้ตัดสินใจ (Decision tree)

เป็นเทคนิควิธีเหมืองข้อมูลที่เป็นที่นิยมกันมาก เนื่องจากเป็นวิธีการวิเคราะห์ที่ง่ายต่อการตีความหมาย ต้นไม้ตัดสินใจนั้นจะประกอบไปด้วยโหนด (Node) ซึ่งจะทำหน้าที่ในการแสดงคุณลักษณะที่ใช้สำหรับการทดสอบข้อมูล กิ่ง (Branch) เป็นส่วนที่จะแสดงคุณสมบัติในโหนดที่ได้มีการแตกออกมา และใบ (Leaf) จะแสดงกลุ่มหรือคลาสที่ได้มีการกำหนดเอาไว้ และในหาความสัมพันธ์ของแต่ละโหนดที่แสดงคุณลักษณะ (Attribute) นั้นจะใช้ค่า Information Gain เพื่อหาความสัมพันธ์ของในแต่ละโหนดคุณลักษณะ โดยการใช้ค่า Information Gain นี้จะช่วยลดจำนวนครั้งที่ใช้ในการทดสอบ

และทำให้ต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินไป

5. เทคนิคต้นไม้ตัดสินใจแบบ C4.5

เป็นเทคนิควิธีหนึ่งที่ได้ทำการพัฒนามาจากเทคนิค ID3 ซึ่งได้กลายมาเป็นอัลกอริทึมพื้นฐานที่ใช้ในการเปรียบเทียบประสิทธิภาพของการทำงานในอัลกอริทึมต่างๆ โดยในการสร้างต้นไม้ตัดสินใจ C4.5 จะใช้ค่ามาตรฐานส่วนเกน (Gain Ratio) เพื่อทำการคัดเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด

วิธีการวิจัย

งานวิจัยการจัดกลุ่มข้อความคิดเห็นนี้ได้นำเอากระบวนการวิเคราะห์เหมืองข้อความ[10] 4 ขั้นตอนดังนี้

1. การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลการแสดงความคิดเห็นได้ทำการแบ่งการรวบรวมข้อมูลในขั้นตอนนี้ ได้แบ่งออกเป็น 2 ส่วน คือ

1.1 การเก็บข้อมูลแสดงความคิดเห็น

ในการเก็บรวบรวมข้อมูล ผู้วิจัยได้ทำการเก็บรวบรวมข้อมูลจากการแสดงความคิดเห็นผ่านการใช้งาน Facebook Graph API โดยทำการเก็บรวบรวมข้อมูลที่มีการแสดงความคิดเห็นตั้งแต่เดือนมิถุนายน 2559 – มกราคม 2560 จำนวน 1,570 ข้อความคิดเห็น โดยในแต่ละความคิดเห็นจะมีการแสดงข้อมูลของผู้แสดงความคิดเห็นคือแสดงชื่อผู้ใช้งานและรหัสประจำตัวของผู้ใช้งาน ในส่วนของข้อความจะแสดงข้อความแสดงความคิดเห็นและรหัสประจำข้อความและทำการบันทึกข้อความคิดเห็นลงฐานข้อมูล ดังแสดงใน Figure 1

1.2 กระบวนการตัดคำและกำจัดคำหยุด

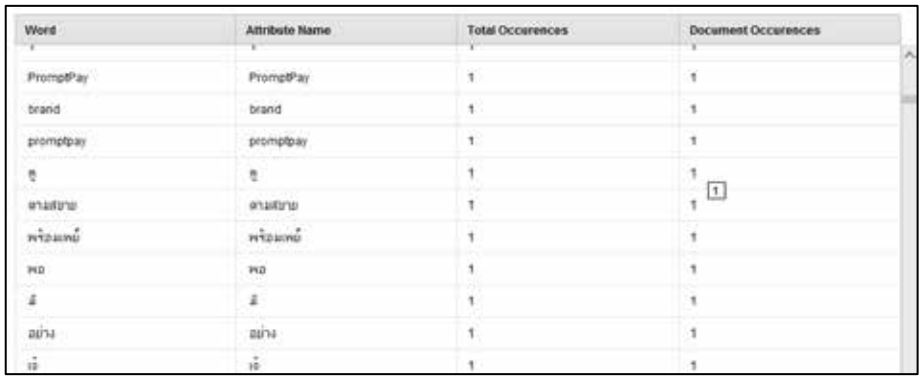
กระบวนการนี้จะเป็นการนำเอาข้อคิดเห็นที่ได้เก็บรวบรวมมาทำการคัดแยกข้อคิดเห็นที่เกี่ยวข้องกับหัวข้อที่กำหนดและจะต้องเป็นข้อความภาษาไทยเท่านั้น จากนั้นนำเข้าสู่กระบวนการตัดคำดังแสดงใน Figure 2 จากนั้นจึงเข้าสู่กระบวนการหาค่าความถี่การใช้คำในเอกสารโดยมีคำทั้งหมดรวมถึงคำภาษาไทยและภาษาอังกฤษจำนวน 3,581 คำ ดังแสดงใน Figure 3 แล้วทำการตัดคำที่เป็นภาษาอังกฤษและตัวเลขออกไปเหลือคำจำนวน 3,302 คำ จากนั้นมาให้ประเภทตามประเภทของพจนานุกรมของแต่ละคำโดยแบ่งเป็นทั้งหมด 11 ประเภท ได้แก่ คำนาม คำสรรพนาม คำกริยา คำวิเศษณ์ คำสันธาน คำบุพบท วลี คำอุทาน คำหยุด คำแสลง และคำไม่ตรงพจนานุกรม

no	comment	date
1	แล้วมันจะพร้อมใช้เมื่อไหร่ครับ รอธนาคารพาณิชย์ก็คงรอไปเรื่อยๆครับ ถ้าแบงค์ชาติไม่ออกมาบังคับ	14-12-16
2	ผมพร้อมจะเพย์ มาตั้งแต่สมัครแล้วครับ เหลือปล่อยให้ระบบpromptใช้งาน	14-12-16
3	ขอบคุนกรุงไทยพร้อมเพย์ค่ะ ฉันได้รับเงินช่วยจากรัฐบาลแล้วค่ะ เมื่อสักครู่นี้เอง1500บาท	09-12-16
4	ใช้เถอะ เพื่อภาครัฐจะได้ติดตามธุรกรรมทางการเงินของท่านได้ง่าย ถูกกฎหมาย ไม่ต้องมีหมาย ไหร	08-12-16
5	อันไหนของรัฐบาลชุดนี้ทำ ไม่ปัง เกิดยาก คิดมาแต่ละอย่าง เอาหวยสามตัวท้าย ไปออกสามตัวหน้า	29-11-16
6	สมัครแล้วแต่ก็ต้องเสียค่าธรรมเนียมโอนเหมือนเดิมเพราะอีกฝ่ายไม่ได้สมัคร	29-11-16
7	อย่าโอนผิดไม่ถูกต้องก็แล้ว,เพราะปัจจุบันนักการเมืองเขาป้องหาเรื่องตลอด,ระวังอย่าพลาดก็แล้วกัน	10-12-16
8	เห็นด้วยกับธนาคาร ที่ต้องก้าวทันโลก เพื่อสนองนักธุรกิจทั่วทุกมุมโลก และเปิดกว้างกับลูกค้าทุกระ	13-12-16
9	ประชาชนจะกลัวธุรกรรมอิเล็กทรอนิกส์ไม่ใช่เพราะอะไร แต่เพราะเวลาเกิดอะไรขึ้นแล้วแบงค์ไม่ค่อย	13-12-16
10	จากอดีตกาลมาสอนให้เราสำนึกว่า เราต้องมีถึงจะนำไปใช้จ่าย เราจึงอยู่กัน #อย่างสุขสบาย อยู่กันอ	06-12-16

Figure 1 Sample comments data saved to database

no	comments	num
1	แล้ว , มัน , จะ , พร้อม , ใช้ , เมื่อไหร่ , ครับ , รอ , ธนาคารพาณิชย์ , ก็ , คง , รอ , ไ	17
2	ผม , พร้อม , จะ , เพย์ , มา , ตั้ง , แต่ , สมัคร , แล้ว , ครับ , เหลือ , แค่,ให้ระบบ, pr	15
3	ขอบ , คุน , กรุงไทย , พร้อม , เพย์ , ค่ะ , ฉัน , ได้ , รับเงิน , ช่วย , จาก , รัฐบาล , แล้	19
4	ใช้ , เถอะ , เพื่อ , ภาครัฐ , จะ , ได้ , ติดตาม , ธุรกรรม , ทางการเงิน , ของ , ท่าน , ไ	42
5	อัน , ไหน , ของ , รัฐบาล , ชุด , นี้ , ทำ , ไม่ , ปัง , เกิด , ยาก , คิด , มา , แต่ละ , อย่	31
6	สมัคร , แล้ว , แต่ , ก็ , ต้อง , เสีย , ค่า , ธรรมเนียม , โอน , เหมือนเดิม , เพราะ , อีก	15
7	อย่า , โอน , ผิด , ไม่ , ถูกต้อง , ก็ , แล้ว , เพราะ , ปัจจุบัน , การเมือง , เขา , ช้อง , ห	18
8	เห็นด้วย , กับ , ธนาคาร , ที่ , ต้อง , ก้าว , ทันโลก , เพื่อ , สนอง , นักธุรกิจ , ทั่วทุก ,	19
9	ประชาชน , จะ , กลัว , ธุรกรรม , อิเล็กทรอนิกส์ , ไม่ใช่ , เพราะอะไร , แต่ , เพราะ ,	17
10	จาก , อดีตกาล , มา , สอน , ให้ , เรา , สำนึก , ว่า , เรา , ต้อง , มี , ถึง , จะ , นำ , ไป	22

Figure 2 The sentences were cut and then saved to database



Word	Attribute Name	Total Occurences	Document Occurences
PromptPay	PromptPay	1	1
brand	brand	1	1
promptpay	promptpay	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1
[illegible]	[illegible]	1	1

Figure 3 Counting the number of words that appear in this document

2. ขั้นตอนก่อนการสร้างแบบจำลอง

เป็นขั้นตอนที่ใช้ในการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลเพื่อเข้าสู่กระบวนการจำแนกข้อความแสดงความคิดเห็นที่มีต่อการใช้บริการพร้อมเพย์ โดยจะนำข้อความคิดเห็นที่ผ่านกระบวนการคัดแยกข้อความที่เกี่ยวข้องกับข้อคิดเห็นและผ่านกระบวนการตัดคำ กำจัดคำหยุด จากนั้นเลือกเอาเฉพาะคำที่เป็นคำวิเศษณ์จำนวน 577 คำ มาคัดแยกตามความหมายเชิงบวก เชิงลบ และเป็นกลางแล้วจะคงเหลือคำเพื่อบ่งชี้ลักษณะจำนวน 400 คำ โดยแบ่งคำหรือกลุ่มคำนั้นออกเป็น 2 กลุ่ม คือ คำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบ

1) วิธีการสร้างตัวแทนเอกสารนั้นจะใช้วิธีในการนำคำบ่งชี้คุณลักษณะในชุดข้อมูลมาเรียงกันเพื่อทำการนับความถี่ของการเกิดขึ้นของคำนั้นๆ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และคำบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 จากนั้นทำการนับจำนวนคำในแต่ละคุณลักษณะ โดยใช้การนับจำนวนความถี่ของค่าคุณลักษณะในแต่ละคุณลักษณะว่ามีจำนวนเท่าใด เพื่อนำค่าความถี่ในการใช้คำของแต่ละคุณลักษณะมาทำการเปรียบเทียบกันโดย

2) เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกมากกว่าความถี่ของคุณลักษณะเชิงลบให้ตัวแปรตามเป็น ความคิดเห็นเชิงบวก แทนด้วย P

3) เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงลบมากกว่าความถี่ของคุณลักษณะเชิงบวกให้ตัวแปรตาม เป็น ความคิดเห็นเชิงลบ แทนด้วย N

4) เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกและความถี่ของคุณลักษณะเชิงลบเท่ากันให้ตัวแปรตาม เป็น ความคิดเห็นเป็นกลาง แทนด้วย B

ซึ่งผลจากการสร้างตัวแทนเอกสารแสดงใน Figure 4

3. การสร้างแบบจำลอง

ในกระบวนการสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการจัดกลุ่มความคิดเห็นที่มีต่อการบริการพร้อมเพย์ได้ใช้ข้อมูลความคิดเห็นที่เป็นภาษาไทย โดยใช้เทคนิคการจำแนกของการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 5 เทคนิค[10] คือ

1) เทคนิค Naïve Bayes
2) เทคนิค Support Vector Machine
3) เทคนิค K-Nearest Neighbor
4) เทคนิคต้นไม้ตัดสินใจ (Decision tree)
5) เทคนิคต้นไม้ตัดสินใจแบบ C4.5

row no	text	กลัว	กับดัก	กิลิ	ไร้กังวล	ไว้ใจ	จำนวนคำ	positive	negative	Process
1	ใช้	0	0	0	0	0	1	1	0	P
2	ใช่ เสียค่าธรรมเนียม	0	0	0	0	0	2	1	1	B
3	ดี ไม่ต้อง รุ่นวาย	0	0	0	0	0	3	1	2	N
4	ไม่เข้าใจ อังไม่ได้ ไม่รู้ ถูกต้อง	0	0	0	0	0	4	1	3	N
5	รุ่นวาย เสียภาษี ยัง รุ่นวาย	0	0	0	0	0	4	0	4	N
6	ได้ ทันที ได้ ไม่ได้ เสีย	0	0	0	0	0	5	3	2	P
7	ไม่ใช้ อังไง บังคับ ใช้ อังไม่ ใช้ ไม่ใช่	0	0	0	0	0	7	2	5	N

Figure 4 The sample data, through the series of features

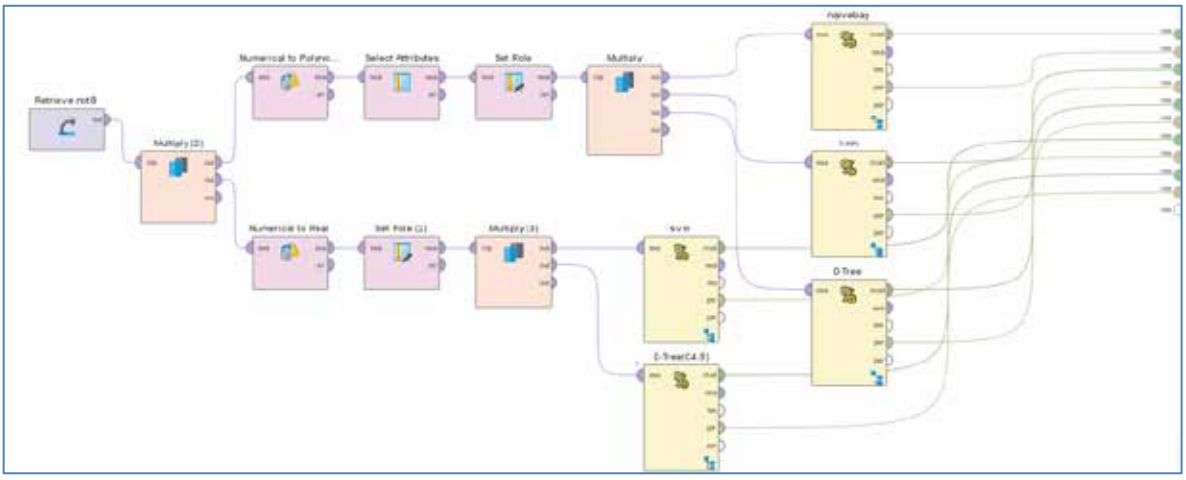


Figure 5 Import data into techniques to model with RapidMiner Studio V7.4.000

4. การวัดประสิทธิภาพของแบบจำลอง

ในการประเมินผลของแบบจำลอง ใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold cross validation โดยทำการแบ่งข้อมูลออกเป็น 10 กลุ่ม เท่าๆ กัน จากนั้นในแต่ละรอบการทดสอบจะใช้ข้อมูล 1 ชุดเป็นทดสอบและให้อีกชุดที่เหลือเป็นชุดฝึกสอน ซึ่งผู้วิจัยได้ทำการทดสอบแบบจำลองในการจัดกลุ่มความคิดเห็นมีต่อการบริการพร้อมเพย์ โดยใช้แบบจำลองที่ละ 1 แบบจำลอง และนำข้อมูลทั้งหมดที่ได้จากการเก็บข้อมูลการแสดงความคิดเห็นซึ่งผ่านกระบวนการเตรียมข้อมูลแล้วจากข้อมูลทั้งหมด 1,015 ชุด ทำการแบ่งออกเป็นทั้งหมด 10 กลุ่ม ทั้งนี้จะแบ่งกลุ่มข้อมูลเพื่อใช้เป็นข้อมูลทดสอบ (Data test) 1 ชุด และที่เหลือจะเป็นข้อมูลฝึก (Data training) ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อปริมาณข้อมูลฝึกคิดเป็นอัตราร้อยละ 10:90 โดยค่าที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง คือได้ ความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) และค่าความระลึก (recall) ดังสมการที่ 2, 3 และ 4 ตามลำดับ

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

โดย TP คือ จำนวนข้อมูลที่ถูกนำมาใช้อย่างถูกต้อง
TN คือ จำนวนข้อมูลที่ผิดที่ถูกนำมาใช้
FP คือ จำนวนข้อมูลที่ถูกต้องแต่ไม่นำมาใช้
FN คือ จำนวนข้อมูลที่ผิดแต่ไม่นำมาใช้

ผลการวิจัย

งานวิจัยนี้ได้ทำการทดลองจำแนกความคิดเห็นเกี่ยวกับการใช้บริการพร้อมเพย์ด้วยเทคนิคการจำแนกของการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 5 เทคนิค คือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิค C4.5 ผ่านการใช้งานโปรแกรม RapidMiner Studio V7.4.000 โดยชุดข้อมูลเอกสารที่ใช้ในงานวิจัยนี้ รวบรวมจากเฟสบุ๊ค จำนวน 1,570 ชุดข้อมูล จากนั้นทำการค้นหาคุณลักษณะของคำเพื่อทำการกำหนดคุณลักษณะของแต่ละประโยค โดยใช้คำเพื่อทำการกำหนดคุณลักษณะทั้งหมด 400 คำ แบ่งออกเป็นคำบ่งชี้คุณลักษณะเชิงบวกจำนวน 123 คำ และคำบ่งชี้คุณลักษณะเชิงลบจำนวน 277 คำ จากนั้นทำการคัดเลือกข้อมูลเข้าสู่โมเดล โดยในการคัดเลือกครั้งแรกทำการตัดข้อมูลที่ไม่ปรากฏคำแสดงคุณลักษณะ คือ มีค่าของจำนวนคำเป็น 0 ซึ่งในขั้นแรกคงเหลือจำนวนข้อมูล 1,215 ชุดข้อมูล จากนั้นเลือกเพียงชุดข้อมูลที่เป็น คลาส P และ N เท่านั้น ซึ่งจำนวนชุดข้อมูลที่เตรียมเข้าสู่โมเดลทั้ง 5 จะเหลือเพียง 1,015 ชุดข้อมูล

ในการประเมินประสิทธิภาพของโมเดลทั้ง 5 โมเดล โดยใช้การทดสอบโมเดลด้วยวิธีการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก

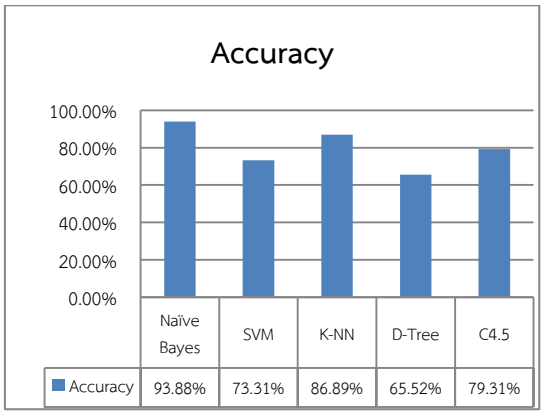


Figure 6 Comparison of accuracy

จาก Figure 6 เป็นการเปรียบเทียบค่าความถูกต้องของ 5 เทคนิคประกอบด้วย เทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และ เทคนิค C4.5 ผลปรากฏว่าว่า เทคนิค Naïve Bayes ให้ความถูกต้องสูงสุดอย่างมีนัยสำคัญที่ 93.885% รองลงมาคือเทคนิค K-NN ที่ 86.89%

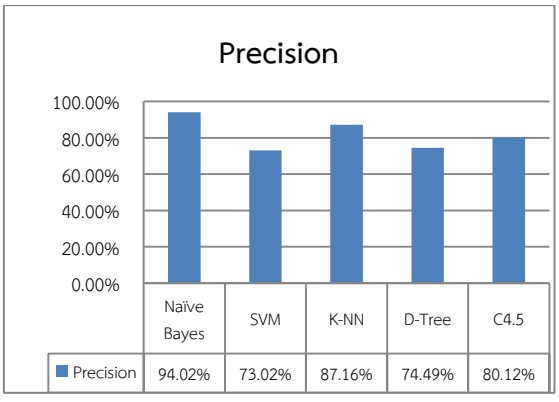


Figure 7 Comparison of precision

ในส่วนของการเปรียบเทียบค่าความแม่นยำใน Figure 7 จะเห็นได้ว่าผลของการเปรียบเทียบนั้น เทคนิค Naïve Bayes ให้ผลของความแม่นยำดีที่สุดอย่างมีนัยสำคัญที่ 94.02% รองลงมาคือเทคนิค K-Nearest Neighbor ที่ค่า 87.16% และใน Figure 8 จะแสดงการเปรียบเทียบค่าความระลึกของโมเดลผลจากการเปรียบเทียบทั้ง 5 เทคนิค จะเห็นได้ว่าเทคนิค Naïve Bayes ให้ผลดีที่สุดอย่างมีนัยสำคัญที่ 93.54% รองลงมาคือเทคนิค K-Nearest Neighbor ที่ 88.24%

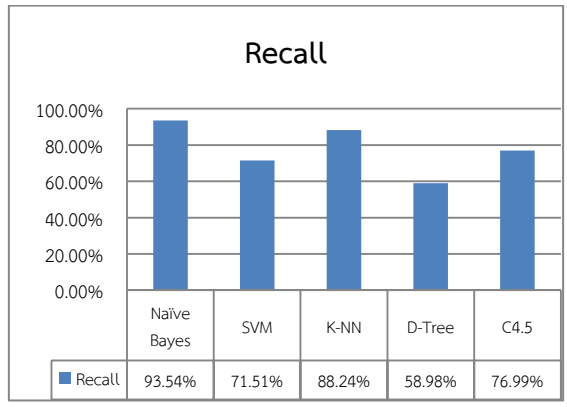


Figure 8 Comparison of recall

จากการเปรียบเทียบของทั้ง 5 โมเดล จะเห็นได้ว่าเทคนิค Naïve Bayes ให้ผลการทดสอบดีที่สุดอย่างมีนัยสำคัญโดยมีค่าความถูกต้อง 93.88% ค่าความแม่นยำ 94.02% และค่าความระลึก 93.54% และรองลงมาจะเป็นเทคนิค K-Nearest Neighbor

แม้ว่าโมเดลการจำแนกกลุ่มของชุดข้อมูลที่ใช้ในการทดสอบนี้จะได้ผลลัพธ์ที่ดี แต่อย่างไรก็ตามข้อผิดพลาดที่เกิดขึ้นจากการจำแนกกลุ่มของโมเดลทั้ง 5 นี้ ได้ใช้วิธีการในการจำแนกกลุ่มตามคุณลักษณะของคำ ซึ่งการดำเนินการเพื่อให้ได้คำที่ใช้ในการจำแนกคุณลักษณะมานั้นใช้วิธีการอิงตามความหมายและประเภทของคำในพจนานุกรมเป็นหลัก อีกทั้งคำที่ใช้ในการจำแนกคุณลักษณะได้ทำการเลือกเพียงคำวิเศษณ์ที่สามารถจำแนกความหมายออกได้เป็นคุณลักษณะเชิงบวกและคุณลักษณะเชิงลบเท่านั้น อีกทั้งข้อผิดพลาดจากการตัดคำ ซึ่งผลของการตัดคำนั้นอาจไม่สามารถให้ความและประเภทของคำคำนั้นได้อย่างถูกต้องตามความหมายและประเภทของพจนานุกรมได้

จากการรวบรวมข้อมูลจากสื่อออนไลน์ซึ่งมีลักษณะเป็นข้อความภาษาธรรมชาติ กล่าวคือ คำบางคำที่พบในเอกสารอาจไม่พบในพจนานุกรม ซึ่งคำดังกล่าวจะไม่สามารถแยกประเภทของคำได้ ถึงแม้ว่าคำดังกล่าวนั้นจะเป็นคำที่สามารถเป็นคำบ่งชี้คุณลักษณะได้ดีก็ตาม จึงไม่สามารถนำมาทดสอบในโมเดลดังกล่าวได้ด้วยเช่นกัน

วิจารณ์และสรุป

งานวิจัยฉบับนี้ เพื่อค้นหาเทคนิคการทำเหมืองข้อความที่มีประสิทธิภาพในการจำแนกข้อความคิดเห็นที่ได้มีการแสดงความคิดเห็นต่อการใช้บริการพร้อมเพย์ที่ถูกเขียนขึ้นด้วยภาษาไทย โดยข้อความคิดเห็นที่ได้นั้นได้ทำการเก็บรวบรวมผ่านการใช้งาน Facebook API จำนวนทั้งหมด 1,570 ข้อความ โดยในงานวิจัยนี้ได้ทำการแบ่งคุณลักษณะออกเป็น

2 กลุ่ม คือ คุณลักษณะเชิงบวกและคุณลักษณะเชิงลบ โดยนำเอาคำวิเศษณ์จากความคิดเห็นต่อการใช้บริการพร้อมเพย์ ซึ่งคำวิเศษณ์นี้จะสามารถแสดงถึงอารมณ์เชิงบวกและเชิงลบได้ดี[11] จากนั้นได้นำเอาเทคนิควิธีการวิเคราะห์เหมืองข้อความมาทำการวิเคราะห์ข้อความคิดเห็นโดยทำการเปรียบเทียบทั้งหมด 5 เทคนิควิธีด้วยกันคือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5

สำหรับขั้นตอนวิธีการวัดประสิทธิภาพของทั้ง 5 โมเดลได้ใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความแม่นยำและค่าความระลึก ผลจากการวิจัยพบว่า เทคนิค Naïve Bayes ให้ผลดีที่สุด โดยให้ค่าความถูกต้อง 93.88% ค่าความแม่นยำ 94.02% และค่าความระลึก 93.54% ผู้วิจัยจึงได้นำเอาหลักการของเทคนิควิธี Naïve Bayes มาประยุกต์ใช้ในการสร้างแบบจำลองเพื่อทำการจำแนกข้อความคิดเห็นภาษาไทย เพื่อให้สามารถพัฒนาใช้งานการจำแนกข้อความคิดเห็นภาษาไทยได้ดีมากยิ่งขึ้น

สำหรับงานวิจัยครั้งต่อไป อาจมีการพิจารณาถึงกระบวนการสร้างคำบ่งชี้จากประเภทของคำ เช่น คำนาม หรือคำกริยา เป็นต้น เพื่อให้เกิดความครอบคุลมในการวิเคราะห์ประโยคจากคำบ่งชี้ และเพื่อเป็นการเพิ่มประสิทธิภาพให้กับการจำแนกประเภทของประโยคและคำได้ดีมากขึ้น

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณอาจารย์คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามที่ให้คำปรึกษาและคำแนะนำต่าง ๆ เพื่อให้งานออกมาได้ดีและมีประสิทธิภาพมากที่สุด

เอกสารอ้างอิง

1. ธนาคารแห่งประเทศไทย. PromtPay 2559 [30 พฤศจิกายน 2559]. สืบค้นจาก: https://www.bot.or.th/Thai/Payment-Systems/PSServices/PromptPay.
2. Nikhil R, Nikhil Tikoo, Sukrit Kurle, Hari Sravan Pisupati, Dr Prasad G R. A. (2015). Survey on Text Mining and Sentiment Analysis for Unstructured Web Data. International Journal of Emerging Technologies and Innovative Research 2(5).
3. Ning Zhong, Yuefeng Li, Sheng-Tang Wu. (2012). Effective Pattern Discovery for Text Mining. IEEE Transactions on Knowledge and Data Engineering 24(1):30-40.
4. กานดา แผ่วัฒนากุล, ดร.ปราโมทย์ ลือนาม. การวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายสังคมออนไลน์. บทความวิชาการ วารสารการจัดการสมัยใหม่ ปีที่ 11. กรกฎาคม-ธันวาคม 2556;ฉบับที่ 2:10.
5. ประพัฒน์ พรมน้ำอ่าง, วสุวรรธน์ พงศ์ขจร, นิเวศ จิระวิชิตชัย. (2559). การจำแนกกลุ่มข้อความรีวิวโดยใช้เทคนิคเหมืองข้อมูล. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี 6(1), 94-102.
6. พัชรนิกานต์ พงษ์ธน, วรารัตน์ รุ่งวรวุฒิ, งามนิจ อาจอินทร์, สมจิตร อาจอินทร์. (2556). วิเคราะห์ความพึงพอใจของลูกค้าจากข้อความคำแนะนำโดยการทำเหมืองความคิดเห็น. Conference on Knowledge and Smart Technology 2012. p. 53-60.
7. อดิเทพ ไชยสาร, รัฐสิทธิ์ สุขะหุต. (2557). การประมาณอารมณ์จากความคิดเห็นภาษาไทยโดยใช้เทคนิคการเรียนรู้ของเครื่อง. การประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 9 (NCCIT 2013), มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพมหานคร, 9-10 พฤษภาคม 2557 p. 260-6.
8. P. Barnaghi, P. Ghaffari, J. G. Breslin. (2016). Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment. IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService); March 29 April 2016.
9. M. Bouazizi, T. Ohtsuki. (2015). Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis. International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 25-28 Aug. 2015.
10. Jiawei Han, Micheline Kamber, Jian Pei. (2012). Data Mining Concepts and Techniques. 3 ed.
11. Yan Sun, Changqin Quan, Xin Kang, Zuopeng Zhang, Fuji Ren. (2014). Customer emotion detection by emotion expression analysis on adverbs. Journal of special topics in Information Technology and Management15(4).