การจำแนกทัศนคติของผู้ซื้อแล็ปท็อปด้วยวิธีเทคนิคเหมืองข้อมูล

Sentiment Recognition of Laptop Reviewby Data Mining Techniques

จิรวรรณ เจริญสุข[1*] อานนท์ ผ่องรัศมีเพ็ญ[1] กานต์ธิดา ฟักนวม[1] และ เปรมปวีร์ ศรีจันทะวงษ์[1]

บทคัดย่อ

งานวิจัยนี้นำเสนอการจำแนกทัศนคติเกี่ยวกับคอมพิวเตอร์แล็ปท็อป ออกเป็น 2 ประเภทคือ ทัศนคติเชิงบวกและทัศนคติเชิงลบ ข้อมูลที่ใช้ในงานวิจัยนี้คือข้อมูลความคิดเห็นเกี่ยวกับคอมพิวเตอร์แล็ปท็อปซึ่งเป็นเอกสารภาษาอังกฤษจาก Semeval 2014 Task4 โดยทำทดสอบประสิทธิภาพการจำแนกทัศนคติของ 3 อัลกอริทึมได้แก่ ซัพพอร์ทเวกเตอร์แมชชีน เคเนียเรสเนอเบอร์และต้นไม้ตัดสินใจ โดยการแทนค่าคุณลักษณะทั้งหมด 2 วิธีคือ ค่าคะแนนจากพจนานุกรมซึ่งเป็นฐานข้อมูลที่ไม่เสียค่าใช้จ่ายและค่าความจริงจากข้อมูลที่ใช้เรียนรู้ จากการทดสอบได้ค่าประสิทธิภาพการจำแนกประสิทธิภาพดีที่สุดเมื่อใช้ค่าคะแนนคุณลักษณะที่ได้จากค่าคะแนนจากพจนานุกรมโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ ซึ่งได้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ 0.833 และค่าประสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ 0.733

คำสำคัญ: การจำแนกทัศนคติ ซัพพอร์ทเวกเตอร์แมชชีน เคเนียเรสเนเบอร์ ต้นไม้ตัดสินใจ แล็ปท็อป

Abstract

This research purposes sentiment recognition model by classifying into 2 two types which are positive and negative sentiment. The data set is computer laptop reviews in English text from SemEval 2014 task 4. There are 3 algorithms that are applied in our model including Support Vector Machine, K-Nearest Neighbors and Decision Tree. And, the main features of Feature Extraction process is composed of 2 features. There are Score of Senti-Wordnet dictionary that is a free database and Boolean values from training data. The experiment result shows Decision Tree gives higher than other algorithm. The F-measure of highest score of positive classification is 0.833 and the highest score of negative classification is 0.733

Keywords: Sentiment Recognition, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Laptop.

[1] คณะวิทยาศาสตร์ ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา ชลบุรี 20230

[1] Faculty of Science at Sri Racha , Kasersart University Sri Racha Campus., Chonburi 20230, Thailand

* Corresponding author. E-mail sfscijrc@src.ku.ac.th

บทนำ

ปัจจุบันการขายสินค้าต่าง ๆ ได้เปลี่ยนรูปแบบมาเป็นการให้บริการผ่านสื่อออนไลน์ทางอินเทอร์เน็ตในรูปแบบเว็บไซต์หรือโมบายแอพพลิเคชัน คอมพิวเตอร์แล็ปท็อปเป็นสินค้าอีกประเภทหนึ่งที่มีการซื้อขายกันบนเว็บไซต์โดยเว็บไซต์ดังกล่าวยังเปิดโอกาสให้ลูกค้าแสดงความคิดเห็นเกี่ยวกับสินค้าผ่านทางเว็บไซต์ความคิดเห็นเหล่านี้เป็นข้อมูลเบื้องต้นสำหรับผู้ที่กำลังตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์แล็ปท็อป นอกจากนั้นผู้ขายเองก็สามารถนำข้อมูลจากความคิดเห็นเหล่านั้นไปใช้ประโยชน์แต่ความคิดเห็นเหล่านั้นมีทั้งในแง่ดีและไม่ดีซึ่งการที่จะทราบได้ว่าเป็นความคิดเห็นแบบใดก็ต้องอ่านความคิดเห็นเหล่านั้นก่อนแต่หากสามารถทำการจำแนกความคิดเห็นออกเป็นทัศนคติของผู้แสดงความคิดเห็นว่าเป็นเชิงบวกหรือเชิงลบได้ จะช่วยให้ผู้อ่านสามารถเลือกอ่านความคิดเห็นได้สะดวกขึ้นและอ่านความคิดเห็นน้อยลง ดังนั้นงานวิจัยนี้จึงนำเสนอการจำแนกทัศนคติออกเป็น 2 ประเภทคือ ทัศนคติเชิงบวกและทัศนคติเชิงลบ ทดสอบประสิทธิภาพการจำแนกทัศนคติโดยใช้อัลกอริทึมเพื่อให้ได้วิธีการในการจำแนกทัศนคติที่มีประสิทธิภาพสูงสุดที่สามารถนำไปใช้ประโยชน์ในการจำแนกทัศนคติจากความคิดเห็นในอนาคตต่อไป

วิธีการศึกษา

เอกสารและงานวิจัยที่เกี่ยวข้อง

ทัศนคติ(Sentiment) หมายถึง ความรู้สึก อารมณ์หรือความคิดเห็น ส่วนการวิเคราะห์ทัศนคติ (Sentiment Analysis) หรือการทำเหมืองแร่ความคิดเห็น (D.K. Kirange ; Ratnadeep R. Deshmukh) หมายถึงการใช้การประมวลผลภาษาธรรมชาติ(Natural Language Processing) เพื่อการวิเคราะห์ทัศนคติโดยใช้ความรู้ทางภาษาศาสตร์และคอมพิวเตอร์ร่วมกัน เพื่อจำแนกประเภทของทัศนคติอารมณ์หรือความคิดเห็น

ฐิติมา เกษมศรีธนาวัฒน์ และ ธนัสนี เพียรตระกูล (2554) ทำการวิจัยเพื่อการจำแนกทัศนคติจากความคิดเห็นซึ่งเขียนเป็นภาษาอังกฤษ โดยเลือกความคิดเห็นของผู้ซื้อหนังสือหมวดการเขียนโปรแกรมด้วยภาษาจาวาจากเว็บไซต์ Amazon จำนวน 100 ความคิดเห็น ประกอบด้วย ทัศนคติที่ดีจำนวน57 ความคิดเห็นและทัศนคติที่ไม่ดีจำนวน43 ความคิดเห็น กระบวนการทำงานของงานวิจัยนี้เริ่มจากนำความคิดเห็นทั้งหมดมาจัดทำทำรายการศัพท์โดยกำจัดคำที่ไม่ต้องการออกจากรายการศัพท์ ได้แก่ +, -, 's', 've', re, article และตัวเลข แต่ละเว้นเครื่องหมาย ? และ ! ออกจากรายการศัพท์จากนั้นเปลี่ยนรูปคำในรายการศัพท์ในอยู่ในรูปเอกพจน์และนำมาแทนค่าคุณลักษณะเป็นค่าความจริง โดยค่า 0 หมายถึง ไม่พบในความคิดเห็น และ 1 หมายถึงพบในความคิดเห็นงานวิจัยนี้จำแนกประเภทของทัศนคติโดยใช้ 3 อัลกอริทึม ได้แก่Multi- Layer Perceptron, Naïve Bayes และ Decision Treeร่วมกับคัดเลือกและลดขนาดมิติของคุณลักษณะด้วยอัลกอริทึมรีลีฟ (Relief Algorithm)ผลการทดลองสรุปว่า Naïve Bayes ใช้ร่วม กับ Relief Algorithm ที่ 2.5% มีค่าความถูกต้อง 68% ดีที่สุด

นิเวศ จิระวิชิตชัยและนรินทร์ พนาวาส (2556)เสนองานวิจัยการจำแนกความคิดเห็น โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine learning) โดยได้นำเสนอแบบจำลองการจำแนกความคิดเห็นแบบอัตโนมัติ และเปรียบเทียบประสิทธิภาพในการจำแนกความคิดเห็นโดยใช้ 4 อัลกอริทึมคือ Support vector Machine, Naïve Bayes,Decision Tree และ K-Nearest Neighbor ซึ่งเป็นข้อมูลของผู้ชมภาพยนตร์ต่างประเทศจากชุดข้อมูลจำนวน 2,000ข้อมูลจากเว็บไซต์ Internet Movie Database (IMDB) สามารถสกัดคุณลักษณะขอข้อมูลได้ 1,166 คุณลักษณะ และใช้อัลกอริทึมInformation Gain ในการคัดเลือกลดขนาดมิติของคุณลักษณะคุณลักษณะโดยแทนค่าคุณลักษณะ 2 วิธีคือแทนค่าเป็นค่าความจริงและความถี่ของคำจากการทดลองปรากฏว่าการใช้อัลกอริทึมNaïve

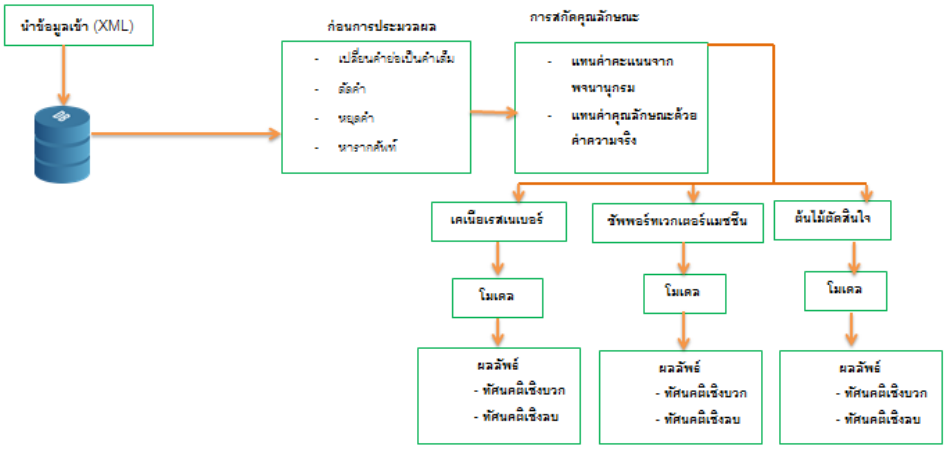
Bayes มีประสิทธิภาพดีที่สุด เมื่อลดคุณลักษณะเหลือ 200 คุณลักษณะ และ SVM มีประสิทธิภาพดีที่สุดเมื่อลดคุณลักษณะเหลือ 300 คุณลักษณะ

D.K. Kirange, Ratnadeep R. Deshmukh (2015) ทำการศึกษาวิเคราะห์ทัศนคติมุ่งเน้นไปที่การรับรู้และการจัดหมวดหมู่ของทัศนคติ4 ประเภท ได้แก่ เชิงบวก เชิงลบ ขัดแย้ง และเป็นกลางโดยเลือกข้อมูลความคิดเห็นจาก Semeval 2014 ซึ่งเป็นข้อมูลในทวิตเตอร์ แบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ข้อมูลการแสดงความคิดเห็นการใช้บริการของร้านอาหารและความคิดเห็นการเลือกซื้อ-ขายคอมพิวเตอร์แล็ปท็อปงานวิจัยนี้นำเสนอการจำแนกทัศนคติโดยมุ่งเน้นการใช้คำจากรายการศัพท์เป็นคีย์เวิร์ดในการหาการปรากฏของคำในเชิงบวกและเชิงลบจากพจนานุกรมชื่อ Linguistic Inquiry and Word Count (LIWC) ซึ่งเป็นพจนานุกรมที่มีการเสียค่าใช้จ่ายในการใช้งาน ซึ่งเปรียบเทียบประสิทธิภาพการจำแนกทัศนคติโดยใช้อัลกอริทึมSupport Vector Machineและ K-Nearest Neighbors ผลการทดลองของงานวิจัยนี้ปรากฏว่า Support Vector Machine มีประสิทธิภาพดีที่สุด

ดังนั้นงานวิจัยนี้จึงมีแนวคิดที่จะทดสอบประสิทธิภาพการจำแนกทัศนคติแบ่งเป็น 2 ประเภทคือ ทัศนคติเชิงบวกและ ทัศนคติเชิงลบ โดยเปรียบเทียบ3 อัลกอริทึมได้แก่Support vector Machine, K-Nearest Neighborsและ Decision Treeซึ่งเริ่มจากกระบวนการก่อนประมวลผลเพื่อจัดเตรียมข้อมูลให้อยู่ในรูปของคำศัพท์ที่ต้องการหลังจากนั้นจะสกัดคุณลักษณะโดยการเก็บเป็นเวกเตอร์ที่แทนค่าพจนานุกรมและค่าความจริง เพื่อช่วยกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของรายการศัพท์ กระบวนการคัดเลือกคุณสมบัติจึงเป็นขั้นตอนที่จำเป็นและงานวิจัยนี้จะเปรียบเทียบการจำแนกทัศนคติโดยมุ่งเน้นใช้ข้อมูลความรู้จากพจนานุกรมภาษาอังกฤษที่ไม่เสียค่าใช้จ่าย

ศึกษาปัญหาและวิเคราะห์

จากการศึกษางานวิจัยที่เกี่ยวข้องและได้ทำการวิเคราะห์ปัญหาพบว่าปัจจุบันมีหลายเว็บไซต์ที่มีผู้ขายสินค้าและบริการ ทำการขายสินค้าและบริการผ่านสื่ออินเทอร์เน็ตออนไลน์ เช่น การซื้อขายคอมพิวเตอร์แล็ปท็อปโดยให้ลูกค้าสามารถเข้ามาเขียนคำวิจารณ์สินค้าและแสดงความคิดเห็นของตนเองผ่านทางเว็บไซต์เพื่อใช้เป็นข้อมูลในการปรับปรุงสินค้า แต่เนื่องจากสินค้าที่มีผู้เข้ามาแสดงความคิดเห็นจำนวนมาก ทำให้ผู้ซื้อสินค้าหรือผู้ขายอ่านความคิดเห็นไม่ครบและใช้เวลานาน ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาการจำแนกทัศนคติของผู้ซื้อแล็ปท็อปโดยจะแบ่งการจำแนกทัศนคติออกเป็น 2ประเภทคือ ทัศนคติเชิงบวก ทัศนคติเชิงลบ ทั้งนี้ผู้วิจัยได้ทำการเปลี่ยนกิริยาช่วยจากคำย่อให้เป็นคำเต็มเพื่อทำการถอดความหมายของแต่ละคำ จะได้ความหมายที่ชัดเจนมากขึ้น ตัดคำตามช่องว่างเพื่อลดจำนวนคำในกระบวนการหยุดคำและเพื่อง่ายต่อการหารากศัพท์ หยุดคำเพื่อลดจำนวนคำที่ไม่บ่งบอกและมีความหมายต่อรูปประโยค และการหารากศัพท์ใช้ Porter Algorithm เพื่อลดจำนวนคำภายในถุงคำพจนานุกรมเพื่อคิดคะแนนเชิงบวกและคะแนนเชิงลบทำร่วมกับ 3 อัลกอริทึมได้แก่ Support vector Machine, K-Nearest Neighbors และDecision Treeแล้วนำเข้าประมวลผลใน โปรแกรม Weka เพื่อจะได้ค่า F-measure ของแต่ละประเภททัศนคติ



ภาพที่ 1 แผนภาพการทำงานของโปรแกรมการจำแนกความคิดเห็นของผู้ซื้อแล็ปท็อป

การทำงานของระบบแบ่งออกเป็น 3 ส่วน

1.**การนำข้อมูลเข้า**(Input)ข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลภาษาอังกฤษโดยข้อมูลดังกล่าวมาจาก Semeval 2015 Task 4 (D.K. Kirange)มีการแบ่งไฟล์ข้อมูลที่ใช้เรียนรู้ (Training data) และข้อมูลที่ใช้ทดสอบ (Testing Data) ไว้อย่างชัดเจนและจำนวนข้อมูลที่ใช้จะแสดงไว้ดังตารางที่1โดยที่ข้อมูลดังกล่าวถูกเก็บให้อยู่ในรูปแบบไฟล์ XML ตัวอย่างข้อมูล XML แสดงดังภาพที่ 2

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<sentences>
    <sentence id="2339">
        <text>I charge it at night and skip taking the cord with me because of the good battery life.</text>
        <aspectTerms>
            <aspectTerm term="cord" polarity="neutral" from="41" to="45"/>
            <aspectTerm term="battery life" polarity="positive" from="74" to="86"/>
        </aspectTerms>
    </sentence>
```

ภาพที่2 ตัวอย่างข้อมูลที่เก็บเป็นไฟล์ XML

ตารางที่ 1 ข้อมูลความคิดเห็นของแต่ละทัศนคติ

ประเภททัศนคติ	จำนวนของข้อมูลเรียนรู้	จำนวนของข้อมูลทดสอบ
เชิงบวก	580	24
เชิงลบ	589	15
รวม	1,169	39

2. **ข้อมูลก่อนประมวลผล**(Preprocessing) เป็นการนำข้อมูลให้เตรียมความพร้อมก่อนขั้นตอนการสกัดคุณลักษณะ ประกอบด้วยขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 2.1 เปลี่ยนกิริยาช่วย (Auxiliary Verb)จากคำย่อให้เป็นคำเต็มเพื่อทำการถอดความหมายของแต่ละคำจะได้ความหมายที่ชัดเจนมากขึ้น ตัวอย่าง กิริยาช่วยเช่น “it’s” เปลี่ยนเป็น “It is” “aren’t” เปลี่ยนเป็น “are not” และ “shouldn’t” เปลี่ยนเป็น “should not” เป็นต้น

ขั้นตอนที่ 2.2 การตัดคำ (Word Segmentation) เป็นการนำข้อความทั่วไปซึ่งอยู่ในรูปแบบประโยคมาแบ่ง ออกเป็นแต่ละคำ เช่น ประโยค “IT IS SO NICE TO LOOK AT AND THE KEYS ARE EASY TO TYPE WITH” ถูกตัดเป็น 14 คำ คือ ”IT”,“IS”, ”SO”,“NICE”,“TO”,“LOOK AT”,” AND”, “THE”, “KEYS”, “ARE”, “EASY”, “TO”, “TYPE”, “WITH“ ซึ่งในประโยคภาษาอังกฤษใช้เว้นวรรคในการตัดคำ

ขั้นตอนที่ 2.3 การกำจัดคำหยุด (Stop Word Removal) เป็นการกำจัดคำที่ไม่มีนัยสำคัญออกไป โดยไม่ทำให้ความหมายในเอกสารหรือข้อความนั้นเสียหรือเปลี่ยนแปลงไป โดยงานวิจัยนี้แบ่งประเภทคำหยุดคำได้แก่ คำสรรพนามคำบุพบท ซึ่งตัวอย่างคำหยุดแสดงดังตารางที่ 2

ตารางที่ 2 ประเภทของคำหยุด

ประเภทคำหยุด	ตัวอย่าง
คำสรรพนาม	it, he, she
คำบุพบท	at,to, for

ขั้นตอนที่ 2.4 การหารากศัพท์ของคำ (Stemming Word) เป็นการหารูปแบบดั้งเดิมหรือรากศัพท์ของคำนั้นๆ โดยปราศจากอุปสรรค (Prefixes) และปัจจัย (Suffixes) โดยส่วนใหญ่จะเป็นคำนามและคำกริยาเช่น คำว่า “walked” และ “walking” รูปแบบเดิมของคำ คือ “walk” แม้ว่า “ed” และ “ing” จะถูกตัดออกไปแต่ความหมายของคำๆ นั้นยังคงเดิม การหารากศัพท์จะช่วยรวมคำดังกล่าวให้เป็นคำเดียวกันเพื่อลดความซ้ำซ้อนของคำหรือคุณลักษณะที่เกิดขึ้นในเอกสารซึ่งงานวิจัยนี้ได้ประยุกต์ใช้ขั้นตอนวิธีการหารากศัพท์ของ Porter(1980) ซึ่งเหมาะสำหรับการหารากศัพท์ของคำที่เป็นภาษาอังกฤษ

รีวิว	กริยาช่วยคำเต็ม	ตัดคำ	หยุดคำ	หารากศัพท์
It's so nice to look at and the keys are easy to type with.	**It is** so nice to look at and the keys are easy to type with.	0 = It 1 = Is 2 = so 3 = nice 4 = to 5 = look 6 = at 7 = and 8 = the 9 = keys 10 = are 11 = easy 12 = to 13 = type 14 = with.	0 = is 1 = so 2 = nice 3 = look 4 = and 5 = the 6 = keys 7 = are 8 = easy 9 = type 10 = with	0 = is 1 = so 2 = nice 3 = look 4 = and 5 = the 6 = kei 7 = ar 8 = easi 9 = type 10 = with

ภาพที่ 3 ตัวอย่างข้อมูลก่อนประมวลผล

3. **การสกัดคุณลักษณะ (Feature Extraction)**คือการดึงคุณลักษณะ (Feature) ของทัศนคติออกมา ซึ่งการดึงคุณลักษณะออกมานั้นต้องกำหนดก่อนว่าใช้อะไรเป็นตัวแทนคุณลักษณะของทัศนคติ งานวิจัยนี้ใช้ค่าคะแนนจากพจนานุกรมและการแทนค่ารายการศัพท์ด้วยค่าความจริง (Boolean) โดยจะมีค่าเป็น 0 หรือ1 (ค่า 0 คือไม่พบความคิดเห็น และ 1 คือพบความคิดเห็น)ในส่วนนี้แบ่งการทำงานออกเป็น 2 ขั้นตอน ได้แก่

ขั้นตอนที่ 1 พจนานุกรมSenti-WordNetv3.0 เป็นฐานข้อมูลที่ไม่เสียค่าใช้จ่ายในการเรียกใช้งานประกอบด้วย คำ, คะแนนแต่ละคำให้คะแนน เชิงบวกและเชิงลบ โดยจะทำการคิดคะแนนของคำจากพจนานุกรม พจนานุกรมจะให้ค่าคะแนน 2 ค่าคือ Positive Score และ Negative Score ของแต่ละคำ

ขั้นตอนที่ 2 แทนค่าคุณลักษณะด้วยค่าความจริง (Boolean)คือการสร้างรายการศัพท์เป็นเวกเตอร์ให้เก็บคุณลักษณะด้วยเวกเตอร์และพจนานุกรม แทนด้วยค่าความจริงเชิงบวก ค่าความจริงเชิงลบ ค่าความจริงขัดแย้ง ค่าความจริงเป็นกลาง โดยจะมีค่าเป็น 0 หรือ1 (ค่า 0 ความหมายคือ ไม่พบความคิดเห็น และ 1ความหมายคือพบความคิดเห็น)

ตารางที่ 3 รายละเอียดปัจจัยที่ส่งผลต่อการจำแนกทัศนคติ

ลำดับ	ชื่อคุณลักษณะ	ค่าตัวแปร	จำนวนคุณลักษณะ	รายละเอียด
1	ค่าคะแนนเชิงบวก	ตัวเลขทศนิยม	49	ค่าคะแนนเชิงบวกของแต่ละคำที่ปรากฏในพจนานุกรม
2	ค่าคะแนนเชิงลบ	ตัวเลขทศนิยม	49	ค่าคะแนนเชิงลบของแต่ละคำที่ปรากฏในพจนานุกรม
3	ผลรวมของคะแนนเชิงบวก	ตัวเลขทศนิยม	1	ผลรวมของค่าคะแนนเชิงบวกทั้งหมดในพจนานุกรม
4	ผลรวมของคะแนนเชิงลบ	ตัวเลขทศนิยม	1	ผลรวมของค่าคะแนนเชิงลบทั้งหมดในพจนานุกรม
5	ผลต่างระหว่างของคะแนนเชิงบวกและคะแนนเชิงลบ	ตัวเลขทศนิยม	1	ผลต่าง = ผลรวมคะแนนเชิงบวก - ผลรวมคะแนนเชิงลบ
6	ผลรวมระหว่างของคะแนนเชิงบวกและคะแนนเชิงลบ	ตัวเลขทศนิยม	1	ผลต่าง = ผลรวมคะแนนเชิงบวก + ผลรวมคะแนนเชิงลบ
7	ค่าความจริงเชิงบวก	ค่าความจริง (0,1)	49	ค่าความจริงเชิงบวกของแต่ละคำที่ปรากฏในแต่ละความคิดเห็น
8	ค่าความจริงเชิงลบ	ค่าความจริง (0,1)	49	ค่าความจริงเชิงลบของแต่ละคำที่ปรากฏในแต่ละความคิดเห็น

เทคนิควิธีเหมืองข้อมูลในการจำแนกทัศนคติ

การทำเหมืองข้อมูล(Data mining) การทำเหมืองข้อมูล (Data Mining) คือกระบวนการกระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล(บุญเสริม กิจศิริกุล 2545) การทำเหมืองข้อมูล

(Data Mining) คือกระบวนการที่นำข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่ มาทำการศึกษา ทำความเข้าใจและนำผลลัพธ์ที่ได้จากการศึกษามาใช้ในการตัดสินใจ(Connolly and E.Begg 2002 : 1115)

ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)แนวคิดของซัพพอร์ทเวกเตอร์แมชชีนนำมาใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วนโดยสร้างเส้นแบ่งกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่ม Support Vector Machineจะใช้ฟังก์ชันแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Spaceใช้สำหรับข้อมูลที่มีมิติของข้อมูลกำหนดให้ $x_i, y_i, \ldots, x_n, y_n$เป็นตัวอย่างที่ใช้สำหรับการสอน n คือจำนวนข้อมูล m คือจำนวนมิติข้อมูลเข้า และ y คือผลลัพธ์ที่มีค่า +1 หรือ –1 (นิเวศ จิระวิชิตชัยและนรินทร์ พนาวาส ,2556)ดังสมการ

$$(x_i, y_i), \ldots, (x_n, y_n) \text{เมื่อ } x \in Rm, y \in \{+1, -1\} \quad (1)$$

ต้นไม้การตัดสินใจ (Decision tree)ต้นไม้ตัดสินใจเป็นแบบจำลองที่มีลักษะคล้ายกับแบบจำลองที่มีลำดับขั้นของการตัดสินใจ (Rokach,2008) เนื่องจากมีความซับซ้อนน้อยกว่าเมื่อเทียบกับอัลกอริทึมอื่นๆ ซึ่งต้นไม้ตัดสินใจเป็นการนำข้อมูลทดสอบ มาสร้างแบบจำลอง มีการทำงานแบบการเรียนรู้แบบมีผู้สอนการแสดงรูปแบบต้นไม้ตัดสินใจประกอบไปด้วย โหนด(Node) และแตกออกเป็นโหนดย่อยจนโหนดสุดท้ายเรียกว่า โหนดปลาย การวิจัยครั้งนี้ผู้วิจัยได้เลือกวิธีการสร้างต้นไม้ตัดสินใจ ด้วยขั้นตอนวิธีแบบ C4.5 (J.R Quinlan,1990) โดยกำหนดให้

$$Entropy(s) - \sum_{i=1}^{k} p_i log\ p_i \quad (2)$$

โดยที่ p_iจำนวนความถี่ของคลาสi ในโหนด sเพื่อใช้สำหรับคำนวณค่าความน่าจะเป็น ซึ่งจะเป็นหนึ่งคลาสผลลัพธ์ โดยค่า Entropy จะมีค่าเป็นนั้นหมายถึงทุกคลาสผลลัพธ์ค่าความน่าจะเป็นที่เท่ากันซึ่งมีโอกาสเกิดขึ้น 1 และมีค่าเป็น 0 ได้โดยนิยาม

$$p = k_i \mid N \quad (3)$$

ที่Nเท่ากับค่าทั้งหมดของรูปแบบกลุ่มของคลาส โดยk_iเท่ากับเหตุการณ์ที่เกิดขึ้นจริงใน Nการคำนวณค่า Information Grain ในการแบ่งกลุ่มpในกลุ่มkด้วยการวัดผล โดยนำค่า Gain ของpที่มีค่าน้อยที่สุดในการรวมกันจากกลุ่มย่อย k แล้วนำไปลบออกจากค่า $Entropy(p)$โดยค่าคุณสมบัติ (Attributes) จะใช้สำหรับการเลือกโหนด(Node) ในการแบ่งกลุ่มโดยเลือกจากค่า Gain ที่มีค่ามากที่สุดของ kโดยมีนิยาม(Tan,2006) ดังนี้

$$Gain = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \quad (4)$$

จากนั้นจึงทำการแจกแจงข้อมูลในแต่ละกลุ่ม pตามโหนด ที่ได้ทำการแบ่งไว้แล้วข้างต้น ตัวจำแนกข้อมูล C4.5 ได้ทำการขยายส่วนของข้อมูลที่เป็นตัวเลข ด้วยการแบ่งข้อมูลที่เป็นช่วง เพื่อใช้ในการสร้างต้นไม้ตัดสินใจ

เคเนียเรสเนเบอร์(K-Nearest Neighbour)คือ วิธีการในการจัดแบ่งคลาสผลลัพธ์โดยจะตัดสินใจ ว่าคลาสผลลัพธ์ใดที่จะแทนเงื่อนไขบ้าง โดยการตรวจสอบจำนวนบางจำนวน ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม จำนวนเงื่อนไข หรือกรณีต่างๆสำหรับแต่ละคลาสผลลัพธ์ และกำหนดเงื่อนไขใหม่ๆ ให้คลาสผลลัพธ์ที่เหมือนกันกับคลาสผลลัพธ์ที่ใกล้เคียงกันมากที่สุดในการนำเทคนิคของ K-NN ไปใช้ในนั้นเป็นการหา

วิธีการวัดระยะห่างระหว่างแต่ละคุณลักษณะในข้อมูลให้ได้ และจากนั้นคำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น หากเงื่อนไขการตัดสินใจมีความซับซ้อนวิธีนี้สามารถสร้างโมเดลที่มีประสิทธิภาพได้(ผศ.วิภาวรรณ บัวทอง)

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^{\ 2}} \qquad (5)$$

ผลการศึกษา

การจำแนกทัศนคติของผู้ซื้อแล็ปท็อปด้วยวิธีเทคนิคเหมืองข้อมูล ใช้ในการจำแนกทัศนคติแบ่งออกเป็น 2 ประเภท คือ ทัศนคติเชิงบวกและทัศนคติเชิงลบนอกจากนั้นยังได้ทำการเปรียบเทียบคุณลักษณะ8 คุณลักษณะ(ดังตารางที่3) โดยทดสอบประสิทธิภาพด้วยอัลกิริทึม Support vector Machine,K-Nearest Neighbors และ Decision Tree การคิดคะแนนค่าความจริงจากรายการศัพท์พบว่าอัลกอริทึม SVM ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.652 และอัลกอริทึมKNNให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ 0.526 การคิดคะแนนของค่าคะแนนจากพจนานุกรมพบว่าอัลกอริทึม J48ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.776 และให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ0.621 การคิดคะแนนจากพจนานุกรมพบว่าอัลกอริทึม J48 ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.833 และให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ0.733 การคิดคะแนนจากพจนานุกรมร่วมกับการคิดคะแนนของค่าคะแนนจากพจนานุกรมพบว่าอัลกอริทึมJ48 ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.760และอัลกอริทึม KNN ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ 0.625 การคิดคะแนนจากทุกคุณลักษณะที่ใช้พบว่าอัลกอริทึม SVM ให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.784 และให้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ0.593

ตารางที่ 4 สรุปผลการศึกษา

ประเภทคุณสมบัติ	F-Measurement						คุณสมบัติที่เลือกใช้
	SVM		DT(J48)		KNN		
	เชิงบวก	**เชิงลบ**	**เชิงบวก**	**เชิงลบ**	**เชิงบวก**	**เชิงลบ**	
1) ค่าความจริงจากรายการศัพท์	**0.652**	0.500	0.571	0.500	0.550	**0.526**	คุณลักษณะที่ 7-8
2) ผลรวมของค่าคะแนนจากพจนานุกรม	0.760	0.571	**0.776**	**0.621**	0.745	0.519	คุณลักษณะที่3-6
3) ค่าคะแนนจากพจนานุกรม	0.564	0.564	**0.833**	**0.733**	0.739	0.625	คุณลักษณะที่ 1-2
4) ผลรวมของค่าคะแนนจากพจนานุกรมร่วมกับค่าคะแนนจากพจนานุกรม	0.745	0.519	**0.760**	0.571	0.739	**0.625**	คุณลักษณะที่ 1-6
5) ทุกคุณลักษณะ	**0.784**	**0.593**	0.583	0.333	0.619	0.556	คุณลักษณะที่ 1-8

สรุป

Decision Tree ใช้อัลกอริทึม C5.0 ในการเลือกโหนดและแตกโหนดย่อยเพื่อใช้ในการจำแนกทัศนคติแต่ละแบบได้ดีจากค่าประสิทธิภาพ(F-Measurement) ที่ได้จาการทดสอบพบว่าใช้ค่าคุณลักษณะที่ได้จากค่าคะแนนจากพจนานุกรมโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ ซึ่งได้ค่าสิทธิภาพในการจำแนกทัศนคติเชิงบวกเท่ากับ0.833และค่าประสิทธิภาพในการจำแนกทัศนคติเชิงลบเท่ากับ 0.733โดยการใช้คุณลักษณะจากค่าความจริงพิจารณา การปรากฏของคำทำให้จำแนกประเภททัศนคติมีค่าสิทธิภาพต่ำกว่าการใช้คุณลักษณะจากค่าคะแนนจากพจนานุกรม เนื่องจากพจนานุกรมประกอบด้วยค่าของคะแนนเชิงบวกและเชิงลบของแต่ละคำซึ่งคะแนนของคำเหล่านี้สามารถใช้ในการจำแนกทัศนคติเชิงบวกและทัศนคติเชิงลบได้ดี

ข้อเสนอแนะ

จาการทดลองพบว่าพจนานุกรมเป็นปัจจัยหลักในการจำแนกทัศนคติ ดังนั้นจึงควรมีการเพิ่มชนิดของพจนานุกรมเข้ามาใช้ร่วมกับคุณลักษณะอื่นๆ นอกจากนี้ยังสามารถนำเอาข้อมูลทางไวยากรณ์ของคำหรือวลีเช่น หน้าที่ของคำ (Part of speech) การระบุขอบเขตของวลีการระบุนิพจน์ระบุนาม (Named Entities) การเพิ่มข้อมูลเชิงความหมายให้กับคำ เช่น การหาคำที่มีความหมายเหมือนกันแต่มีรูปแบบของคำไม่เหมือนกัน (synonym) การหาความสัมพันธ์ระหว่างคำศัพท์ จากฐานข้อมูล Word-Net เป็นต้น

เอกสารอ้างอิง

คมคิด ชัชราภรณ์, ธรา อังสกุลและจิติมนต์ อั่งสกุล.2555. ”แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้

ฐิติมา เกษมศรีธนาวัฒน์ ธนัสนี เพียรตระกูล 2554. การจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบเบย์ร่วมมือกับการเลือกคุณลักษณะด้วยอัลกอริทึมรีลีฟ. CIT2011&UniNOMS2011

เทคนิคการเรียนรู้ของเครื่อง” .สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยสุรนารี. 24 มกราคม 2555.

นิเวศ จิระวิชิตชัย นรินทร์ พนาวาส 2556. การจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง

D.K. Kirange; Ratnadeep R. Deshmukh, Ph.D. 2015. Emotion Classification of Restaurant and Laptop

“K-Nearest-Neighbors.”.(Online).Available.ออนไลน์ https://wipawanblog.files.wordpress.com/2014/06/chapter-6-k-nearest-neighbors.pdf

MF. Porter.1980. “An algorithm for suffix stripping” . Program.no3, PP 130-137, July 1980

Quin, J.R. (1990) Decision Trees and decision-making. IEEE Transaction on Systems, Man and Cybernetics, 20(2), 339-346

Review Dataset: Semeval 2014 Task4. International Journal of Computer Application (0975-8887) Volume 113-No.6, March 2015

Rokach,L, L. (2008) Data Mining with Decision Trees: Theory and Application. World Scientific

“SemEval 2014 Task 4.”(Online).Available.ออนไลน์http://alt.qcri.org/semeval2014/task4/

“SentiWordNet v3.0.”.(Online).Available.is based on WordNet version 3.0. WordNet website: (ออนไลน์).http://wordnet.princeton.edu/

“Weka”. (Online).Available.(ออนไลน์). http://www.cs.waikato.ac.nz/ml/weka/