การประยุกต์ใช้การวิเคราะห์ความหมายแฝงกับการจำแนกประเภทอารมณ์ในข้อความภาษาไทย

นางสาวปิยธิดา อินทร์รักษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFICATION OF EMOTIONS IN THAI TEXT

Ms. Piyatida Inrak

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2009

หัวข้อวิทยานิพนธ์	การประยุกต์ใช้การวิเคราะห์ความหมายแฝงกับการจำแนกประเภทอารมณ์ในข้อความภาษาไทย
โดย	นางสาวปิยธิดา อินทร์รักษ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.ชลวิช นัทธี)

ปิยธิดา อินทร์รักษ์ : การประยุกต์ใช้การวิเคราะห์ความหมายแฝงกับการจำแนกประเภทอารมณ์ในข้อความภาษาไทย. (APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFICATION OF EMOTIONS IN THAI TEXT) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร.สุกรี สินธุภิญโญ, 74 หน้า.

ในยุคที่การติดต่อสื่อสารข้อมูลผ่านเครือข่ายอินเทอร์เน็ตเติบโตขึ้นอย่างต่อเนื่อง ข้อมูลประเภทตัวอักษรก็ถูกผลิตขึ้นมาเป็นจำนวนมากเช่นกัน ข้อมูลเหล่านี้สามารถถูกถ่ายทอดออกมาและจำแนกหมวดหมู่ของตัวอักษรได้ การจำแนกด้านอารมณ์ก็เป็นอีกหัวข้อที่น่าสนใจในปัจจุบัน แต่การจำแนกด้านอารมณ์จากตัวอักษรภาษาไทยนั้นยังไม่มีประสิทธิภาพที่ดีพอ หัวข้อวิจัยนี้ได้แบ่งการจำแนกประเภทอารมณ์จากข้อความสั้นภาษาไทยออกมาเป็น ๖ อารมณ์สากลพื้นฐาน ได้แก่ โกรธ ขยะแขยง กลัว มีความสุข เศร้า และประหลาดใจ ซึ่งอ้างอิงจากข้อมูลการวิจัย ในการวิจัยนี้ได้เปรียบเทียบผลลัพธ์ของ ๒ ตัวแบบที่สร้างมาจากประโยครูปแบบต่างๆ และประยุกต์ใช้กับ ๓ ระเบียบวิธีได้แก่นาอีฟเบส์ (Naïve Bayes), เครื่องจักรเวกเตอร์สนับสนุน (Support Vector Machine, SVM) และต้นไม้ตัดสินใจ (Decision Tree) โดยตัวแบบที่หนึ่งใช้การจำแนกโดยการวิเคราะห์ความหมายแฝงของคำเดี่ยว ส่วนตัวแบบที่สองใช้การประยุกต์การวิเคราะห์ความหมายแฝงของคำคู่ที่มักปรากฏคู่กันร่วมกับระนาบความหมายของคำเดี่ยว ผลการเปรียบเทียบผลลัพธ์แสดงให้เห็นว่า ตัวแบบที่สองให้ความถูกต้องได้สูงกว่าตัวแบบที่หนึ่ง อ้างอิงจากระเบียบวิธีการจำแนกของนาอีฟเบส์ที่ให้ผลสูงกว่าระเบียบวิธีการอื่น

ภาควิชา .....วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต .....ปิยธิดา อินทร์รักษ์.....
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา ....2552....

5171427521 : MAJOR COMPUTER SCIENCE

KEYWORDS: EMOTIONS IN TEXT / LATENT SEMANTIC ANALYSIS / AFFECTIVE COMPUTING

PIYATIDA INRAK : APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFICATION OF EMOTIONS IN THAI TEXT. THESIS ADVISOR : ASST. PROF. SUKREE SINTHUPINYO, Ph.D., 74 pp.

With a rapid growth of the internet communication, many types of text are produced. They can convey the meanings that can contribute to text categorization. Moreover, emotion classification becomes more interesting, but emotion classification in Thai text is still not able to be correctly classified. Thus, this paper proposes a novel approach that takes advantage of bi-words occurrence to classify emotion hidden in a short sentence. In this paper, we classify Thai text into six basic universal emotions including anger, disgust, fear, happiness, sadness, and surprise based on Latent Semantic Analysis (LSA) approach. We compared the results between two models which construct features from the sentences and applied both models to three classification methods, i.e. Naïve Bayes, SVM, and Decision Tree. The first feature model uses only single word occurrence in the classification. The second model uses single word combined with bi-words occurrence in the classification. The results show that the second model yielded higher accuracy than the first model based on the Naïve Bayes classification method.

Department : Computer Engineering Student's Signature .....ปิยะโถ อินทร์อิก.....
Field of Study : Computer Science Advisor's Signature .....
Academic Year : ...2009....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษา ซึ่งท่านได้จุดประกายแนวคิดในการเริ่มต้นทำวิจัย ให้ความรู้ แนะนำแนวทางการวิจัย ตรวจสอบ ให้คำแนะนำ และสนับสนุนเป็นอย่างดี รวมไปถึงการให้กำลังใจ ผลักดันจนทำให้การวิจัยในครั้งนี้สำเร็จออกมาด้วยดี

ผู้วิจัยขอขอบพระคุณ ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล ผู้ช่วยศาสตราจารย์ ดร. ญาใจ ลิ่มปิยะกรณ์ และผู้ช่วยศาสตราจารย์ ดร. ชลวิช นัทธี กรรมการสอบวิทยานิพนธ์ ที่กรุณาเสียสละเวลา ให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้ ตลอดจนคณาจารย์ทุกท่านตั้งแต่อดีตจนถึงปัจจุบัน ที่ได้ประสิทธิ์ประสาทวิชาความรู้หลากหลายแขนงให้ผู้วิจัย

ผู้วิจัยขอขอบคุณเพื่อน ๆ ร่วมรุ่น CT19 ทุกคน ที่คอยให้กำลังใจ และให้ความช่วยเหลือในทุกรูปแบบ เพื่ออำนวยความสะดวกในระหว่างการทำวิจัย

ผู้วิจัยขอขอบคุณเพื่อนร่วมงานทุกคน และผู้บังคับบัญชาในสายงาน ที่คอยติดตาม ให้กำลังใจและสนับสนุน มีส่วนช่วยให้วิทยานิพนธ์สำเร็จได้ด้วยดี

สุดท้ายนี้ ผู้วิจัยขอขอบคุณคุณพ่อ คุณแม่ พี่สาว คุณป้า คุณตา และสมาชิกในครอบครัวทุกคน ที่คอยให้กำลังใจ สละเวลา และสนับสนุนในทุก ๆ เรื่อง และทุก ๆ วินาทีตั้งแต่ก้าวเข้าสู่รั้วมหาวิทยาลัย รวมถึงท่านอื่น ๆ ที่มิได้กล่าวชื่อไว้ ณ ที่นี้ที่มีส่วนช่วยให้วิทยานิพนธ์สำเร็จได้ด้วยดี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในทางจิตวิทยา อารมณ์ หมายถึง สภาพทางจิตใจ (Mental state) หรือ ความรู้สึก (Feeling) ที่เกิดขึ้นภายในจิตใจของแต่ละบุคคลตามเหตุการณ์และสิ่งแวดล้อม เช่น ความสนุกสนาน (Joy) ความเศร้า (Sorrow) ความนับถือ (Respect) ความเกลียดชัง (Hate) ความรัก (Love) [1] ซึ่งเป็นพฤติกรรมที่เกิดขึ้นเองโดยธรรมชาติ ทำให้เป็นสิ่งที่ยากในการจับต้องและวัดได้ด้วยการทดลอง [2] ต่อมาพบว่ามีงานวิจัยจำนวนไม่น้อยที่ให้ความสนใจเกี่ยวกับอารมณ์ ไม่ว่าจะเป็นสร้างพจนานุกรมความสัมพันธ์ของคำศัพท์กับอารมณ์ [1,3,4] หรือการจำแนกอารมณ์จากสื่อต่างๆรอบตัวเรา เช่น การจำแนกอารมณ์จากเพลง การจำแนกอารมณ์จากการเคลื่อนไหวของใบหน้า การจำแนกอารมณ์จากการแสดงท่าทางด้วยมือ เป็นต้น เนื่องจากไม่ว่าจะเป็นเสียง หรือเพลง [5] หรือลักษณะของใบหน้า หรือท่าทางของมือก็แล้วแต่ สามารถนำคุณสมบัติทางกายภาพได้แก่คลื่นเสียง ลักษณะใบหน้า ลักษณะมือ เหล่านี้มาหาความแตกต่างและจำแนกเป็นอารมณ์ได้ แต่เมื่อกล่าวถึงการจำแนกอารมณ์จากข้อความ กลับพบว่าเราไม่สามารถจำแนกได้ด้วยคุณลักษณะทางกายภาพ แต่จะต้องวิเคราะห์ไปถึงความหมายของข้อความที่ประกอบขึ้นเป็นคำหรือประโยคหรือบทความนั้น ซึ่งการหาความสัมพันธ์ของคำที่ปรากฏร่วมกันภายใต้ระนาบเดียวกันของความหมายเป็นวิธีการที่ดีที่สุดในการวิเคราะห์ความหมายแฝงของข้อความ [6]

ปัจจุบันข้อมูลบนอินเทอร์เน็ตมีมหาศาลและหลากหลาย การจัดกลุ่มข้อมูลเป็นสิ่งที่มีประโยชน์ ช่วยในการอ้างอิงหรือการค้นหาได้อย่างมีประสิทธิภาพ ทั้งนี้การจัดกลุ่มข้อมูลมีหลายรูปแบบตามแต่ความเหมาะสม ซึ่งสามารถนำไปใช้ประโยชน์ได้ไม่น้อย

งานวิจัยฉบับนี้ เลือกนำการวิเคราะห์ความหมายแฝงของคำมาประยุกต์ใช้ในการหาความสัมพันธ์ของคำที่ปรากฏคู่กัน (Co-Occurrence) ในข้อความ และวิเคราะห์เปรียบเทียบประสิทธิภาพของการจำแนกอารมณ์จากข้อความบนอินเทอร์เน็ต ทั้งนี้ผู้วิจัยเชื่อว่า คำเดี่ยวเพียงหนึ่งคำไม่สามารถบอกอารมณ์ของทั้งข้อความหรือบทความได้อย่างมีประสิทธิภาพ [2] เท่ากับการคำนึงถึงคำคู่ที่มักปรากฏคู่กันในข้อความ นอกจากนี้ความน่าสนใจในการประยุกต์ใช้วิธีการจำแนกประเภทอารมณ์ของข้อความภาษาไทยก็เป็นสิ่งที่ท้าทายเนื่องจากในปัจจุบันยังไม่มีงานวิจัยที่จำแนกประเภทอารมณ์ของข้อความภาษาไทย ไม่มีพจนานุกรมอารมณ์ที่ใช้อ้างอิงเหมือนกับภาษาอังกฤษหรือภาษาจีน อีกทั้งโครงสร้างภาษาไทยที่มีความละเอียดอ่อนแตกต่างจากภาษาอื่น ดังนั้นการที่สามารถสื่ออารมณ์จากข้อความได้นั้นจะเป็นประโยชน์กับการนำไป

ประยุกต์ใช้กับการสื่อสารเพื่อให้เกิดความสมจริงมากขึ้นด้วย เช่น ระบบการแปลงข้อความเป็นเสียง (Text to Speech) ในการเล่านิทานอัตโนมัติ สามารถจับคู่ข้อความที่ถูกกำกับว่ามีอารมณ์หนึ่ง กับโทนเสียงที่เกี่ยวข้องกับอารมณ์นั้นได้ เป็นต้น

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอการประยุกต์ใช้การวิเคราะห์ความหมายแฝงจากคำที่ปรากฏคู่กัน ในการจำแนกประเภทอารมณ์ในข้อความภาษาไทย เพื่อสื่อสารอารมณ์ของทั้งข้อความได้อย่างมีประสิทธิภาพใกล้เคียงกับความรู้สึกที่ได้จากผู้อ่าน

1.3 ขอบเขตของการวิจัย

1. ศึกษาวิธีการในการวิเคราะห์ความหมายแฝงของคำในข้อความภาษาไทย และนำไปประยุกต์ใช้เพื่อการปรับปรุงการจำแนกประเภทของข้อความภาษาไทยได้
2. เตรียมชุดข้อมูลเป็นข้อความภาษาไทยเพื่อการฝึกฝนและทดสอบระบบจำนวน 150 ข้อความ ซึ่งได้จากการเก็บตัวอย่างจากบล็อกและกระทู้ในอินเทอร์เน็ต โดยภาษาที่นำมาวิเคราะห์จะเป็นภาษาธรรมชาติ ไม่มีส่วนของไอคอนอารมณ์หรือการประกอบกันของตัวอักษรเป็นรูปหรืออารมณ์ใดๆ และไม่สนใจรูปแบบตัวอักษรที่ต่างกัน
3. การเตรียมชุดข้อมูลฝึกฝน ได้จากการเลือกโดยมนุษย์จากข้อความจำนวน 200 ข้อความ และกำหนดให้ผู้อ่านจำนวน 10 คน ระบุอารมณ์ที่เด่นชัดที่สุดของแต่ละข้อความ จากนั้นคัดเลือกข้อความที่ไม่แสดงอารมณ์ที่เด่นชัดออกไปให้เหลือเพียง 150 ข้อความ
4. การจำแนกประเภทอารมณ์ จะเปรียบเทียบประสิทธิภาพความถูกต้องในการจำแนกของอัลกอริทึม 3 แบบ ได้แก่ นาอีฟเบส์, เครื่องจักรเวกเตอร์สนับสนุน, และต้นไม้ตัดสินใจ ด้วยวิธีการทดสอบแบบไขว้ข้ามสิบพับ (10-fold cross validation)
5. อารมณ์ที่ใช้อ้างอิงเพื่อวิเคราะห์จำแนกอารมณ์จากข้อความภาษาไทยในงานวิจัยฉบับนี้ แบ่งเป็น 6 ประเภท ตามหลักการจำแนกอารมณ์สากล (Universal Emotions) ได้แก่ ขยะแขยง (Disgust) โกรธ (Anger) กลัว (Fear) เศร้า (Sadness) มีความสุข (Happiness) และ ประหลาดใจ (Surprise)
6. งานวิจัยนี้เป็นการทดลองเปรียบเทียบ 2 กระบวนการ ได้แก่ การวิเคราะห์ความหมายแฝงของคำเดี่ยว และการวิเคราะห์ความหมายแฝงของคำเดี่ยวร่วมกับคำที่มักปรากฏคู่กัน โดยอ้างอิงอารมณ์จากผู้อ่าน

7. การทดสอบประสิทธิภาพของตัวเรียนรู้ในระบบการจำแนกอารมณ์จากข้อความภาษาไทยด้วยวิธีการวิเคราะห์ความหมายแฝงของคำในข้อความ จะเปรียบเทียบความถูกต้อง (Accuracy) ในการจำแนกของอัลกอริทึม 3 แบบ ได้แก่ นาอีฟเบส์, เครื่องจักรเวกเตอร์สนับสนุน, และต้นไม้ตัดสินใจ (Decision Tree) ด้วยวิธีการทดสอบแบบไขว้ข้ามสิบพับ

1.4 แนวคิดและวิธีการดำเนินงาน

งานวิจัยนี้มีแนวคิดที่จะนำเสนอเปรียบเทียบตัวแบบวิธีการการวิเคราะห์ความหมายแฝงจากข้อความภาษาไทยของคำเดี่ยวกับการประยุกต์การวิเคราะห์ความหมายแฝงของคำคู่ที่มักปรากฏคู่กันร่วมกับระนาบความหมายของคำเดี่ยว ดังขั้นตอนต่อไปนี้

1. จัดเตรียมเอกสารข้อความภาษาไทยสำหรับการเรียนรู้ ที่เก็บรวบรวมได้จากบล็อกหรือกระทู้บนอินเทอร์เน็ต
2. ให้ผู้อ่าน อ่านเอกสารข้อความภาษาไทย และวิเคราะห์ให้ค่าอารมณ์ที่เหมาะสมกับข้อความนั้นมากที่สุด
3. นำข้อความมาตัดคำด้วยโปรแกรมการตัดคำภาษาไทย (SWATH) โดยเลือกวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบ ไบแกรม กับลำดับของหน้าที่ของคำ เพื่อคัดเฉพาะคำที่มีบทบาทต่ออารมณ์ซึ่งได้แก่ คำกริยา และ คำนามเท่านั้น
4. สร้างเมตริกซ์ความสัมพันธ์ของแต่ละคำกับชุดข้อความ
5. นำเมตริกซ์ที่ได้ผ่านกระบวนการการวิเคราะห์ความหมายแฝง
6. จำแนกประเภทอารมณ์ ด้วยการใช้วิธีการจำแนกประเภทเอกสาร 3 วิธี ได้แก่ วิธีนา-อีฟเบส์ วิธีเครื่องจักรเวกเตอร์สนับสนุน และวิธีต้นไม้ตัดสินใจ
7. เปรียบเทียบผลลัพธ์ ระหว่างตัวแบบทั้งสองด้วยการสร้างชุดข้อมูลฝึกฝนและทดสอบระบบด้วยวิธีการทดสอบแบบไขว้ข้ามสิบพับ
8. สรุปผลการทดลอง

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. นำเสนอวิธีการใหม่ในการประยุกต์การวิเคราะห์ความหมายแฝงของคำเดี่ยวที่ปรากฏในข้อความภาษาไทยเพื่อเพิ่มประสิทธิภาพการสื่อสารออกมาเป็นอารมณ์ของทั้งข้อความได้อย่างถูกต้องใกล้เคียงกับความรู้สึกที่ได้จากผู้อ่าน
2. สามารถนำวิธีการใหม่นี้ไปใช้อ้างอิง เพื่อการนำไปประยุกต์ใช้กับการวิเคราะห์ข้อความในรูปแบบอื่นๆ ได้ เช่น นวนิยาย บทความ เป็นต้น
3. สามารถนำไปพัฒนาใช้ร่วมกับระบบอื่นได้อย่างมีประสิทธิภาพ เช่น ระบบการแปลงข้อความเป็นเสียง (Text to Speech) ในการเล่านิทานภาษาไทย เพื่อการปรับโทนเสียงให้สอดคล้องกับอารมณ์ที่เกี่ยวข้องกับข้อความได้
4. สามารถนำไปใช้อ้างอิงหรือประยุกต์ใช้เพื่อการจัดลำดับ (Rating) หรือการจำแนกประเภทอารมณ์ของข้อความภาษาไทยบนโลกไซเบอร์ได้ เพื่อง่ายต่อการค้นหาและประโยชน์อื่นในอนาคตได้ เช่น การทำ RSS feed ข้อความบนบล็อกของเพื่อน + อารมณ์ได้ เป็นต้น

1.6 โครงสร้างของเนื้อหาในวิทยานิพนธ์

เนื้อหาในวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บท ดังนี้คือ บทที่ 1 บทนำ บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการทำงานวิจัยชิ้นนี้ บทที่ 3 กล่าวถึงการดำเนินงานวิจัยโดยอธิบายเป็นขั้นตอนต่างๆ อย่างละเอียด ส่วนในบทที่ 4 เป็นการทดลองและผลที่ได้จากการทดลอง และบทที่ 5 เป็นบทสรุปผลการทดลองและข้อเสนอแนะของงานวิจัยซึ่งจะเป็นประโยชน์ต่องานวิจัยอื่นๆ ในอนาคต

1.7 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของงานวิทยานิพนธ์นี้ ได้รับการตีพิมพ์เป็นบทความทางวิชาการ ดังนี้

“APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFY EMOTIONS IN THAI TEXT” โดย ปิยธิดา อินทร์รักษ์ และ สุกรี สินธุภิญโญ ในงานประชุมวิชาการระดับนานาชาติ “2010 The 2nd International Conference on Computer Engineering and Technology (ICCET 2010)” ซึ่งจัดขึ้น ณ เมืองเฉิงตู ประเทศสาธารณรัฐประชาชนจีน ระหว่างวันที่ 16-18 เมษายน พ.ศ. 2553 ดังภาคผนวก ก หน้า 58-63

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การวิเคราะห์ความหมายแฝงของคำ

การวิเคราะห์ความหมายแฝงของคำ เป็นหลักการพีชคณิตเชิงเส้นและวิธีการคำนวณทางสถิติที่มีประสิทธิภาพสูงเพื่อแยกดึงความหมายของคำจากข้อความหรือกลุ่มข้อความที่มีขนาดใหญ่ [7] การวิเคราะห์ความหมายแฝงของคำนี้จะแสดงให้เห็นถึงค่าความสัมพันธ์ระหว่างคำกับคำ คำกับเอกสารหรือข้อความ และเอกสารกับเอกสารในรูปของเมตริกซ์ที่มีมิติขนาดใหญ่ได้ประกอบด้วย 2 ขั้นตอนหลัก ดังนี้

2.1.1.1 อัลกอริทึมการแยกค่าแบบเดี่ยว (Singular Value Decomposition, SVD)

เริ่มจากการสร้างเมตริกซ์ A ที่มีค่าการปรากฏขึ้นของคำในแต่ละข้อความ ขนาด t x d ดังตารางที่ 1 โดยที่ แต่ละสดมภ์ (Column) แสดงเอกสารหรือชุดข้อความ (Documents) ที่นำมาวิเคราะห์ จำนวน d เอกสาร แต่ละแถว (Row) แสดงถึงคำ (Terms) แต่ละคำในข้อความไม่ซ้ำกัน (Unique word) จำนวน t คำ และแต่ละช่อง (Cell) แสดงความถี่ของคำ a(i, j) ที่ปรากฏในแต่ละชุดข้อความ

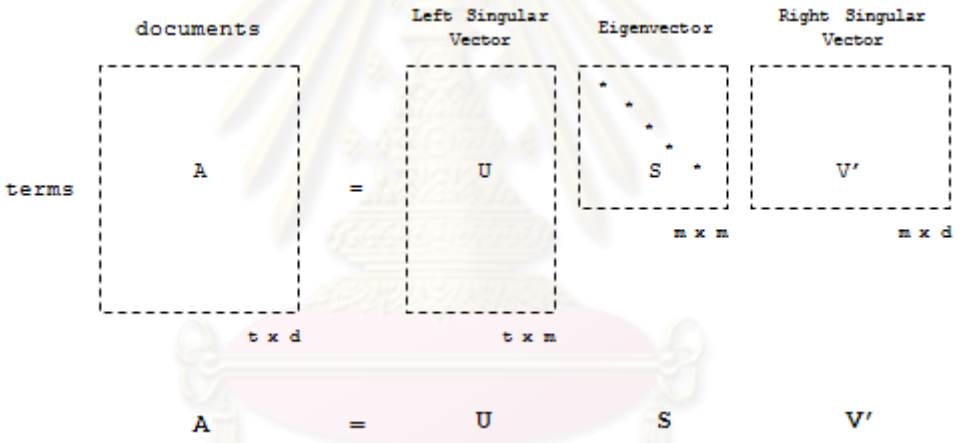
ตารางที่ 1เมตริกซ์ความสัมพันธ์ของคำกับชุดข้อความ

	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	...	ชุดข้อความ d
คำ 1	a(1,1)	a(2,1)	a(3,1)		a(m,1)
คำ 2	a(1,2)				…
คำ 3	a(1,3)				…
คำ 4	a(1,4)				…
คำ 5	a(1,5)				…
…					…
…					…
…					…
คำ t	a(1,n)	…	…	…	a(m,n)

จากนั้นนำเมตริกซ์ A มาผ่านอัลกอริทึมการแยกค่าแบบเดี่ยวที่ถูกนิยามในสมการที่ (1) จะได้เวกเตอร์เดี่ยวซ้าย (Left Singular Vector) เวกเตอร์เดี่ยวขวา (Right Singular Vector) และเวกเตอร์ไอเกน (Eigenvectors) ซึ่งเป็นเวกเตอร์แบบทแยงมุม [7]

$$A = USV^T \qquad (1)$$

อัลกอริทึมการแยกค่าแบบเดี่ยวประกอบด้วยการคำนวณหาค่าไอเกน (Eigenvalues) และเวกเตอร์ไอเกน ของ AA^T และ A^TA จะได้ว่าเวกเตอร์ไอเกนของ AA^T คือสดมภ์ของเมตริกซ์ U และเวกเตอร์ไอเกนของ A^TA คือสดมภ์ของเมตริกซ์ V และเวกเตอร์ค่าเดี่ยว S คือค่ารากที่ 2 ของค่าเดี่ยวจาก AA^T และ A^TA กล่าวคือ เมตริกซ์ S ซึ่งเป็นเมตริกซ์ทแยงมุมที่มีค่าเป็น $diag(s_1, s_2, ..., s_m)$ โดยที่ s_i ใดๆมีค่ามากกว่า 0 เสมอ สำหรับ $1 \le i \le r$ และ $s_j = 0$ สำหรับ $j \ge r+1$ โดยที่ $r \le \min(t,d)$ แสดงดังรูปที่ 1



รูปที่ 1 เมตริกซ์เมื่อผ่านอัลกอริทึมการแยกค่าแบบเดี่ยว

นอกจากนี้ $U^TU = V^TV = I_n$ และจะแสดงวิธีการคำนวณหาค่าเวกเตอร์ไอเกนโดยละเอียดได้ดังนี้

การคำนวณหาค่าเวกเตอร์ไอเกน (Computation of Eigenvectors)

หลักการของเวกเตอร์ไอเกนคือเมื่อนำเมตริกซ์ Q มาคูณแล้ว ผลลัพธ์ที่ได้ยังคงมีทิศทางเดิม [8] กล่าวคือ เวกเตอร์ไอเกน $X = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$ ซึ่งทำให้ $Q\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$ กล่าวคือ

$$(Q - \lambda I)\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} = 0 \qquad (2)$$

โดยที่ λ เป็นจำนวนจริงที่ไม่เท่ากับ 0 หรือเรียกว่าค่าเดี่ยว (Eigenvalue) ของเมตริกซ์ Q และสามารถหาได้จาก $\det(Q - \lambda I) = 0$ นั่นเอง หลังจากทราบค่า λ แล้ว จะสามารถหาเมตริกซ์ X ได้ดังตัวอย่างการคำนวณหาค่าเวกเตอร์เดี่ยวขวาของเมตริกซ์ A หรือเมตริกซ์ U

กำหนดให้ เมตริกซ์ $A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 1 & 4 \end{bmatrix}$

เนื่องจากเมตริกซ์ U คือเวกเตอร์เดี่ยวขวาของ AA^T จึงต้องหาค่า AA^T ก่อน จะได้เมตริกซ์จัตุรัส (Square Matrix) ที่มีขนาด $t \times t$ ในที่นี้กำหนดให้เป็นเมตริกซ์ Q

$$Q = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 1 & 4 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 8 & 9 \\ 8 & 13 & 14 \\ 9 & 14 & 17 \end{bmatrix}$$

จากนั้นหาค่า $\det(Q - \lambda I) = 0$ จะได้ว่า

$$\begin{vmatrix} 1-\lambda & 8 & 9 \\ 8 & 1-\lambda & 14 \\ 9 & 14 & 1-\lambda \end{vmatrix} = 0$$

ทราบค่า λ แล้ว จะสามารถหาเมตริกซ์ X ได้ และจากสมการที่ (2) จะได้เวกเตอร์ไอเกน X ที่มีขนาด $t \times 1$ ซึ่งสามารถเขียนได้ว่า $\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$ ซึ่งก็คือ เวกเตอร์เดี่ยวขวาของ AA^T นั่นเอง

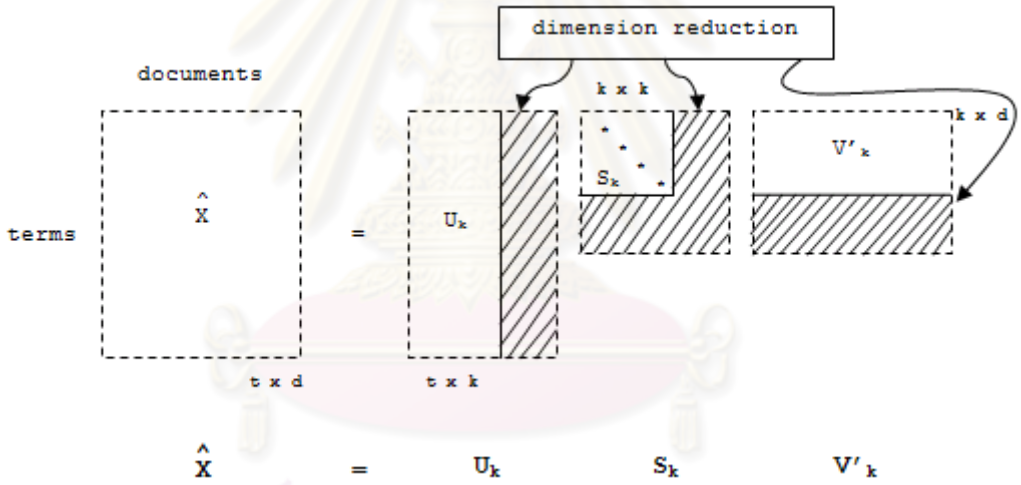
ในทำนองเดียวกันเราจะสามารถหาค่าเวกเตอร์เดี่ยวซ้ายของ $A^T A$ ได้เช่นกัน

2.1.1.2 การประมาณค่าการจัดอันดับ k (k-rank Approximations)

จากนั้นเป็นการลดขนาดมิติ (Dimension Reduction) โดยการเลือกขนาดของ k ที่เหมาะสม เพื่อเหลือไว้เพียงส่วนที่มีความหมายกับเอกสารนั้น [9] ในขั้นตอนนี้จะทำให้ได้ เมตริกซ์ของความสัมพันธ์ใหม่ที่มีขนาดที่เหมาะสม (The Least Square Best Fit) ดังสมการที่ (3) ซึ่งเมตริกซ์ใหม่ให้ชื่อว่า $\hat{X}$ นี้จะแสดงถึงระนาบที่พยากรณ์ความถี่ที่เหมาะสมของแต่ละคำที่มีแนวโน้มการเกิดขึ้นในแต่ละชุดข้อความได้ มีขนาดเท่ากับ k นั่นเอง

$$\hat{X} = \sum_{i=1}^{k} u_i \cdot s_i \cdot v_i^T \qquad (3)$$

จากนั้นนำเมตริกซ์ที่ผ่านการกำจัดสิ่งรบกวนทั้งสามมาคูณกัน เกิดเป็นเมตริกซ์ใหม่ ที่แสดงค่าความสัมพันธ์ระหว่างคำกับชุดข้อความอย่างมีนัย ดังรูปที่ 2



รูปที่ 2 เมตริกซ์ใหม่ที่ได้จากผลคูณของเมตริกซ์ที่ผ่านการลดขนาดมิติอย่างเหมาะสมเพื่อการกำจัดสิ่งรบกวน

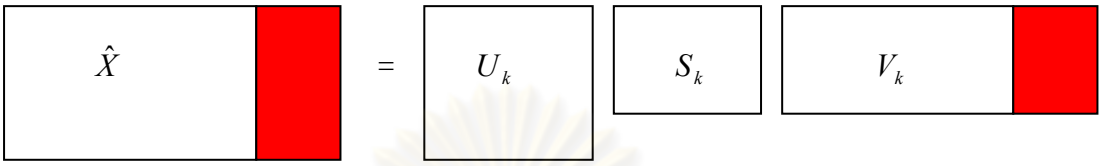
2.1.1.3 การค้นหา (Queries) ข้อมูลบนระนาบความหมาย

เมื่อเราได้เมตริกซ์ใหม่ $\hat{X}$ ซึ่งเป็นระนาบความหมาย (Semantic Space) ที่มีขนาดเหมาะสม k แล้ว สามารถแสดงค่าของชุดข้อความหนึ่งๆได้ในรูปของผลรวมของน้ำหนักของเวกเตอร์คำ หรือเวกเตอร์ $\hat{q}$ ดังสมการที่ (4) [7]

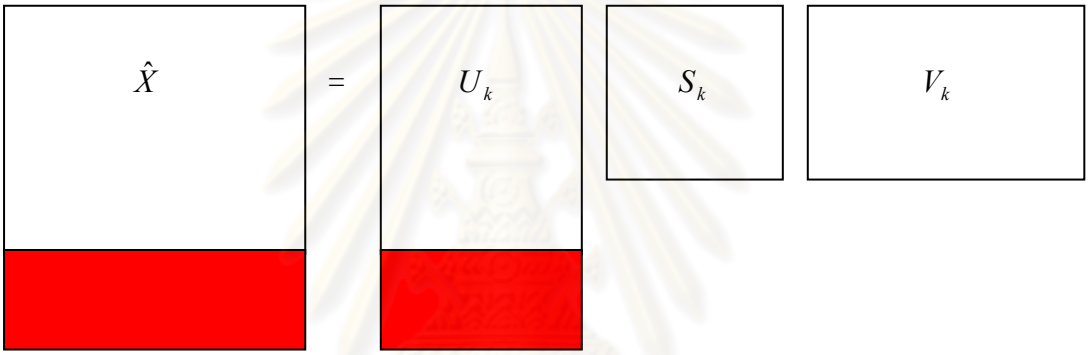
$$\hat{q} = q^T U_k S_k^{-1} \qquad (4)$$

2.1.1.4 การปรับปรุง (Updating) ข้อมูลบนระนาบความหมาย

การปรับปรุงเพิ่มเติมข้อมูล (Updating) บนระนาบความหมายได้ สามารถทำได้ 2 วิธี ได้แก่ การเพิ่มจำนวนชุดข้อมูล ดังรูปที่ 3 และการเพิ่มจำนวนคำ ดังรูปที่ 4 [7]



รูปที่ 3 การเพิ่มจำนวนชุดข้อมูลบนระนาบความหมาย



รูปที่ 4 การเพิ่มจำนวนคำบนระนาบความหมาย

ทั้งนี้จะต้องแปลงค่าให้อยู่ในรูปของผลรวมของน้ำหนักของเวกเตอร์ก่อนปรับปรุงข้อมูลบนระนาบความหมาย สมการที่ (5) แสดงการหาค่าผลรวมของน้ำหนักของเวกเตอร์ชุดข้อมูล

$$\hat{d} = d^T U_k S_k^{-1} \quad (5)$$

และสมการที่ (6) แสดงการหาค่าผลรวมของน้ำหนักของเวกเตอร์คำ

$$\hat{t} = tV_k S_k^{-1} \quad (6)$$

2.1.1.5 การแปลผล (Interpretation) จากระนาบความหมาย

นำเมตริกซ์ใหม่ที่ได้จากการทำอัลกอริทึมการแยกค่าแบบเดี่ยว มาแปลผลลัพธ์โดยการหาค่าความสัมพันธ์ต่างๆ ด้วยวิธีการจัดลำดับสหสัมพันธ์สัมประสิทธิ์ของสเปียแมน (Spearman's Rank Correlation Coefficient) [7] ซึ่งผลลัพธ์ที่ได้จะมีค่าอยู่ระหว่าง -1 และ +1 สามารถแปลผลได้ดังนี้

- เข้าใกล้ -1 : มีความสัมพันธ์ในทางตรงกันข้ามกัน (Negative correlation)
- เข้าใกล้ 0 : มีความสัมพันธ์กันน้อยมาก (No linear correlation)
- เข้าใกล้ +1 : มีความสัมพันธ์ไปในทางเดียวกัน (Positive correlation)

ทั้งนี้ ความสัมพันธ์สามารถหาค่าได้ 3 แบบ ดังนี้

แบบที่ 1 ความสัมพันธ์ระหว่างคำและคำ สามารถคำนวณหาได้จากระนาบความหมาย $\hat{X}\hat{X}^T$ [10] ตัวอย่างเช่น ความสัมพันธ์ระหว่างคำว่า “กลัว” และ “อกหัก” มีค่าเท่ากับ -0.70 ซึ่งมีค่าเข้าใกล้ -1 กล่าวคือ คำว่า “กลัว” และ “อกหัก” นี้มีความสัมพันธ์ในทางตรงกันข้ามกัน สามารถจำแนกเป็นคนละกลุ่มข้อมูลได้

หรือ ความสัมพันธ์ระหว่างคำว่า “ระทม” และ “อกหัก” ซึ่งมีค่าเท่ากับ 0.90 ซึ่งมีค่าเข้าใกล้ +1 กล่าวคือ คำว่า “ระทม” และ “อกหัก” นี้มีความสัมพันธ์ไปในทางเดียวกัน สามารถจำแนกเป็นคำในกลุ่มข้อมูลเดียวกันได้ ดังแสดงในตารางที่ 2 เป็นต้น

ตารางที่ 2 ตัวอย่างบางส่วนของผลลัพธ์จากการทดลองเพื่อแสดงความสัมพันธ์ระหว่างคำกับคำ

ความสัมพันธ์ระหว่างคำ (กลัว, เจ็บ)	0.40
ความสัมพันธ์ระหว่างคำ (กลัว, ระทม)	-0.40
ความสัมพันธ์ระหว่างคำ (กลัว, อกหัก)	-0.70
ความสัมพันธ์ระหว่างคำ (เจ็บ, ระทม)	0.30
ความสัมพันธ์ระหว่างคำ (รัก, อกหัก)	1.00
ความสัมพันธ์ระหว่างคำ (แพ้, ตาย)	0.38
ความสัมพันธ์ระหว่างคำ (เจ็บ, อกหัก)	-0.10
ความสัมพันธ์ระหว่างคำ (ระทม, อกหัก)	0.90

ในอีกมุมมองสำหรับความสัมพันธ์ระหว่างคำว่า “รัก” และ “อกหัก” ซึ่งมีค่าเท่ากับ 1.00 เข้าข่ายค่าเข้าใกล้ +1 เต็มๆ สามารถตีความได้ว่า เนื่องจาก คำว่า “รัก” และ “อกหัก” มักเกิดขึ้น

ในเอกสารเดียวกันจึงให้ค่าความสัมพันธ์ที่สูง แต่หากนำมาจำแนกอารมณ์ของแต่ละคำอาจจะพบว่า คำสองคำนี้ไม่สามารถทำให้เกิดเป็นอารมณ์ที่มีเปอร์เซนต์ความรู้สึกแบบเดียวกันได้

แบบที่ 2 ความสัมพันธ์ระหว่างข้อความและข้อความ สามารถคำนวณหาได้จากระนาบความหมาย $\hat{X}^T\hat{X}$ [10] ตัวอย่างเช่น ความสัมพันธ์ระหว่างข้อความ 2 และข้อความ 3 มีค่าเท่ากับ 0.96 ซึ่งมีค่าเข้าใกล้ +1 สามารถอธิบายได้ว่า ข้อความ 2 และข้อความ 3 นี้มีความสัมพันธ์ไปในทางเดียวกัน สามารถจำแนกเป็นข้อความในกลุ่มข้อมูลเดียวกันได้ ดังแสดงในตารางที่ 3

ตารางที่ 3 ตัวอย่างบางส่วนของผลลัพธ์จากการทดลองเพื่อแสดงความสัมพันธ์ระหว่างข้อความกับข้อความ

ความสัมพันธ์ระหว่างข้อความ (ข้อความ 1, ข้อความ 2)	0.85
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 1, ข้อความ 3)	0.74
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 1, ข้อความ 4)	0.75
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 1, ข้อความ 5)	-0.06
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 2, ข้อความ 3)	0.96
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 2, ข้อความ 4)	0.96
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 2, ข้อความ 5)	-0.41
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 3, ข้อความ 4)	0.99
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 3, ข้อความ 5)	-0.53
ความสัมพันธ์ระหว่างข้อความ (ข้อความ 4, ข้อความ 5)	-0.51

อีกตัวอย่างหนึ่งเช่น ความสัมพันธ์ระหว่างข้อความ 1 และข้อความ 5 มีค่าเท่ากับ -0.06 ซึ่งมีค่าเข้าใกล้ 0 กล่าวคือ มีความสัมพันธ์กันน้อยมาก ซึ่งถึงแม้ทั้งสองข้อความจะกล่าวถึง “เจ็บ” และ “แผล” เหมือนกัน แต่ก็ถูกตัดสินว่าเป็นข้อความที่ไม่เกี่ยวข้องกัน

อีกตัวอย่างหนึ่งเช่น ความสัมพันธ์ระหว่างข้อความ 3 และข้อความ 4 มีค่าเท่ากับ 0.99 ซึ่งมีค่าเข้าใกล้ +1 ทั้งสองชุดข้อความควรถูกจัดประเภทให้อยู่ในกลุ่มเดียวกัน แต่เมื่อเราอ่านเนื้อหาจริงๆแล้วกลับพบว่า ชุดข้อความทั้งสองไม่ได้เกี่ยวข้องกันเลย ให้ความรู้สึกที่ต่างกันด้วยซ้ำ

แบบที่ 3 ความสัมพันธ์ระหว่างคำและชุดข้อความ จากเมตริกซ์ใหม่ $\hat{X}$ ที่ได้จากการทำอัลกอริทึมการแยกค่าแบบเดี่ยว แต่ละจุด (Point) คือค่าความสัมพันธ์ระหว่างคำและชุดข้อความ ตัวอย่างเช่น ความสัมพันธ์ระหว่างคำและชุดข้อความ (รู้สึก, ชุดข้อความ 5) = 0.99 เป็นต้น

2.1.2 **ความถี่คำกับส่วนกลับเอกสาร** (Term Frequency – Inverse Document Frequency, TF-IDF)

การหาดัชนีความหมายแฝงของคำ จะต้องสร้างเมตริกซ์ $A = [a_{ij}]$ ซึ่งประกอบด้วยความถี่ของคำที่ปรากฏในแต่ละชุดข้อความ โดยที่ $a(i, j)$ เป็นค่าความถี่ของคำ i ที่ปรากฏขึ้นในชุดข้อความ j แต่เนื่องจากแต่ละคำไม่ได้ปรากฏขึ้นในทุกๆชุดข้อความทำให้ค่าที่ปรากฏในเมตริกซ์มีความห่างกันมาก ไม่หนาแน่น เราจึงคำนวณหาค่าน้ำหนักของคำนั้นกับทุกๆชุดข้อความด้วยวิธีการพิจารณาความถี่คำกับส่วนกลับเอกสาร

ดังนั้นค่าความถี่ในแต่ละช่องของเมตริกซ์ จะถูกแปลงโดยการให้ค่าน้ำหนักคำตามการประมาณค่าความสำคัญของคำนั้นกับเอกสาร ซึ่งสามารถหาได้จากการคำนวณด้วยวิธีการพิจารณาความถี่คำกับส่วนกลับเอกสาร ได้ค่าใหม่ที่แสดงถึงความสำคัญของคำนั้นที่มีต่อเอกสารในแต่ละสดมภ์ และการที่คำนั้นแสดงค่าในรูปทั่วไปเมื่อเทียบกับคำอื่นในเอกสารอื่นได้ [8] มีรายละเอียดดังนี้

TF คือ ความถี่ของแต่ละคำที่ปรากฏในแต่ละเอกสาร (Term Frequency) ดังสมการที่ (7)

$$TF = \frac{n}{N} \tag{7}$$

โดยที่

n คือ ความถี่ที่คำนั้นปรากฏในเอกสารหนึ่ง

N คือ จำนวนคำทั้งหมดในเอกสารหนึ่ง

IDF คือ ส่วนกลับความถี่ของเอกสาร (Inverse Document Frequency) ดังสมการที่ (8)

$$IDF = 1 + \log\frac{|D|}{DF} \tag{8}$$

โดยที่

$|D|$ คือ จำนวนเอกสารทั้งหมด

DF คือ จำนวนเอกสารที่มีคำนั้นปรากฏอยู่

กล่าวคือ การหาค่าความสำคัญของแต่ละคำที่ปรากฏ จะสามารถหาผลลัพธ์จากสมการ TF-IDF นี้ได้ด้วยการหาผลคูณของค่า TF และ IDF (TF x IDF) ดังสมการที่ (9)

$$(TF - IDF)_{i,j} = TF_{i,j} \times IDF_i \quad (9)$$

ค่าน้ำหนักที่ได้นี้มีแนวโน้มที่จะช่วยกรองคำที่ปรากฏทั่วไปในทุก ๆ ชุดข้อความออกไป เพื่อหาคำที่มีความสำคัญกับเอกสารนั้นไว้

2.1.3 การตัดคำในข้อความภาษาไทย

ภาษาไทยมีความยากในการตัดคำ เนื่องจากภาษาไทยไม่ระบุขอบเขตของคำ [12] มีการเขียนอย่างต่อเนื่อง ซึ่งในงานวิจัยฉบับนี้จะเลือกใช้โปรแกรม SWATH (Smart Word Analysis for THai) ซึ่งเป็นโปรแกรมการตัดคำภาษาไทยที่ได้รับความนิยมมากที่สุด ถูกพัฒนาขึ้นโดย S. Meknavin, P. Charoenpornsawat, และ B. Kijsirikul [13] เพื่อมุ่งหวังให้คอมพิวเตอร์สามารถประมวลผลภาษาไทยได้อย่างมีประสิทธิภาพ เนื่องจากภาษาไทยมีลักษณะการเขียนต่อกันเรื่อยๆ โดยไม่แบ่งแยกคำเหมือนกับการเขียนภาษาอังกฤษ ซึ่งถือเป็นอุปสรรคอย่างหนึ่งในการนำภาษาไทยไปใช้ในการวิเคราะห์ลักษณะอื่นต่อไป อัลกอริทึมที่นำเสนอในการตัดคำภาษาไทยมี 4 รูปแบบ ได้แก่ วิธีการตัดคำแบบยาวที่สุด (Longest Matching) วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) วิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบ ไบแกรม (Bigram) และ วิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบ ไบแกรมกับลำดับของหน้าที่ของคำ (Bigram with Part of Speech) นอกจากนี้ โปรแกรม SWATH ยังสามารถรองรับไฟล์อินพุตได้หลายแบบไม่ว่าจะเป็น utf8, text format, LaTex, RTF and HTML ตัวอย่างการตัดคำที่ทดลองใช้ทั้ง 4 แบบด้วยต้นแบบของข้อความเริ่มต้นเป็น “หัวใจเปียกปอน”

ผลลัพธ์แบบที่ 1 ด้วยวิธีการตัดคำแบบยาวที่สุด จะค้นหาคำโดยเริ่มจากตัวอักษรซ้ายสุดของข้อความไปยังตัวอักษรถัดไปจนกว่าจะพบคำที่มีอยู่ในพจนานุกรมภาษาไทย และค้นหาคำถัดไปจนกว่าจะจบข้อความ จะได้ หัวใจ | เปีย | ก | ปอน |

ผลลัพธ์แบบที่ 2 ด้วยวิธีการตัดคำแบบสอดคล้องมากที่สุด จะได้ หัวใจ | เปีย | ก | ปอน |

ผลลัพธ์แบบที่ 3 ด้วยวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบไบแกรม จะได้ หัวใจ | เปียก | ปอน |

ผลลัพธ์แบบที่ 4 ด้วยวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบไบแกรม กับลำดับของหน้าที่ของคำ ซึ่งเป็นวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของ

คำร่วมกับลำดับของหน้าที่ของคำในประโยค มาช่วยในการคำนวณหาความน่าจะเป็นที่จะเกิดเป็นคำที่มีความหมายและถูกต้องตามหน้าที่ของคำในประโยค จะได้ หัวใจ@NCMN | เปียก@VSTA | ปอน@ADVN | ทั้งนี้ตัวย่อที่แสดงหน้าที่ของคำนี้สามารถดูความหมายได้จากภาคผนวก ข หน้า 64-65

2.1.4 **การจำแนกประเภทของข้อมูล** (Data Classification)

การจำแนกประเภทของข้อมูล เป็นกระบวนการสร้างตัวแบบจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นความแตกต่างระหว่างคลาสหรือกลุ่มของข้อมูลได้ และเพื่อทำนายว่าข้อมูลนี้ ควรจัดอยู่ในคลาสใด ซึ่งโมเดลที่ใช้จำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ จะขึ้นอยู่กับการวิเคราะห์เซตของข้อมูลฝึกฝนโดยนำชุดข้อมูลฝึกฝนมาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในคลาสเดียวกันบ้าง [11]

ผลลัพธ์ที่ได้จากการเรียนรู้ คือ ตัวแบบจัดประเภทข้อมูล ซึ่งตัวแบบนี้สามารถแทนได้ในหลายรูปแบบ เช่น การจำแนกด้วยกฎถ้า-แล้ว (IF-THEN rules) ต้นไม้ตัดสินใจ (Decision Tree) หรือ นิวรอลเน็ตเวิร์ก (Neural networks) เป็นต้น และจะนำข้อมูลส่วนที่เหลือจากการฝึกฝนเป็นข้อมูลที่ใช้ทดสอบ ซึ่งเป็นกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากตัวแบบเพื่อทดสอบความถูกต้อง โดยเราจะปรับปรุงตัวแบบจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลผ่านตัวแบบ โดยตัวแบบจะสามารถทำนายกลุ่มของข้อมูลนี้ได้

2.1.4.1 **นาอีฟเบส์** (Naïve Bayes)

นาอีฟเบย์ เป็นเทคนิคที่ใช้ทฤษฎีเบย์ (Bayes' Theorem) ในการคำนวณหาค่าความน่าจะเป็น เพื่อการทำนายผล ซึ่งทำได้โดยการรวมผลของตัวแปรอิสระที่มีผลต่อตัวแปรตาม ซึ่งเทคนิคนี้ถูกนำมาใช้ในการจำแนกประเภท (Classification) [11]

จากทฤษฎีเบย์ กำหนดให้ P(H) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H และ P(H|E) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E

ดังนั้น จากตัวแปรที่กำหนดและแนวคิดของทฤษฎีเบย์ เราสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆได้ดังสมการที่ (10)

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)} \qquad (10)$$

และเราจะสามารถแสดงการคำนวณการจำแนกประเภทของเหตุการณ์ที่มีการเกิดของเหตุการณ์ต่างๆที่ใช้ในการจำแนกประเภทมากกว่า 1 ชนิด ได้ดังสมการที่ (11)

$$P(H \mid E1, E2, ..., En) = \frac{P(E1, E2, ..., En \mid H) \times P(H)}{P(E1, E2, ..., En)} \qquad (11)$$

ตัวอย่าง การคำนวณหาความน่าจะเป็น P(Emotion=Sadness|Word) ได้ดังต่อไปนี้

P(Emotion=Sadness|Word) = P(Word “เจ็บ” | Sadness) x P(Word “น่าเป็นห่วง” | Sadness) x P(Word “ไม่เหลียวแล” | Sadness) x P(Word “ระทม” | Sadness) x P(Word “เหยียบย่ำ” | Sadness) x P(Word “อกหัก” | Sadness) x P(Word “เหวอะหวะ” | Sadness) x P(Word “วิวาห์” | Sadness) x P(Word “หวาน” | Sadness) x P(Sadness) x 1/P(Word)

ตัวอย่าง การคำนวณหาความน่าจะเป็น P(Emotion=Happiness|Word) ได้ดังต่อไปนี้

P(Emotion=Happiness|Word) = P(Word “เจ็บ” | Happiness) x P(Word “น่าเป็นห่วง” | Happiness) x P(Word “ไม่เหลียวแล” | Happiness) x P(Word “ระทม” | Happiness) x P(Word “เหยียบย่ำ” | Happiness) x P(Word “อกหัก” | Happiness) x P(Word “เหวอะหวะ” | Happiness) x P(Word “วิวาห์” | Happiness) x P(Word “หวาน” | Happiness) x P(Happiness) x 1/P(Word)

เมื่อกำหนดให้เหตุการณ์ E1, E2,....,En คือเหตุการณ์ n เหตุการณ์ที่ใช้ในการจำแนกประเภท และจากสมมติฐานที่เรากำหนดให้แต่ละเหตุการณ์ต่างๆที่ใช้ในการจำแนกประเภทเป็นอิสระต่อกันแล้วนั้น เราจะสามารถแสดงการคำนวณโดยใช้ทฤษฎีเบย์ได้ดังสมการที่ (12)

$$P(E1, E2, ..., En \mid H) = \frac{P(E1 \mid H) \times P(E2 \mid H) \times ... \times P(En \mid H) \times P(H)}{P(E1) \times P(E2) \times P(E3) \times ... \times P(En)} \qquad (12)$$

หากนำมาประยุกต์ใช้กับคำสองคำที่เป็นอิสระต่อกันของการปรากฏคู่กันในประโยคหนึ่ง ทำให้เกิดเป็นอารมณ์หนึ่ง จะแสดงได้ดังสมการที่ (13)

$$P(E1, E2 \mid H) = \frac{P(E1 \mid H) P(E2 \mid H) \times P(H)}{P(E1) P(E2)} \qquad (13)$$

กล่าวคือ ความน่าจะเป็นของคำว่า “เจ็บ” ที่อาจปรากฏคู่กันกับคำว่า “อกหัก” ในประโยคหนึ่งแล้วส่งผลให้ประโยคมีอารมณ์ “เศร้า” คือผลคูณของค่าความน่าจะเป็นของคำว่า “เจ็บ” ที่ทำให้เกิดอารมณ์ “เศร้า” กับผลคูณของค่าความน่าจะเป็นของคำว่า “อกหัก” ที่ทำให้เกิดอารมณ์ “เศร้า” และค่าความน่าจะเป็นของอารมณ์ “เศร้า” ที่เกิดขึ้นในชุดข้อมูล โดยที่ คำว่า “เจ็บ” และคำว่า “อกหัก” เป็นอิสระต่อกัน

P(Word “เจ็บ”, Word “อกหัก” | Sadness) = (P(Word “เจ็บ” | Sadness) x P(Word “อกหัก” | Sadness) x P(Sadness)) / (P(Word “เจ็บ”) x P(Word “อกหัก”))

ในขั้นตอนการจำแนกประเภทว่าในสถานการณ์ที่ยกตัวอย่างมานั้นเราจะพิจารณาว่าเอกสารหนึ่งมีค่าความน่าจะเป็นค่าใดมีค่ามากที่สุดเราก็จะกำหนดให้เอกสารนั้นอยู่ในกลุ่มนั้น

2.1.4.2 **ต้นไม้ตัดสินใจ** (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) เป็นวิธีหนึ่งที่สำคัญในจำแนกประเภท โดยแผนภาพต้นไม้จะมีลักษณะเป็นแผนภาพการไหล (Flow-Chart) เหมือนโครงสร้างต้นไม้ ที่แต่ละจุดตัดสินใจ (Decision Node) แสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งแสดงผลในการทดสอบและจุดใบ (Leaf Node) แสดงกลุ่มหรือคลาสที่กำหนดไว้ ซึ่งต้นไม้ตัดสินใจนี้ง่ายต่อการปรับเปลี่ยนเป็นกฎการจำแนก เมื่อมีข้อมูลที่ต้องการที่จะจัดกลุ่มก็จะนำคุณลักษณะต่างๆของข้อมูลนั้นไปเทียบกับต้นไม้ตัดสินใจนี้ตามเส้นทางในต้นไม้จนกระทั่งได้คลาสปลายทาง ซึ่งก็คือกลุ่มของข้อมูลที่เหมือนกัน

ข้อดีของวิธีการนี้คือ สามารถตีความและเข้าใจลักษณะของรูปแบบข้อมูล (Pattern) ได้ง่าย เพราะ มีการแยกออกเป็นกฎ หรือข้อกำหนดต่างๆ ข้อเสียคือเรื่องของการให้น้ำหนักความน่าเชื่อถือหรือการให้ค่าน้ำหนักในแต่ละจุด ซึ่งถ้าให้น้ำหนักผิดไป อาจจะทำให้การตีความผิดไปได้

2.1.4.3 **เครื่องจักรเวกเตอร์สนับสนุน** (Support Vector Machine)

เครื่องจักรเวกเตอร์สนับสนุน เป็นวิธีการจำแนกประเภทข้อมูลด้วยการสร้างสมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 คลาสออกจากกัน โดยการย้ายข้อมูลจากระนาบอินพุตไปยังระนาบคุณลักษณะ และสร้างฟังก์ชันวัดความคล้ายกันของข้อมูลที่เรียกว่าฟังก์ชันเคอร์เนล (Kernel Function) บนระนาบคุณลักษณะ ตัวจำแนกที่เหมาะที่สุด (Optimal) ควรจะมีอยู่ห่างจากข้อมูลที่อยู่ใกล้ที่สุดจากทั้ง 2 คลาสด้วยระยะทางที่มากที่สุด

ข้อดีของวิธีการนี้คือ สามารถรองรับจำนวนคุณลักษณะได้มาก และมีความถูกต้องสูง ข้อเสียคือ ต้องเลือกฟังก์ชันเคอร์เนลที่เหมาะสม

การประยุกต์ใช้เครื่องจักรเวกเตอร์สนับสนุนแบบหลายคลาส (Multiclass SVM) ในการจำแนกข้อมูลที่ไม่ได้จำกัดเพียง 2 คลาส สามารถทำได้ 3 แบบ

(1) ใช้ k หนึ่งตัวแยกแยะส่วนที่เหลือ (k One-to-Rest Classifiers)

(2) ใช้ k(k-1)/2 แยกแยะเป็นคู่ (k(k-1)/2 Pairwise Classifiers) กับการออกเสียง ดังนี้

- ลงคะแนนเสียงส่วนใหญ่ (Majority Voting)
- การเทียบเคียงเข้าคู่ (Coupling Pairwise)

(3) ขยายสูตรเครื่องจักรเวกเตอร์สนับสนุนเพื่อรองรับปัญหาหลายคลาส

- สร้างฟังก์ชันการตัดสินใจโดยพิจารณาจากทุกคลาสทั้งหมดในครั้งเดียว
- สร้างฟังก์ชันการตัดสินใจสำหรับแต่ละคลาสโดยพิจารณาเฉพาะจุดข้อมูลการฝึกฝนที่อยู่ในคลาสนั้น

2.1.5 **การจำแนกประเภทอารมณ์** (Emotion Classification)

การเลียนแบบความสามารถในการแสดงอารมณ์ของมนุษย์เป็นสิ่งที่ยาก แต่สามารถทำได้โดยเลือกสิ่งที่เป็นพื้นฐานทางอารมณ์ที่เหมาะสมกับความมุ่งหมาย เช่น การรู้จำ (Recognition) ความเข้าใจ (Understanding) และการจำลอง (Simulation) และปรับปรุงกระบวนการเพื่อให้การติดต่อสื่อสารกันระหว่างมนุษย์กับคอมพิวเตอร์มีประสิทธิภาพดีขึ้น [1]

(1) การวิเคราะห์อารมณ์จากข้อความมีรูปแบบทั่วไปที่น่าสนใจ [14] ได้แก่

- การวิเคราะห์ด้วยการใช้ความรู้สึก (Sentiment Analysis)
- การใช้คอมพิวเตอร์ช่วยในด้านความคิดสร้างสรรค์ (Computer Assisted Creativity) และ
- การวิเคราะห์จากคำกริยาของประโยคที่มีผลต่อความรู้สึกในการแสดงอารมณ์ในการปฏิสัมพันธ์ระหว่างคอมพิวเตอร์กับมนุษย์ (Verbal Expressivity in Human Computer Interaction) ซึ่งวิธีนี้เน้นการดึงอารมณ์มาจากภาษาธรรมชาติของมนุษย์ให้มีประสิทธิภาพมากขึ้น เนื่องจากคำกริยาสามารถบ่งบอกถึงระดับความเป็นไปได้ในการรู้จำอารมณ์ของมนุษย์ และสามารถที่จะบอกเป็นนัยให้รู้ว่าประโยคดังกล่าวแสดงอารมณ์อย่างไร

(2) จากวิธีการจำแนกอารมณ์ข้างต้นสามารถแบ่งประเภทอารมณ์ได้ ดังนี้

- แบ่งแบบหยาบ (course-grained) เช่น แบ่งเป็นขั้วบวก-ขั้วลบ
- แบ่งแบบละเอียด (fine-grained) ซึ่งแบ่งเป็นอารมณ์ต่างๆ [15]

- แบ่งทั้งสองแบบโดยเลือกแบ่งเป็นขั้วก่อนแล้วจึงแบ่งเป็นอารมณ์ที่มีความหมายไปในทางเดียวกับขั้วนั้น ซึ่งพบว่าให้ผลดีกว่า ทำให้ผลการจำแนกที่ได้มีความชัดเจนกว่า [16]

ทั้งนี้ในการศึกษาเกี่ยวกับข้อความที่มีผลต่อความรู้สึก (Affective Text) มักจะกล่าวถึงการตกลงกันด้วยเงื่อนไขการวัดความสัมพันธ์ของคำแบบตัววัดสหสัมพันธ์ของเพียร์ซัน (Pearson Correlation Measure) ซึ่งประกอบด้วย 6 อารมณ์หลัก ได้แก่ โกรธ (Anger) ขยะแขยง (Disgust) กลัว (Fear) สนุกสนาน (Joy) เศร้า (Sadness) และ ประหลาดใจ (Surprise) ซึ่งเชื่อว่าน่าจะบอกอารมณ์ที่ดีที่สุดในการจำแนกข้อความ [14,17] และเป็นเหตุผลให้การเลือกอารมณ์เหล่านี้มาใช้วิเคราะห์ในงานวิจัยอารมณ์จากข้อความภาษาไทยในงานวิจัยฉบับนี้

2.1.6 **การจำแนกประเภทของข้อความ** (Text Classification)

ในการจำแนกประเภทของข้อความ มีปัจจัยแวดล้อมได้หลายรูปแบบที่เอื้อประโยชน์ในการวิเคราะห์จำแนก ได้แก่

(1) ความยาวของข้อความ ได้แก่ ระดับคำ [18] ระดับประโยค [16] ระดับข้อความสั้น [2] หรือ ระดับบทความยาว [2,19] เป็นต้น

(2) ลักษณะของแหล่งที่มาของข้อความ เช่น หนังสือนิทานเด็ก [20] ข่าว หัวข้อข่าว [4] ไดอารี่/บล็อก [2,3,16,18] กระทู้ ห้องสนทนา (chat environment) [21] นวนิยาย [19,22] เป็นต้น

การจำแนกประเภทของข้อความ จะต้องประกอบด้วย อินพุต คือชุดข้อความ X และเอาต์พุตเป็นคลาส Y ที่ต้องการทำนาย ตัวอย่างเช่น การจำแนกประเภทข้อความตลก (Jokes) เป็นข้อความที่มีความตลก (Funny) และไม่ตลก (Not Funny) การจำแนกประเภทรายงานของนักศึกษาให้อยู่ในคลาส A, B, C, D, หรือ F การจำแนกประเภทอีเมลว่าเป็นสแปม (spam) หรือไม่เป็นสแปม เป็นต้น [23]

จากการศึกษางานวิจัยที่ผ่านมา พบว่า การจำแนกประเภทของข้อความด้วยวิธีนาอีฟเบส์เป็นวิธีที่มีประสิทธิภาพที่สุดในการจำแนกประเภทข้อความภาษาอังกฤษ [24] แต่เนื่องจากงานวิจัยฉบับนี้ จำแนกประเภทเอกสารข้อความภาษาไทย ผู้วิจัยจึงทดลองเปรียบเทียบประสิทธิภาพกับวิธีการจำแนกประเภทเอกสาร 3 วิธี ได้แก่ วิธีนาอีฟเบส์ วิธีเครื่องจักรเวกเตอร์สนับสนุน และวิธีต้นไม้ตัดสินใจ ทั้งนี้ระนาบความหมายที่ได้จากการคำนวณหาความหมายแฝงนั้นเป็นระนาบความสัมพันธ์ที่มีขนาดใหญ่มาก ดังนั้นเราจะแบ่งข้อมูลออกเป็น 10 ชุดข้อมูลและทดสอบชุดข้อมูลด้วยวิธีการทดสอบแบบไขว้ข้ามสิบพับ กล่าวคือในหนึ่งรอบของการทดสอบ 9 ชุดข้อมูลเป็นชุดฝึกฝน และ 1 ชุดข้อมูลเป็นชุดทดสอบ แล้วดำเนินการทดสอบซ้ำเป็นจำนวน 10 ครั้ง เพื่อให้ได้ผลเป็นเปอร์เซ็นต์ความถูกต้องของการจำแนกประเภทของเอกสารข้อความ

จากนั้น จำแนกประเภทของคำตามคลาสที่กำหนด กล่าวคือ สำหรับคลาส Y จะสร้างรูปแบบความน่าจะเป็นของคำในเอกสารให้เป็นคลาส Y ได้ดังนี้ P(X|Y = y) กล่าวคือ ความน่าจะเป็นของเอกสาร X จะอยู่ในคลาส Y มีค่าเท่ากับ y เช่น ต้องการจำแนกสัตว์เป็นสัตว์บกกับสัตว์น้ำ จะเขียนความน่าจะเป็นได้ว่า P(X={เสือ, ช้าง, ม้า, วัว, ควาย, ...} | Y=สัตว์บก) = 0.89 หรือ P(X={ปลา, ปลาหมึก, กุ้ง, หอย, ปู, ...} | Y=สัตว์น้ำ) = 0.93 เป็นต้น นอกจากนี้พบว่า ข้อดีของการจำแนกประเภทข้อความด้วยวิธีนาอีฟเบส์ คือ สามารถใช้สถิติในการจำแนก ซึ่งสามารถทนต่อสิ่งรบกวนอื่นได้ดี และ สามารถใช้เมื่อชุดข้อมูลฝึกฝนมีขนาดใหญ่ได้ดี [24] อีกทั้ง เร็วและง่ายในการนำไปปฏิบัติ และมีรูปแบบอย่างเป็นทางการที่ดีในการเข้าใจจากประสบการณ์การเรียนรู้ข้อมูล [23]

2.1.7 การจำแนกอารมณ์จากข้อความ (Emotion Classification from Text)

ในการจำแนกอารมณ์จากข้อความ มีปัจจัยแวดล้อมได้หลายรูปแบบที่เอื้อประโยชน์ในการวิเคราะห์ ได้แก่

(1) ตัวบ่งชี้อารมณ์จากข้อความ สามารถแบ่งได้ 2 ประเภทคือ ทางกายภาพ เช่น ไอคอนอารมณ์ (Emotion Icon หรือ Emoticon) [16] หรือตัวหนังสือที่หนาอาจจะบ่งบอกว่าเป็นคำที่เน้นและต้องการสื่ออารมณ์บางอย่าง และในเชิงความหมายของข้อความ ซึ่งข้อความเกิดจากคำที่ประกอบขึ้นเป็นประโยคโดยที่คำเหล่านี้เป็นส่วนหนึ่งของประโยค (Part of Speech, POS) และทำหน้าที่ต่างๆในประโยค สามารถใช้เป็นตัวคัดกรองเบื้องต้นได้ [1,19] เชื่อว่าอารมณ์และความรู้สึกจะสามารถอธิบายได้ด้วยคำนาม (Noun) และคำคุณศัพท์ (Adjective) ของภาษา ส่วนเหตุการณ์ที่มีผลต่อการแสดงออกของอารมณ์ (Affective Events) จะแสดงความรู้สึกด้วยคำกริยา (Verb) [1] เป็นต้น

(2) วิธีการที่ใช้ในการจำแนกอารมณ์จากข้อความ [14] พบว่ามี 3 วิธีการหลัก ได้แก่

- ใช้เทคนิคของการเรียนรู้ของเครื่อง (Machine Learning) [16]
- ใช้การอ้างอิงเทียบจากพจนานุกรมระหว่างคำและอารมณ์ที่ถูกสร้างขึ้นและมีการเผยแพร่สู่สาธารณะให้คนทั่วไปสามารถใช้อ้างอิงได้ เช่น WordNet [14,25], ConceptNet [3,4], ANEW (Affective Norms English Words) [3] หรือสร้างพจนานุกรมอารมณ์ขึ้นมาใหม่โดยการอ้างอิงพจนานุกรมที่บอกเพียงวิธีการใช้คำมาหาความสัมพันธ์ใหม่ที่เกี่ยวข้องกับอารมณ์ เช่น พจนานุกรมภาษาจีนที่ชื่อว่า HowNet [1] เป็นต้น สำหรับ WordNet, ConceptNet และ ANEW มีข้อจำกัดคือคำจะต้องเป็นภาษาอังกฤษเท่านั้น และ HowNet ก็มีข้อจำกัดคือคำจะต้องเป็นภาษาจีนเท่านั้น สำหรับพจนานุกรมภาษาไทยยังไม่มีส่วนที่กล่าวถึงอารมณ์ของคำ [26]

- ใช้คนให้คะแนนหรือระบุ ซึ่งในกรณีนี้ จะได้อารมณ์ความรู้สึกที่เกิดขึ้นจากความรู้สึกจริงๆของคน สามารถแบ่งได้ออกเป็น 2 ประเภท คือ การจำแนกอารมณ์โดยอ้างอิงการให้คะแนนจากผู้อ่าน [17,27] เช่น ให้ผู้อ่านให้คะแนนจากวงล้อการประเมินอารมณ์ [2] หรือ การจำแนกอารมณ์จากการกำกับอารมณ์ (Emotion Tagging) จากผู้เขียน [15,22,28] ทั้งนี้พบว่าการนำอารมณ์ที่ผู้เขียนกำกับไว้กับข้อความที่เขียนนั้น อาจจะไม่สื่อความรู้สึกที่แท้จริงที่ต้องการสื่อในข้อความได้ [28] เนื่องจากในขณะที่ผู้เขียนกำลังเขียนข้อความอาจเกิดความสับสน หรือไม่ได้กำกับความรู้สึกที่แท้จริง ดังนั้นเชื่อว่าอารมณ์ที่เกิดจากผู้อ่าน (Readers' Emotion) จะให้ความถูกต้องมากกว่า

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยที่เกี่ยวกับการจำแนกอารมณ์จากข้อความ

F. Sugimoto และ Y. Masahide [19] จำแนกอารมณ์จากนวนิยายภาษาญี่ปุ่นจากการคำนวณความถี่ที่ปรากฏขึ้นและรูปแบบของหน้าที่ของคำในประโยคโดยไม่จำเป็นต้องรู้ความหมายของคำ ข้อมูลในงานวิจัยนี้ เป็นบทความซึ่งประกอบด้วยคำและข้อความที่มีความยาวมาก ดังนั้นวิธีการแรกคือการแบ่งข้อความที่ยาวนี้เป็นช่วงย่อยโดยพิจารณาจากน้ำหนักของความถี่ของคำที่มีน้ำหนักของอารมณ์มากน้อยต่างกันที่เกิดขึ้นในแต่ละประโยค จากนั้นคำนวณหาความคล้ายกันในประโยคที่มีน้ำหนักอารมณ์มากกว่า และจำแนกอารมณ์ด้วยการระบุหน้าที่ของคำในประโยค (คำนาม คำขยายคำนาม หรือคำกริยา) และระบุอารมณ์ของแต่ละคำ โดยแบ่งแบบหยาบก่อนได้เป็น 3 กลุ่มคือ พอใจ (Pleasantness) ไม่พอใจ (Unpleasantness) หรือ เป็นกลาง (Neutral) เมื่อแบ่งกลุ่มใหญ่ได้แล้ว จึงแบ่งเป็นอารมณ์ย่อยได้เป็น 5 อารมณ์ได้แก่ สนุกสนาน (Joy) ซึ่งจัดอยู่ในกลุ่มของกลุ่มอารมณ์พอใจ และ เสียใจ (Sorrow) หรือ โกรธ (Anger) ในกลุ่มของกลุ่มอารมณ์ไม่พอใจ อารมณ์ตื่นเต้น (Surprise) จะอยู่ในกลุ่มอารมณ์พอใจหรือไม่พอใจนั้นขึ้นกับความหมายของคำ จะต้องพิจารณาอีกครั้งหนึ่ง ทั้งนี้ในกระบวนการการจำแนกอารมณ์สามารถกล่าวได้ 3 วิธี ได้แก่ การวิเคราะห์ความสัมพันธ์กับคำอื่นโดยพิจารณาหน้าที่ของคำ การสร้างกฎขึ้นมาจากหน้าที่ของคำเพื่อระบุอารมณ์ให้กับคำภาษาญี่ปุ่นโดยอ้างอิงจากพจนานุกรมภาษาอังกฤษ และสุดท้าย วิเคราะห์รูปแบบของหน้าที่ของคำในประโยคที่มีผลต่ออารมณ์ โดยมีรายละเอียดและตัวอย่างดังนี้ ตัวอย่างเช่น

(1) This book is interesting. เป็นประโยค ที่ประกอบด้วยคำนามและคำขยายคำนามจะมีรูปแบบเป็น S+V+C เมื่อ S แทนประธาน V แทนคำกริยา และ C แทนคำขยายคำนาม

สามารถคำนวณการปรากฏของคำที่แสดงอารมณ์กับคำทั้งหมดได้คะแนนอารมณ์ของทั้งประโยคนี้ เท่ากับ สนุกสนาน 1.0 สามารถแตกประโยคเป็นหน้าที่ของคำได้ดังนี้

ประธาน: This (Neutral) book ให้อารมณ์เป็นกลาง (Neutral)

กริยา: is

คำขยายคำนาม: interesting ให้อารมณ์ สนุกสนาน 1.0

(2) A beautiful lady weeps. เป็นประโยคที่ประกอบด้วยคำกริยาจะมีรูปแบบเป็น S+V สามารถคำนวณการปรากฏของคำที่แสดงอารมณ์กับคำทั้งหมดได้คะแนนอารมณ์ของทั้งประโยคนี้ เท่ากับ เสียใจ 1.0 สามารถแตกประโยคเป็นหน้าที่ของคำได้ดังนี้

ประธาน : A beautiful (Joy 0.7) lady ให้อารมณ์ สนุกสนาน 0.7

กริยา: weeps ให้อารมณ์เสียใจ 0.9

(3) A hateful enemy loses. เป็นประโยคที่ประกอบด้วยคำกริยาแต่มีรูปแบบเป็น S+V+O คือ ประธาน กริยาและกรรม สามารถคำนวณการปรากฏของคำที่แสดงอารมณ์กับคำทั้งหมดได้คะแนนอารมณ์ของทั้งประโยคนี้ เท่ากับ สนุกสนาน 0.7 สามารถแตกประโยคเป็นหน้าที่ของคำได้ดังนี้

ประธาน : A hateful (Anger 0.9) enemy ให้อารมณ์ โกรธ 0.9

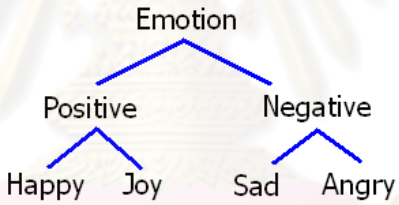
กริยา: loses ให้อารมณ์เสียใจ 0.8

เป็นต้น

C. Strapparava และ R. Mihalcea [14] ได้เสนอทฤษฎีในการวิเคราะห์อารมณ์จากข้อความขึ้นจากการทดลองที่เกี่ยวข้องในหลายรูปแบบ โดยแบ่งประเภทของข้อความออกเป็น 6 อารมณ์สากล (Universal Emotion) ได้แก่ โกรธ (Anger), ขยะแขยง (Disgust), กลัว (Fear), สนุกสนาน (Joy), เศร้า (Sadness) และ ตื่นเต้น (Surprise) ศึกษาการวิเคราะห์อารมณ์ 2 แบบ ได้แก่ การวิเคราะห์จากพื้นฐานความรู้ของคำ (Knowledge-based) ด้วยการเลือกวิเคราะห์คำจากหัวข้อข่าว โดยเริ่มจากการตัดประโยคออกเป็นคำๆ พิจารณาลำดับการเรียงลำดับของคำต่างๆ โดยคำนึงถึงความหมายที่คล้ายกันของคำนั้น (Synonym) และเปรียบเทียบด้วยวิธีการประเมินผลหลายแบบ (WordNet lexicon, LSA, Naïve Bayes) พบว่าวิธีการดึงข้อมูลจากพจนานุกรม WordNet ให้ค่าความถูกต้องสูงที่สุด และ การพิจารณาความหมายแฝงของคำที่สื่ออารมณ์ต่างๆ ร่วมกัน (LSA all emotion words) ให้ค่าความระลึก และ ค่าการวัดประสิทธิภาพด้วยค่า F-Measure ที่สูงที่สุด จากนั้นก็ศึกษาแบบที่เป็นการวิเคราะห์โดยมองภาพรวมทั้งหมด (Corpus-based) ด้วยการเลือกวิเคราะห์ข้อความยาวๆ เช่นข้อความจากบล็อกมาพิจารณาเปรียบเทียบด้วยวิธีการประเมินผลหลายแบบเช่นกัน (SWAT, UA, UPAR7) พบว่าวิธี UPAR7 เป็นวิธีในการ

ประเมินอารมณ์จากภาพรวมของข้อความได้ดีที่สุด เนื่องจากเป็นการใช้กฎของการสร้างภาษาในประโยคเข้ามาพิจารณาด้วย นอกจากนี้ยังมีงานวิจัย [28] ในรูปแบบคล้ายกันที่นำเสนอวิธีการจำแนกอารมณ์ของข้อความจากการพิจารณาอารมณ์ที่เกิดขึ้นร่วมกันของคำนาม, คำขยายคำนาม, คำกริยา และ คำขยายคำกริยา โดยนำคะแนนของอารมณ์ในแต่ละคำที่ได้จากพจนานุกรมอารมณ์มาหาค่าเฉลี่ยอารมณ์ของทั้งประโยค และ งานวิจัย [21] ที่วิเคราะห์โดยสนใจที่ตำแหน่งของแต่ละคำที่แสดงอารมณ์ บวกกับโครงสร้างของประโยค

C. Yang, K. H. Lin, และ H. Chen [16] แบ่งประเภทอารมณ์ของบล็อกโดยใช้เทคนิคของการเรียนรู้ของเครื่อง ด้วยวิธีเครื่องจักรเวกเตอร์สนับสนุน และ วิธีเขตข้อมูลแบบสุ่มอย่างมีเงื่อนไข (Conditional Random Field, CRF) ในการแบ่งประเภทของบทความหนึ่งเป็นอารมณ์เพียงหนึ่งอารมณ์ โดยการแบ่งเป็น 2 กลุ่มหลักก่อนคือกลุ่มที่แสดงอารมณ์เป็นบวก หรือกลุ่มที่แสดงอารมณ์เป็นลบ และแบ่งต่อในระดับถัดไป ซึ่งหากแบ่งครั้งแรกปรากฏอยู่ในกลุ่มอารมณ์เป็นบวก ก็จะพิจารณาต่อว่าเป็นมีความสุข (Happy) หรือสนุกสนาน (Joy) และหากแบ่งครั้งแรกปรากฏอยู่ในกลุ่มอารมณ์เป็นลบ ก็จะพิจารณาต่อว่าเป็นเศร้า (Sad) หรือโกรธ (Angry) ดังรูปที่ 5



รูปที่ 5 การแบ่งกลุ่มอารมณ์จากกลุ่มหลักก่อนแล้วจึงแยกเป็นอารมณ์ย่อยภายใต้อารมณ์หลักอย่างชัดเจน

การแบ่งประเภทอารมณ์ทำในระดับประโยค และประยุกต์เป็นอารมณ์ของทั้งบทความด้วยการออกแบบเกณฑ์การจำแนกประเภท 3 แบบ ดังนี้

c1 : กำหนดอารมณ์ของบทความ ด้วยประโยคที่จัดแบ่งประเภทแล้ว มีมากที่สุด

c2 : กำหนดอารมณ์ของบทความ ด้วยประโยคที่จัดแบ่งประเภทแล้ว มีความยาวของตัวบ่งชี้อารมณ์มากที่สุด

c3 : กำหนดอารมณ์ของบทความ ด้วยประโยคสุดท้ายของบทความ

โดยใช้ไอคอนอารมณ์เป็นตัววัดการบ่งชี้อารมณ์ของบทความที่ได้มาจากการเก็บข้อมูลจากบล็อก Taiwan's Yahoo Kimo Blog ทั้งนี้ ไอคอนอารมณ์เป็นสิ่งที่ผู้เขียนบล็อกนิยมใช้กันมาก เนื่องเปรียบเสมือนเป็นตัวแทนการแสดงอารมณ์ของคนได้อย่างรวบรัดและกระทัดรัดขึ้น พบว่าอารมณ์ที่เกิดจากประโยคสุดท้ายของบทความ สามารถแสดงบทบาทสำคัญที่สุดในการตัดสิน

อารมณ์ของทั้งบทความ และเทคนิควิธี CRF ให้ผลการแบ่งประเภทอารมณ์ที่มีความถูกต้องกว่าวิธีเครื่องจักรเวกเตอร์สนับสนุน

J. Yan, D. B. Bracewell, F. Ren, และ S. Kuroiwa [1] เชื่อว่า อารมณ์เกิดขึ้นภายในจิตใจของแต่ละบุคคลตามเหตุการณ์และสิ่งแวดล้อม เช่น ความสนุกสนาน (Joy) ความเศร้า (Sorrow) ความนับถือ (Respect) ความเกลียดชัง (Hate) ความรัก (Love) อาการทางจิตและความรู้สึกดังกล่าวนี้จะอธิบายด้วยคำนาม (Noun) และคำคุณศัพท์ (Adjective) ของภาษา ส่วนเหตุการณ์ที่มีผลต่อการแสดงออกของอารมณ์ (Affective events) จะแสดงความรู้สึกด้วยคำกริยา (Verb) โดยอ้างอิงจาก HowNet เป็นพจนานุกรมคำศัพท์ของจีน (Chinese lexical dictionary) ที่อธิบายความสัมพันธ์ทั้งแบบ ระหว่างแนวคิด กับ ระหว่างค่าคุณลักษณะ ระหว่างคำศัพท์ ซึ่งมักถูกนำไปอ้างอิงในการประมวลผลภาษาธรรมชาติ เช่นเดียวกับ WordNet ที่เป็นภาษาอังกฤษ จึงสามารถนำมาใช้อ้างอิงเพื่อสร้างพจนานุกรมอารมณ์ของคำศัพท์แบบกึ่งอัตโนมัติ

ขั้นตอนแรกในการสร้างพจนานุกรมอารมณ์นี้แบ่งประเภทอารมณ์ได้เป็น 3 ชนิดคือกลุ่มที่แสดงอารมณ์, กลุ่มที่เป็นทางเลือก คือมีหรือไม่มีอารมณ์ก็ได้ (ได้จากการที่คำศัพท์ที่ “desired” สามารถบ่งบอกได้หรือ “undesired” ไม่สามารถบ่งบอกได้), และกลุ่มที่ไม่แสดงอารมณ์ จากนั้นขั้นตอนที่สอง คัดเลือกอารมณ์จากหลักการของการสร้างกฎการวิเคราะห์ความหมายของคำ

วิธีการประเมินความถูกต้องทำโดยการใช้แรงงานคนประเมินซึ่งพบว่าทุกคำกริยา (5,498 คำ) ถูกประเมินได้ถูกต้อง และถูกประเภททั้งหมด พบว่า เมื่อเปรียบเทียบผลการประเมินอารมณ์ของภาษาอังกฤษกับภาษาจีน มีบางคำที่แสดงอารมณ์แตกต่างกัน อาจเนื่องจากวัฒนธรรมประเพณีในการดำรงชีวิตของแต่ละชนชาติที่ได้นำมากำหนดเป็นคำจำกัดความอารมณ์ของคำศัพท์แต่ละคำมีความแตกต่างกัน

A. J. Gill, D. Gergle, R. M. French, และ J. Oberlander [2] วิจัยการให้คะแนนอารมณ์จากผู้อ่านไดอารีหรือบล็อกที่มีความยาวประมาณ 50-200 คำ นำเสนอวิธีการให้คะแนนจากผู้อ่านโดยใช้วงล้อซึ่งผู้อ่านสามารถให้เลือกอารมณ์ได้มากกว่า 1 อารมณ์และสามารถเลือกน้ำหนักของอารมณ์ต่างๆได้อย่างครบถ้วน นอกจากนี้ยังกล่าวสนับสนุนอีกว่า คำเพียงหนึ่งคำไม่สามารถบอกอารมณ์ของทั้งข้อความหรือบทความได้ แต่ประโยคจะเป็นตัวกำหนดอารมณ์โดยรวมของบทความได้

C. Strapparava และ R. Mihalcea [17] กล่าวถึง ข้อความที่สื่ออารมณ์ ซึ่งการวิเคราะห์อารมณ์จากข้อความมีรูปแบบทั่วไปที่น่าสนใจ ได้แก่ การใช้ความรู้สึกวิเคราะห์ (sentiment analysis) การใช้คอมพิวเตอร์ช่วยในการสร้างอารมณ์อัตโนมัติ (computer assisted creativity) และ การวิเคราะห์จากคำกริยาที่มีผลต่อความรู้สึกในการแสดงอารมณ์ในการปฏิสัมพันธ์ระหว่าง

คอมพิวเตอร์กับมนุษย์ (verbal expressivity in human computer interaction) ซึ่งวิธีนี้เน้นการดึงอารมณ์มาจากภาษาธรรมชาติของมนุษย์ให้มีประสิทธิภาพมากขึ้น เนื่องจากคำกริยาสามารถบ่งบอกถึงระดับความเป็นไปได้ในการรู้จำอารมณ์ของมนุษย์ และสามารถที่จะบอกเป็นนัยให้รู้ว่าประโยคดังกล่าวแสดงอารมณ์อย่างไร นอกจากนี้ งานวิจัยยังเน้นเรื่องการจำแนกอารมณ์จากหัวข้อข่าวว่าเป็นขั้วบวกหรือขั้วลบ โดยการวินิจฉัยที่ความเชื่อมต่อกันระหว่างอารมณ์และความหมายที่เป็นรากศัพท์ เนื่องจากศัพท์แต่ละคำมีการสื่อความหมายของอารมณ์อยู่ แม้ว่าจะให้อารมณ์เป็นเฉยๆ ก็ตาม หรือแม้กระทั่งการให้ความรู้สึกพึงพอใจ หรือเสียใจ ก็สามารถสื่อได้จากความสัมพันธ์ของความหมายจากการแบ่งกลุ่มอารมณ์ คำบางคำให้อารมณ์ที่สื่อความหมายเฉพาะกับเรื่องที่กล่าวถึงเท่านั้น ในขณะที่มีอีกหลายคำที่การสื่ออารมณ์เป็นเรื่องของการจินตนาการ ตัวอย่างเช่น “mum” (แม่) “ghost” (ผี) “war” (สงคราม) เป็นต้น ทั้งนี้ งานวิจัยมีจุดประสงค์เพื่อกำหนดประเภทอารมณ์ของประโยค และจำแนกว่าประโยคนั้นให้ความหมายที่เป็นบวกหรือลบ

การเลือกหัวข้อข่าวมาเป็นตัวอย่างข้อมูลของการทดลองด้วยเหตุผล 2 ข้อคือ หัวข้อข่าวมักจะใช้คำที่สื่อถึงอารมณ์อย่างชัดเจน และมักจะถูกเขียนขึ้นในรูปแบบของการดึงดูดความสนใจจากผู้อ่าน และ โครงสร้างของหัวข้อข่าวเหมาะสมกับจุดประสงค์ที่สนใจจะทดลองกับข้อมูลในระดับที่เป็นประโยคเท่านั้น

งานวิจัยนี้นำเสนอวิธีการ โดยเริ่มจากการพัฒนาเว็บขึ้นมาเพื่อแสดงหัวข้อข่าวที่กำหนดไว้และแสดงแถบเลื่อนเพื่อให้ผู้อ่านใส่คะแนนการให้คำจำกัดความอารมณ์ มีคะแนนอยู่ระหว่าง [0,100] โดยที่ 0 หมายความว่า หัวข้อข่าวนี้ไม่สื่ออารมณ์ใดๆ และ 100 แสดงถึงการที่หัวข้อข่าวสื่ออารมณ์ได้อย่างชัดเจน ทั้งนี้การให้คำจำกัดความขั้วของอารมณ์ว่าเป็นบวกหรือลบนั้น มีคะแนนอยู่ระหว่าง [-100,100] โดยที่ -100 แสดงถึงหัวข้อข่าวที่ให้อารมณ์ที่เป็นลบมากที่สุด และ 100 แสดงถึงหัวข้อข่าวที่ให้อารมณ์ที่เป็นบวกมากที่สุดนั่นเอง

ระบบมีการตรวจสอบประสิทธิภาพของกระบวนการได้จากการหาค่าความถูกต้อง ความถูกต้อง และความระลึก จากฐานข้อมูลอารมณ์ที่มีอยู่แล้ว ได้แก่ UPAR7, SICS, CLaC, CLaC-NB, UA, SWAT

B. Yu และ J. Unsworth [29] เสนอข้อมูลที่ทดลองเปรียบเทียบประสิทธิภาพความถูกต้องในการจำแนกประเภทของข้อความระหว่างวิธีเครื่องจักรเวกเตอร์สนับสนุน และนาอีฟเบส์ ได้ว่าเครื่องจักรเวกเตอร์สนับสนุนให้ผลที่มีประสิทธิภาพความถูกต้องสูงกว่านาอีฟเบส์ เมื่อเปรียบเทียบข้อความในระดับประโยค และนาอีฟเบส์ ให้ผลที่มีประสิทธิภาพความถูกต้องสูงกว่าเครื่องจักรเวกเตอร์สนับสนุน เมื่อเปรียบเทียบข้อความในระดับบทความ

S. Steidl, M. Levit, A. Batliner, E. Nöth, และ H. Niemann [5] เปรียบเทียบการจำแนกประเภทของมนุษย์กับการจำแนกด้วยเครื่องเรียนรู้ พบว่า ในการจำแนกประเภทของมนุษย์สามารถเกิดความผิดพลาดได้เสมอแต่สุดท้ายมนุษย์มักจะเป็นผู้ตัดสินว่าการจำแนกประเภทสามารถทำได้ดีเพียงใด

2.2.2 งานวิจัยที่เกี่ยวกับการจำแนกประเภทจากข้อความภาษาไทย

ภาษาไทยมีโครงสร้างของภาษาที่ต่างจากภาษาอื่นๆ ซึ่งได้แก่ ระบบพยัญชนะภาษาไทย, การไม่บอกขอบเขตของคำในภาษาไทย, การไม่บอกความยาวของประโยคหนึ่งๆ, การที่ไม่มีคำกริยาที่ผันไปตามบุรุษและพจน์ เป็นต้น เพื่อสามารถแสดงคุณสมบัติของคำให้ได้มากที่สุด จึงมีการค้นคว้าวิจัยเพื่อการสร้างพจนานุกรมสำหรับภาษาไทยให้มีรายละเอียดมากขึ้นโดยการพิจารณาถึงความหมายของคำ งานวิจัยที่เกี่ยวข้องกับภาษาไทยกับการประยุกต์ใช้งานคอมพิวเตอร์ ได้แก่ การแปลงข้อความภาษาไทยเป็นเสียง (Thai Text to Speech) การตรวจตัวสะกด (Spell Checking) [13] การสร้างคลังข่าวและเหตุการณ์ในประเทศไทยบนอินเทอร์เน็ต การสรุปใจความสำคัญของเอกสารด้วยวิธีการนับความถี่คำกับส่วนกลับเอกสาร เป็นต้น

J. Polpinij, C. Sibunruang, R. Chamchong, A. Chotthanom, และ S. Puangpronpitag [31] จำแนกเว็บอนาจารโดยใช้ข้อความภาษาไทยภายในเว็บในการตรวจสอบประเภทของเว็บ เรียกว่า การกรองด้วยเนื้อหาเป็นพื้นฐาน (Content-based) มีการอ้างอิงการศึกษาเทคนิค โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ว่าเป็นวิธีการที่มีประสิทธิภาพ อยู่ในระดับที่ยอมรับได้ แต่ใช้เวลาในการประมวลผลมาก ไม่เหมาะในการที่จะประมวลผลแบบเวลาจริง (Real time) และศึกษางานวิจัยของ Du และคณะ ที่เสนอการจำแนกประเภทที่เรียกว่าการจำแนกหนึ่งคลาส (One-class classification) ซึ่งเป็นวิธีการเพื่อพิจารณาผลลัพธ์ในการจำแนกประเภทว่าใช่หรือไม่ หรือ เป็นอารมณ์บวกหรือลบ โดยกำหนดให้กลุ่มเอกสารที่สนใจเป็นบวก และนอกเหนือจากนั้นเป็นลบ แต่วิธีการนี้ ไม่เหมาะสมกับสภาพปัญหาเนื่องจากเนื้อหาอาจจะมีความหลากหลายมากเกินไป งานวิจัยฉบับนี้จึงนำเสนอ วิธีการสร้างตัวแยกประเภทในเชิงสถิติ ที่เรียกว่าตัวจำแนกความน่าจะเป็น (Probabilistic Classifier) จาก นาอีฟเบส์ เพื่อกรองเว็บอนาจารจากข้อความภาษาไทย กระบวนการของงานวิจัยดังกล่าวนี้ ได้นำเสนอการประยุกต์การจำแนกประเภทเอกสารข้อความ สำหรับการกรองเว็บอนาจาร คือ

1. เตรียมเอกสารข้อความสำหรับการเรียนรู้

2. แปลงข้อความให้อยู่ในรูปของหน่วยที่เล็กที่สุด นั่นคือ "คำ" ด้วยการตัดคำแบบเทียบคำยาวที่สุดซึ่งให้ความถูกต้องมากกว่า 80%

3. หาความถี่ของทุกคำที่ไม่ซ้ำกัน (Term Frequency: TF)
เลือกตัดคำที่ไม่จำเป็นออกจากระบบ ซึ่งมักจะเป็นคำที่มีความถี่สูง แต่ไม่มีความจำเป็นต่อการประมวลผลเลย

4. นำคำมาเทียบคำในคลังคำอนาจารที่พบ เพื่อการให้ค่าน้ำหนักของคำโดยใช้สมการความถี่คำกับส่วนกลับเอกสารซึ่งเป็นวิธีการในการให้น้ำหนักคำจากความถี่ที่ปรากฏในเอกสารข้อความอย่างง่าย

5. ใช้ Vector Space Model ในการแสดงความสัมพันธ์ระหว่างแต่ละเอกสารและคำอนาจารทั้งหมดที่ปรากฏอยู่ (Document Word Matrix)

6. ใช้ตัวจำแนกนาอีฟเบส์สำหรับการกรองเว็บ เนื่องจากนาอีฟเบส์มักนิยมใช้ในงานวิจัยที่มีกระบวนการการจำแนกประเภท โดยตัวแยกประเภทเรียกว่า "นาอีฟ" ผลลัพธ์จะได้จากการประมวลผลในเชิงสถิติ บางครั้งเรียกว่า ตัวจำแนกความน่าจะเป็น ซึ่งนาอีฟเบส์นิยมใช้กับการจำแนกประเภทเอกสาร [30] โดยการใช้ชุดเอกสารของการเรียนรู้มาสร้างเป็นตัวแบบในการประมาณค่าพารามิเตอร์ แล้วจึงนำตัวแบบนี้ไปใช้กับเอกสารใหม่เพื่อการจำแนกประเภท

การวัดประสิทธิภาพของงานวิจัยฉบับนี้ หาค่าความถูกต้อง (Precision: P) ค่าความระลึก (Recall: R) การวัดคุณภาพ (F-Measure) จากชุดข้อมูลซึ่งเป็นเว็บที่ใช้ข้อความภาษาไทยรวบรวมไว้ระหว่างเดือนมีนาคม 2547 ถึงมกราคม 2548 จำนวน 400 หน้าเว็บ โดยแบ่งเป็นชุดข้อมูลฝึกฝน = 300 เว็บ และชุดข้อมูลทดสอบ = 100 เว็บ และแสดงผลการทดสอบดังตารางที่ 4

ตารางที่ 4 แสดงการวัดประสิทธิภาพของงานวิจัยที่เกี่ยวข้อง

Cross Validation	ชุดข้อมูล	Precision (%)	Recall (%)	F-Measure (%)
1-fold	เว็บทั่วไป	97	93	94.96
	เว็บอนาจาร	98	95	96.48
2-fold	เว็บทั่วไป	97	94	95.48
	เว็บอนาจาร	96	93	94.46
3-fold	เว็บทั่วไป	98	92	94.91
	เว็บอนาจาร	97	96	96.50

T. Charoenporn, C. Kruengkrai, V. Sornlertlamvanich, H. Isahara และ T. Theeramunkong [26] นำเสนองานวิจัยที่วิเคราะห์คำศัพท์ภาษาไทยและสร้างเป็นพจนานุกรมภาษาไทยขึ้นเพื่อให้มีการระบุข้อมูลที่เกี่ยวข้องกับคำในภาษาไทยให้ละเอียดมากขึ้น โดยมีการ

เก็บรวบรวมข้อจำกัดแบ่งเป็น 2 ประเภทได้แก่ เงื่อนไขเชิงตรรกะ (logical constraint) ซึ่งประกอบด้วย ความหมายของคำ (meaning), คำที่มีความหมายเหมือนกัน (same meaning), คำที่มีความหมายตรงข้าม (opposite meaning), หน้าที่ของคำในประโยค (part of speech) และประเภทเงื่อนไขเชิงความหมาย (semantic constraint) ซึ่งประกอบด้วย คำประธานที่ใช้ร่วมกับคำนี้, คำที่ใช้เป็นกรรมของคำนี้, เครื่องมือที่ใช้ร่วมกับคำนี้, สถานที่ที่เกี่ยวข้องกับคำนี้ และเวลาหรือเหตุการณ์ที่ใช้ร่วมกับคำนี้ ด้วยวิธีการเก็บรวบรวมข้อมูลจากเว็บไซต์ต่างๆ นำมาแยกคำ และหาค่าความน่าจะเป็นของการเกิดร่วมกันของคำ (co-occurrence) โดยใช้คนช่วยระบุความสัมพันธ์ แต่ทั้งนี้การเก็บรวมข้อมูลของพจนานุกรมฉบับนี้ก็ยังไม่มีส่วนของการระบุอารมณ์ของคำต่างๆ

นอกจากที่กล่าวมาแล้ว สามารถอธิบายเพิ่มเติมได้อีกว่าทำไมต้องภาษาไทย?

(1) เพราะคนไทยสามารถเข้าใจภาษาไทยได้ดีกว่าภาษาอื่นๆ

(2) ในปัจจุบันมีงานวิจัยจำนวนไม่น้อยที่กล่าวถึงการจำแนกประเภทจากข้อความภาษาไทย [12,31] แต่ก็ยังไม่มีงานวิจัยที่วิเคราะห์จำแนกอารมณ์จากข้อความภาษาไทย

(3) อารมณ์ที่เกิดจากการตีความของคนไทย (ด้วยสังคมและวัฒนธรรม) ในประโยคแบบเดียวกันกับภาษาอังกฤษอาจจะให้ความหมายที่แสดงอารมณ์ได้ต่างกัน

(4) คำภาษาไทย มีการแสดงอารมณ์ได้ลึกซึ้งกว่าภาษาอังกฤษ

(5) นอกจากนี้ โครงสร้างการเขียนข้อความภาษาไทยก็แตกต่างจากภาษาอื่น กล่าวคือประโยคภาษาไทยมีระบบการเขียนที่ต่อเนื่องกันโดยไม่มีการเว้นระยะจนกว่าจะจบประโยค หากเขียนวรรคตอนผิดจะทำให้เสียความหมายหรือทำให้ความหมายเปลี่ยนไปได้ [7]

H. Bacan, I. S. Pandzic และ D. Gulija [32] สร้างระบบที่แบ่งประเภทของข่าว โดยที่ระบบรับอินพุตเป็นข้อความในเนื้อหาข่าว และเอาต์พุตเป็นประเภทข่าว ประกอบด้วย 2 ขั้นตอนหลัก ก็คือ ขั้นตอนที่ 1 การทำอินเด็กซ์ให้เอกสาร (Document Indexing) ประกอบด้วย 2 ขั้นตอนย่อย ได้แก่ การเลือกคำสำคัญที่เหมาะสมจากคำทั้งหมดของเอกสาร ซึ่งควรเลือกคำที่สื่อความหมาย (carry the meaning) และตัดคำที่ไม่ได้ช่วยจำแนกเอกสารให้มีความแตกต่าง ถ้าคำที่เลือกปรากฏเป็นจำนวนน้อยเกินไป ก็จะไม่สนใจ จัดว่าเป็นคำที่ไม่มีความหมายในการจัดประเภท ถ้าคำที่เลือกปรากฏเป็นจำนวนมากเกินไป ก็จัดว่าเป็นคำทั่วไป ไม่สามารถนำมาจัดประเภทได้ จากนั้นก็ ให้น้ำหนักกับคำนั้นๆ (Assign weights to keyword) โดยวิธีการหาความถี่คำกับส่วนกลับเอกสาร

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยนี้นำเสนอวิธีการจำแนกอารมณ์ในข้อความภาษาไทย ประกอบด้วยขั้นตอนหลัก 4 ขั้นตอน ดังนี้

3.1 ขั้นตอนการเตรียมข้อความภาษาไทย

(1) จัดเตรียมข้อความภาษาไทยสำหรับการเรียนรู้และการทดสอบ ที่เก็บรวบรวมได้จากบล็อกหรือกระทู้บนอินเทอร์เน็ต

(2) ให้ผู้อ่าน อ่านข้อความภาษาไทย และวิเคราะห์ให้ค่าอารมณ์ที่เหมาะสมกับข้อความนั้นมากที่สุด แสดงดังตารางที่ 5

ตารางที่ 5 แสดงรูปแบบตารางความสัมพันธ์ระหว่างชุดข้อความกับอารมณ์ที่ให้ผู้อ่านระบุ

	โกรธ	ตื่นเต้น	กลัว	มีความสุข	เศร้า
ข้อความ 1	√				
ข้อความ 2			√		
ข้อความ 3					√
ข้อความ 4		√			

(3) นำข้อความที่เตรียมไว้มาตัดคำด้วยโปรแกรมการตัดคำภาษาไทย (SWATH) โดยเลือกวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบ ไบแกรม กับลำดับของหน้าที่ของคำ เพื่อคัดเฉพาะคำที่มีบทบาทต่ออารมณ์ซึ่งได้แก่ คำกริยา คำนาม เท่านั้น

3.2 ขั้นตอนการวิเคราะห์ความหมายแฝงของคำ

(1) สร้างเมตริกซ์ความสัมพันธ์ของแต่ละคำกับชุดข้อความ โดยการนับความถี่ของคำที่ปรากฏในแต่ละชุดข้อความตัวอย่างการนับความถี่ของคำนี้แสดงไว้ในตารางที่ 6

ตารางที่ 6 ตัวอย่างการนับความถี่ของคำในแต่ละชุดข้อความ

อารมณ์สูงสุด	เศร้า	มีความสุข	กลัว	เศร้า	โกรธ
คำ	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4	ชุดข้อความ 5
เจ็บ	2	0	1	0	0
น่าเป็นห่วง	0	0	1	0	0
ไม่เหลียวแล	1	0	0	0	3
ระทม	1	0	0	3	0
เหยียบย่ำ	1	0	0	0	0
เหวอะหวะ	0	0	1	0	0
อกหัก	0	0	0	1	0
วิวาห์	0	1	0	0	1
หวาน	0	2	0	0	0

(2) ให้ค่าน้ำหนักคำด้วยวิธีการคำนวณหาความถี่คำกับส่วนกลับเอกสาร

ตัวอย่างการคำนวณหาค่าน้ำหนักของคำ จากข้อมูลในตารางที่ 6 แสดงได้ดังนี้

ข้อความที่ 1 ประกอบด้วยคำทั้งหมด 5 คำ ซึ่งมีคำว่า “เจ็บ” ที่ปรากฏในข้อความเป็นความถี่ 2 ครั้ง

จะได้ว่า

n คือ ความถี่ที่คำนั้นปรากฏในเอกสารหนึ่ง มีค่าเท่ากับ 2

N คือ จำนวนคำทั้งหมดในเอกสารหนึ่ง มีค่าเท่ากับ 5

ดังนั้น TF มีค่าเท่ากับ $\frac{2}{5}$

$|D|$ คือ จำนวนเอกสารทั้งหมด มีค่าเท่ากับ 5

DF คือ จำนวนเอกสารที่มีคำนั้นปรากฏอยู่ มีค่าเท่ากับ 3

กล่าวคือ

$$(TF - IDF)_{\text{“เจ็บ”}} = \frac{2}{5} \times \left(1 + \log \frac{5}{3}\right) = 0.62$$

จากการคำนวณหาค่าน้ำหนักคำ จะทำให้ได้ผลลัพธ์เป็นเมตริกซ์ดังตารางที่ 7

ตารางที่ 7 เมตริกซ์ความสัมพันธ์ของคำกับชุดข้อความที่ผ่านการให้ค่าน้ำหนัก

คำ	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4	ชุดข้อความ 5
เจ็บ	0.56	0.00	0.47	0.00	0.00
น่าเป็นห่วง	0.00	0.00	0.57	0.00	0.00
ไม่เหลียวแล	0.28	0.00	0.00	0.00	1.05
ระทม	0.28	0.00	0.00	1.05	0.00
เหยียบย่ำ	0.34	0.00	0.00	0.00	0.00
เหวอะหวะ	0.00	0.00	0.57	0.00	0.00
อกหัก	0.00	0.00	0.00	0.42	0.00
วิวาห์	0.00	0.47	0.00	0.00	0.35
หวาน	0.00	1.13	0.00	0.00	0.00

(3) นำเมตริกซ์ที่ได้ผ่านอัลกอริทึมการแยกค่าแบบเดี่ยว (Singular Value Decomposition, SVD) จะได้ผลลัพธ์เป็นระนาบความหมายที่แสดงค่าความหมายของความสัมพันธ์ของคำและข้อความ ดังตารางที่ 8

ตารางที่ 8 เมตริกซ์ความสัมพันธ์ของคำกับชุดข้อความที่ผ่านอัลกอริทึมการแยกค่าแบบเดี่ยว

คำ	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4	ชุดข้อความ 5
เจ็บ	-0.2268672	- 0.5926964	-0.3449062	0.1114663	0.0189002
น่าเป็นห่วง	0.1533002	-0.5053966	0.0538319	-0.1084684	0.1769466
ไม่เหลียวแล	-0.0623466	-0.3447682	0.4760374	0.8775234	-0.2479696
ระทม	-0.7095491	-0.3175689	0.4375773	-0.4453423	-0.4321880
เหยียบย่ำ	-0.2144869	-0.1068364	-0.1824577	0.1219781	-0.1000593
เหวอะหวะ	0.1533002	-0.5053966	-0.0538319	-0.1084684	0.1769466
อกหัก	-0.2131651	-0.0918344	0.2351346	-0.2183179	-0.1399145
วิวาห์	0.2778715	-0.1262388	0.1222392	0.1974145	-0.4430985
หวาน	0.5764801	-0.0977178	-0.2080315	-0.1481240	-0.9326329

จากระนาบความหมาย สามารถคำนวณหาความสัมพันธ์ระหว่างคำกับคำได้จาก $\hat{X}\hat{X}^T$ และหาความสัมพันธ์ระหว่างข้อความกับข้อความได้จาก $\hat{X}^T\hat{X}$

ตารางที่ 9 ตารางค่าความสัมพันธ์ระหว่างคำกับคำจากระนาบความหมาย

	น่าเป็นห่วง	ไม่เหลียวแล	ระทม	เหยียบย่ำ	เหวอะหวะ	อกหัก	วิวาห์	หวาน
เจ็บ	0.237	0.147	0.140	0.187	0.275	-0.005	-0.017	-0.035
น่าเป็นห่วง		0.051	0.047	-0.020	0.319	0.025	0.013	-0.022
ไม่เหลียวแล			0.078	0.095	-4.678E-09	4.035E-08	0.368	-2.751E-08
ระทม				0.095	-2.450E-08	0.441	-2.976E-08	-3.332E-08
เหยียบย่ำ					-8.966E-09	4.165E-09	-9.619E-09	4.847E-08
เหวอะหวะ						-2.975E-09	4.045E-08	4.199E-08
อกหัก							9.775E-09	3.346E-10
วิวาห์								0.531

ตารางที่ 10 ตารางค่าความสัมพันธ์ระหว่างชุดข้อความกับชุดข้อความจากระนาบความหมาย

	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4	ชุดข้อความ 5
ชุดข้อความ 1	0.1774148	-0.358861	0.1925775	-0.2374
ชุดข้อความ 2		-0.095862	-0.120972	0.203293
ชุดข้อความ 3			0.1657732	-0.188466
ชุดข้อความ 4				0.007605

3.3 ขั้นตอนการจับคู่ความหมายบนระนาบกับคลาสอารมณ์

(1) จากระนาบความหมาย สามารถคำนวณหาค่าผลรวมของน้ำหนักของเวกเตอร์คำได้เพื่อเป็นตัวแทนของชุดข้อความ

$$\hat{d}_1 = [0.2311857\ ,\ 0.7912070\ ,\ 0.0827602\ ,\ 0.2078338\ ,\ 0.5200966]$$

$$\hat{d}_2 = [-0.3870819\ ,\ 0.4559054\ ,\ -0.1824021\ ,\ -0.7569342\ ,\ -0.1899946]$$

$$\hat{d}_3 = [0.1490384\ ,\ -0.3960734\ ,\ 0.1207096\ ,\ -0.5264551\ ,\ 0.7274528]$$

$$\hat{d}_4 = [0.2403508\ ,\ 0.0889287\ ,\ 0.8944349\ ,\ -0.2093887\ ,\ -0.3007753]$$

$$\hat{d}_5 = [-0.8466049\ ,\ -0.0368685\ ,\ 0.3811766\ ,\ 0.2507132\ ,\ 0.2715661]$$

(2) จับคู่แต่ละชุดข้อความกับคลาสอารมณ์สูงสุดที่ผู้อ่านระบุไว้ สร้างเป็นชุดข้อมูลสำหรับการฝึกฝนและทดสอบให้กับระบบ สำหรับตัวแบบที่ 1

$\hat{d}_1 = [0.2311857\ ,\ 0.7912070\ ,\ 0.0827602\ ,\ 0.2078338\ ,\ 0.5200966\ ,\ \#เศร้า]$

$\hat{d}_2 = [-0.3870819\ ,\ 0.4559054\ ,\ -0.1824021\ ,\ -0.7569342\ ,\ -0.1899946\ ,\ \#มีความสุข]$

$\hat{d}_3 = [0.1490384\ ,\ -0.3960734\ ,\ 0.1207096\ ,\ -0.5264551\ ,\ 0.7274528\ ,\ \#กลัว]$

$\hat{d}_4 = [0.2403508\ ,\ 0.0889287\ ,\ 0.8944349\ ,\ -0.2093887\ ,\ -0.3007753\ ,\ \#เศร้า]$

$\hat{d}_5 = [-0.8466049\ ,\ -0.0368685\ ,\ 0.3811766\ ,\ 0.2507132\ ,\ 0.2715661\ ,\ \#โกรธ]$

(3) สร้างเป็นชุดข้อมูลสำหรับการฝึกฝนและทดสอบให้กับระบบสำหรับตัวแบบที่ 2

(3.1) หาค่าความน่าจะเป็นของการเกิดขึ้นคู่กันของคำนามและคำกริยาที่เป็นอิสระต่อกัน ดังตารางที่ 11

ตารางที่ 11 แสดงผลลัพธ์ความน่าจะเป็นของการเกิดขึ้นคู่กันของคำนามและคำกริยา จากตัวอย่าง

คำที่ 1	คำที่ 2	ค่าความน่าจะเป็น			
		เศร้า	มีความสุข	กลัว	โกรธ
เจ็บ	น่าเป็นห่วง	0	0	0.333	0
เจ็บ	ไม่เหลียวแล	0.167	0	0	0
เจ็บ	ระทม	0.667	0	0	0
เจ็บ	เหยียบย่ำ	0.667	0	0	0
เจ็บ	เหวอะหวะ	0	0	0.333	0
เจ็บ	อกหัก	0	0	0	0
เจ็บ	วิวาห์	0	0	0	0
เจ็บ	หวาน	0	0	0	0
น่าเป็นห่วง	ไม่เหลียวแล	0	0	0	0
น่าเป็นห่วง	ระทม	0	0	0	0
น่าเป็นห่วง	เหยียบย่ำ	0	0	0	0
น่าเป็นห่วง	เหวอะหวะ	0	0	1.0	0
น่าเป็นห่วง	อกหัก	0	0	0	0
น่าเป็นห่วง	วิวาห์	0	0	0	0
น่าเป็นห่วง	หวาน	0	0	0	0
ไม่เหลียวแล	ระทม	0.25	0	0	0
ไม่เหลียวแล	เหยียบย่ำ	0.25	0	0	0
ไม่เหลียวแล	เหวอะหวะ	0	0	0	0

ไม่เหลียวแล	อกหัก	0	0	0	0
ไม่เหลียวแล	วิวาห์	0	0	0	0.375
ไม่เหลียวแล	หวาน	0	0	0	0
ระทม	เหยียบย่ำ	1.0	0	0	0
ระทม	เหวอะหวะ	0	0	0	0
ระทม	อกหัก	1.0	0	0	0
ระทม	วิวาห์	0	0	0	0
ระทม	หวาน	0	0	0	0
เหยียบย่ำ	เหวอะหวะ	0	0	0	0
เหยียบย่ำ	อกหัก	0	0	0	0
เหยียบย่ำ	วิวาห์	0	0	0	0
เหยียบย่ำ	หวาน	0	0	0	0
เหวอะหวะ	อกหัก	0	0	0	0
เหวอะหวะ	วิวาห์	0	0	0	0
เหวอะหวะ	หวาน	0	0	0	0
อกหัก	วิวาห์	0	0	0	0
อกหัก	หวาน	0	0	0	0
วิวาห์	หวาน	0	0.5	0	0

(3.2) เลือกเฉพาะคำคู่ที่มีความน่าจะเป็นที่จะปรากฏคู่กันและสะท้อนอารมณ์ ดังตารางที่ 12

ตารางที่ 12 ผลลัพธ์การคัดเลือกเฉพาะคำคู่ที่มีความน่าจะเป็นที่จะปรากฏคู่กันและสะท้อนอารมณ์

คำที่ 1	คำที่ 2	ค่าความน่าจะเป็น			
		เศร้า	มีความสุข	กลัว	โกรธ
เจ็บ	น่าเป็นห่วง	0	0	0.333	0
เจ็บ	ไม่เหลียวแล	0.167	0	0	0
เจ็บ	ระทม	0.667	0	0	0
เจ็บ	เหยียบย่ำ	0.667	0	0	0
เจ็บ	เหวอะหวะ	0	0	0.333	0
น่าเป็นห่วง	เหวอะหวะ	0	0	1.0	0
ไม่เหลียวแล	ระทม	0.25	0	0	0
ไม่เหลียวแล	เหยียบย่ำ	0.25	0	0	0
ไม่เหลียวแล	วิวาห์	0	0	0	0.375

ระทม	เหยียบย่ำ	1.0	0	0	0
ระทม	อกหัก	1.0	0	0	0
วิวาห์	หวาน	0	0.5	0	0

(3.3) นำคำคู่ต่อในตารางระนาบความหมาย และระบุค่าเป็นผลคูณของความน่าจะเป็นของคำคู่กับขนาดมุมระหว่างคำคู่ซึ่งหมายถึงการค่าน้ำหนักของความหมายของคำคู่ต่อชุดข้อความ ดังตารางที่ 13

ตารางที่ 13 ระนาบความหมายที่นำคำคู่มาต่อกับระนาบเดิมที่เป็นค่าความหมายของคำเดี่ยว

คำ	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4	ชุดข้อความ 5
เจ็บ	-0.2268672	- 0.5926964	-0.3449062	0.1114663	0.0189002
น่าเป็นห่วง	0.1533002	-0.5053966	0.0538319	-0.1084684	0.1769466
ไม่เหลียวแล	-0.0623466	-0.3447682	0.4760374	0.8775234	-0.2479696
ระทม	-0.7095491	-0.3175689	0.4375773	-0.4453423	-0.4321880
เหยียบย่ำ	-0.2144869	-0.1068364	-0.1824577	0.1219781	-0.1000593
เหวอะหวะ	0.1533002	-0.5053966	-0.0538319	-0.1084684	0.1769466
อกหัก	-0.2131651	-0.0918344	0.2351346	-0.2183179	-0.1399145
วิวาห์	0.2778715	-0.1262388	0.1222392	0.1974145	-0.4430985
หวาน	0.5764801	-0.0977178	-0.2080315	-0.1481240	-0.9326329
เจ็บ+น่าเป็นห่วง	0	0	0.078921	0	0
เจ็บ+ไม่เหลียวแล	0.024549	0	0	0.024549	0
เจ็บ+ระทม	0.09338	0	0	0.09338	0
เจ็บ+เหยียบย่ำ	0.124729	0	0	0.124729	0
เจ็บ+เหวอะหวะ	0	0	0.091575	0	0
น่าเป็นห่วง+ เหวอะหวะ	0	0	0.319	0	0
ไม่เหลียวแล+ระทม	0.0195	0	0	0.0195	0
ไม่เหลียวแล+ เหยียบย่ำ	0.02375	0	0	0.02375	0
ไม่เหลียวแล+วิวาห์	0	0	0	0	0.138
ระทม+เหยียบย่ำ	0.095	0	0	0.095	0
ระทม+อกหัก	0.441	0	0	0.441	0
วิวาห์+หวาน	0	0.2655	0	0	0

(3.4) จากระนาบความหมาย สามารถคำนวณหาค่าเวกเตอร์คำได้เพื่อเป็นตัวแทนของชุดข้อความได้ ดังนี้

$\hat{d}_1$ = [-0.2268672, 0.1533002, -0.0623466, -0.7095491, -0.2144869, 0.1533002, -0.2131651, 0.2778715, 0.5764801, 0, 0.024549, 0.09338, 0.124729, 0, 0, 0.0195, 0.02375, 0, 0.095, 0.441, 0]

$\hat{d}_2$ = [-0.5926964, -0.5053966, -0.3447682, -0.3175689, -0.1068364, -0.5053966, -0.0918344, -0.1262388, -0.0977178, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2655]

$\hat{d}_3$ = [-0.3449062, 0.0538319, 0.4760374, 0.4375773,-0.1824577, -0.0538319, 0.2351346, 0.1222392, -0.2080315, 0.078921, 0, 0, 0, 0.091575, 0.319, 0, 0, 0, 0, 0, 0]

$\hat{d}_4$ = [0.1114663, -0.1084684, 0.8775234, -0.4453423, 0.1219781, -0.1084684, -0.2183179, 0.1974145, -0.1481240, 0, 0.024549, 0.09338, 0.124729, 0, 0, 0.0195, 0.02375, 0, 0.095, 0.441, 0]

$\hat{d}_5$ = [0.0189002, 0.1769466, -0.2479696, -0.4321880, -0.1000593, 0.1769466, -0.1399145, -0.4430985, -0.9326329, 0, 0, 0, 0, 0, 0, 0, 0, 0.138, 0, 0, 0]

(3.5) จับคู่แต่ละชุดข้อความกับคลาสอารมณ์สูงสุดที่ผู้อ่านระบุไว้ สร้างเป็นชุดข้อมูลสำหรับการฝึกฝนและทดสอบให้กับระบบสำหรับตัวแบบที่ 2

$\hat{d}_1$ = [-0.2268672, 0.1533002, -0.0623466, -0.7095491, -0.2144869, 0.1533002, -0.2131651, 0.2778715, 0.5764801, 0, 0.024549, 0.09338, 0.124729, 0, 0, 0.0195, 0.02375, 0, 0.095, 0.441, 0, #เศร้า]

$\hat{d}_2$ = [-0.5926964, -0.5053966, -0.3447682, -0.3175689, -0.1068364, -0.5053966, -0.0918344, -0.1262388, -0.0977178, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2655, #มีความสุข]

$\hat{d}_3$ = [-0.3449062, 0.0538319, 0.4760374, 0.4375773,-0.1824577, -0.0538319, 0.2351346, 0.1222392, -0.2080315, 0.078921, 0, 0, 0, 0.091575, 0.319, 0, 0, 0, 0, 0, 0, #กลัว]

$\hat{d}_4$ = [0.1114663, -0.1084684, 0.8775234, -0.4453423, 0.1219781, -0.1084684, -0.2183179, 0.1974145, -0.1481240, 0, 0.024549, 0.09338, 0.124729, 0, 0, 0.0195, 0.02375, 0, 0.095, 0.441, 0, #เศร้า]

$\hat{d}_5$ = [0.0189002, 0.1769466, -0.2479696, -0.4321880, -0.1000593, 0.1769466, -0.1399145, -0.4430985, -0.9326329, 0, 0, 0, 0, 0, 0, 0, 0, 0.138, 0, 0, 0, #โกรธ]

3.4 ขั้นตอนการจำแนกประเภทอารมณ์

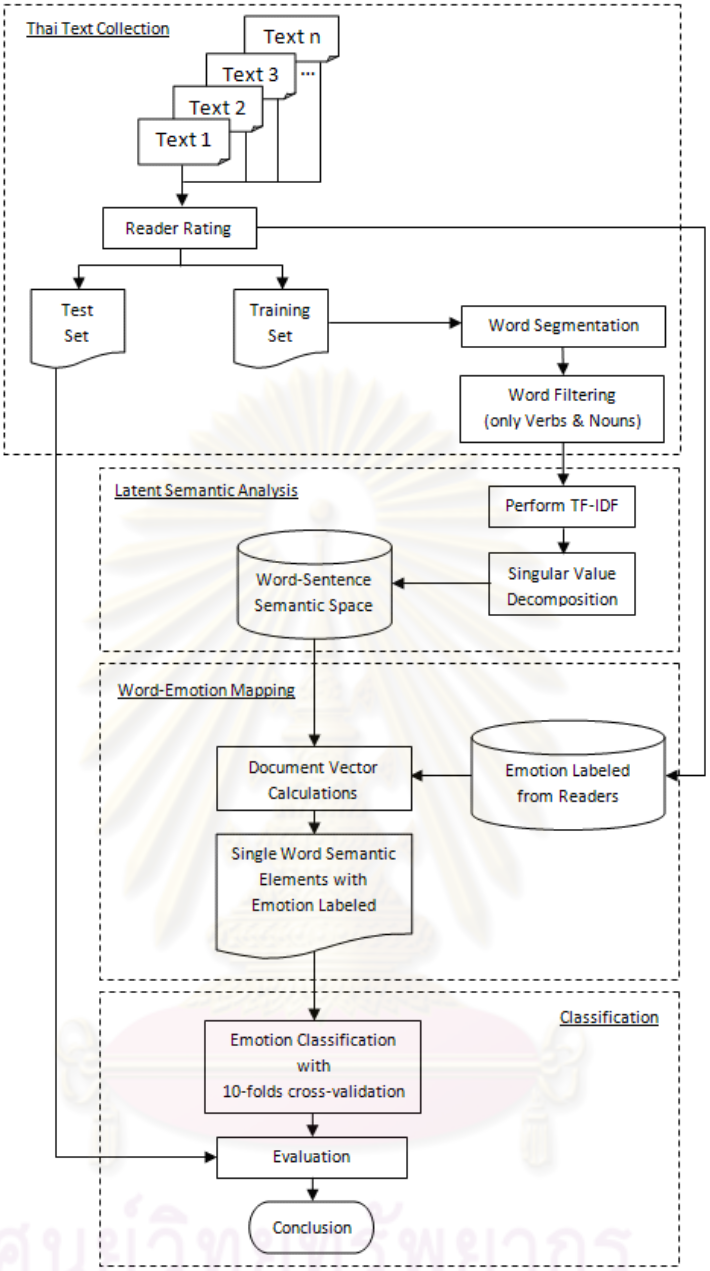
(1) จากการศึกษางานวิจัยที่ผ่านมา พบว่า การจำแนกประเภทของข้อความด้วยวิธีนาอีฟเบส์ เป็นวิธีที่มีประสิทธิภาพที่สุดในการจำแนกประเภทข้อความภาษาอังกฤษ [24] แต่เนื่องจากงานวิจัยฉบับนี้ จำแนกประเภทเอกสารข้อความภาษาไทย ผู้วิจัยจึงทดลองเปรียบเทียบประสิทธิภาพกับวิธีการจำแนกประเภทเอกสาร 3 วิธี ได้แก่ วิธีนาอีฟเบส์ วิธีเครื่องจักรเวกเตอร์สนับสนุน และวิธีต้นไม้ตัดสินใจ ทั้งนี้ระนาบความหมายที่ได้จากการคำนวณหาความหมายแฝงนั้นเป็นระนาบความสัมพันธ์ที่มีขนาดใหญ่มาก ดังนั้นเราจะแบ่งข้อมูลออกเป็น 10 ชุดข้อมูลและทดสอบชุดข้อมูลด้วยวิธีการทดสอบแบบไขว้ข้ามสิบพับ กล่าวคือในหนึ่งรอบของการทดสอบ 9 ชุดข้อมูลเป็นชุดฝึกฝน และ 1 ชุดข้อมูลเป็นชุดทดสอบ แล้วดำเนินการทดสอบซ้ำเป็นจำนวน 10 ครั้ง เพื่อให้ได้ผลเป็นเปอร์เซ็นต์ความถูกต้องของการจำแนกประเภทของเอกสารข้อความ

(2) สรุปผลการทดลอง

งานวิจัยนี้ทดลองเปรียบเทียบการจำแนกอารมณ์จากข้อความภาษาไทยระหว่าง 2 ตัวแบบ ได้แก่ ตัวแบบที่ 1 เป็นการหาความหมายแฝงระหว่างคำเดี่ยวกับชุดข้อความ และ ตัวแบบที่ 2 เป็นการประยุกต์การหาความหมายแฝงด้วยการพิจารณาคำคู่ที่มักปรากฏคู่กันในชุดข้อความด้วย

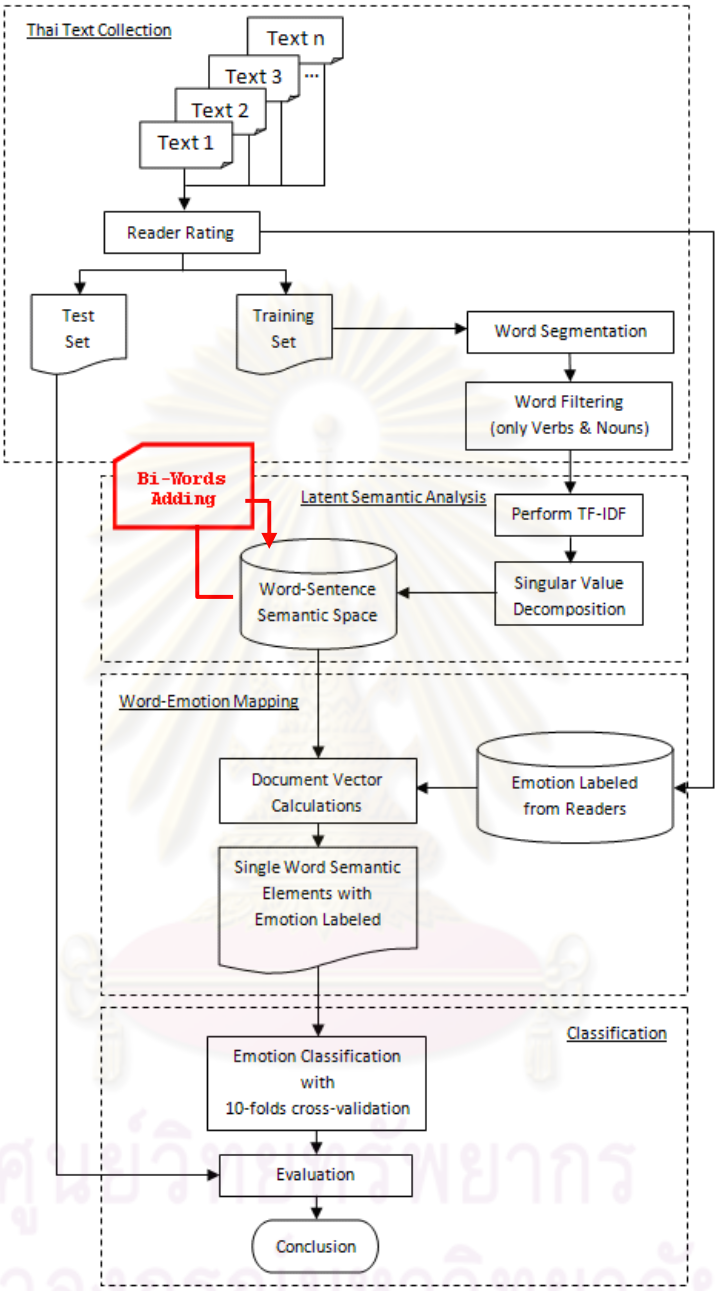
3.5 คำอธิบายตัวแบบที่ 1

แต่ละชุดข้อความจะถูกแปลงให้อยู่ในรูปของเวกเตอร์ของชุดข้อความ *dj = [wj1, wj2, …, wj489]* ซึ่งประกอบด้วยค่าความหมายของแต่ละคำ จากนั้นจึงทำจับคู่กับอารมณ์สูงสุดที่ผู้อ่านเป็นผู้ระบุ ได้เป็นเวกเตอร์ใหม่ *dj = [wj1, wj2, …, wj489, #emotion]* สำหรับนำไปเป็นชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบต่อไป สามารถแสดงได้ดังแผนภาพรวมของระบบในรูปที่ 6



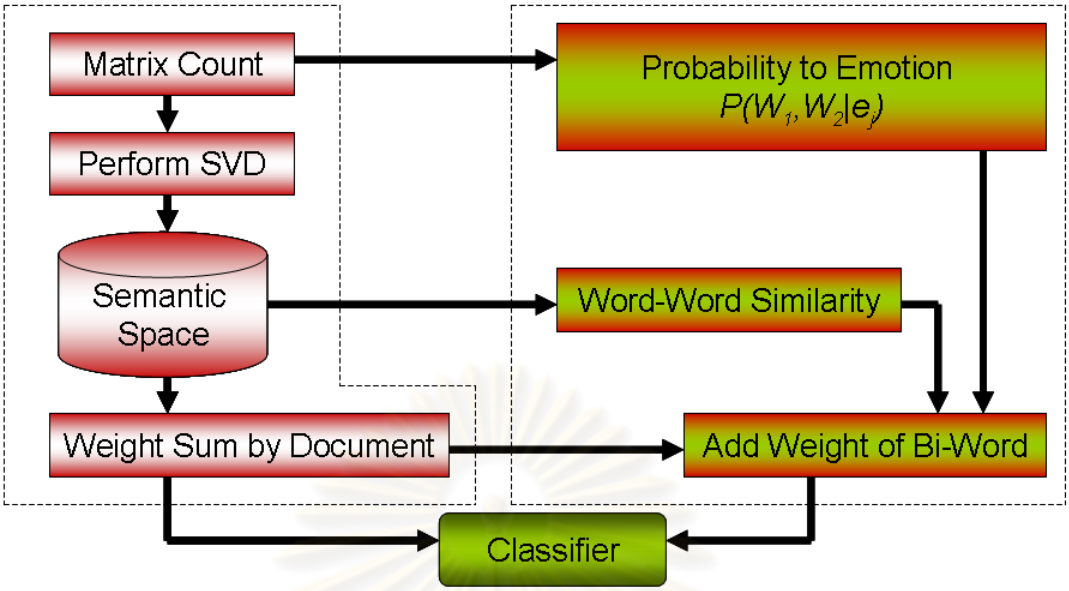
รูปที่ 6 แผนภาพรวมของระบบ ตัวแบบที่ 1

3.6 คำอธิบายตัวแบบที่ 2



รูปที่ 7 แผนภาพรวมของระบบ ตัวแบบที่ 2

ตัวแบบที่ 2 เป็นการประยุกต์ใช้การวิเคราะห์ความหมายแฝงกับการจำแนกประเภทอารมณ์ในข้อความภาษาไทย โดยสามารถเขียนแผนภาพการประยุกต์ได้ ดังรูปที่ 8 และสามารถอธิบายรายละเอียดได้ดังนี้



รูปที่ 8 การประยุกต์ใช้การวิเคราะห์ความหมายแฝงเพื่อการเปรียบเทียบการจำแนก

ส่วนที่เพิ่มจากตัวแบบที่ 1 คือการเพิ่มการพิจารณาคำที่มักปรากฏคู่กันด้วยการหาค่าความน่าจะเป็นของการเกิดขึ้นคู่กันของคำนามและคำกริยา แล้วส่งผลให้เกิดเป็นอารมณ์ใดๆ สามารถหาได้จากสมการที่ (14) ซึ่งในที่นี้จะไม่สนใจลำดับของการเกิดคู่กัน

$$P(e_j \mid W_1, W_2) = \frac{P(W_1, W_2 \mid e_j)P(e_j)}{P(W_1, W_2)}$$
$$= \frac{P(W_1 \mid e_j)P(W_2 \mid e_j)P(e_j)}{P(W_1, W_2)}$$

ซึ่งเมื่อพิจารณาโดยให้ความน่าจะเป็นของอารมณ์มีค่าเท่ากันและความน่าจะเป็นของคำมีค่าเท่ากันเพื่อประโยชน์ในการเปรียบเทียบ เราจะได้

$$\hat{P}(W_1, W_2 \mid e_j) = P(W_1 \mid e_j)P(W_2 \mid e_j) \qquad (14)$$

โดยที่ W_1 , W_2 หมายถึงคำที่ 1 แล ะ คำที่ 2 ตามลำดับ

e_j คืออารมณ์ใดๆ อันได้แก่ โกรธ (Anger) ขยะแขยง (Disgust) กลัว (Fear) เศร้า (Sadness) มีความสุข (Happiness) และ ประหลาดใจ (Surprise)

กล่าวคือความน่าจะเป็นของคำที่ 1 และคำที่ 2 ที่ส่งผลให้เกิดอารมณ์ใดๆ มีค่าเท่ากับ ผลคูณของความน่าจะเป็นของการเกิดขึ้นของอารมณ์นั้นกับความน่าจะเป็นของการที่คำทั้งสองส่งผลให้เกิดอารมณ์นั้น

ต่อมา หาผลคูณระหว่างค่าความน่าจะเป็นกับค่าความเหมือนระหว่างคู่คำ เพื่อนำมาใช้เป็นค่าน้ำหนักของคำคู่นั้น

จากนั้น นำแต่ละชุดข้อความแปลงให้อยู่ในรูปของเวกเตอร์ของชุดข้อความ *dj = [wj1, wj2, …, wj7110]* ซึ่งประกอบด้วยค่าน้ำหนักของแต่ละคำเดี่ยว ต่อด้วยค่าน้ำหนักของแต่ละคู่ของคำที่มักปรากฏคู่กัน และทำจับคู่กับอารมณ์สูงสุดที่ผู้อ่านเป็นผู้ระบุ ได้เป็นเวกเตอร์ใหม่ *dj = [wj1, wj2, …, wj7110, #emotion]* สำหรับนำไปเป็นชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบเปรียบเทียบกับตัวแบบที่ 1 ต่อไป

บทที่ 4
การทดลองและผลการทดลอง

4.1 การทดลอง

ผู้วิจัยทดลองโดยจัดเตรียมชุดข้อความจำนวน 150 ชุด และให้ผู้อ่านจำนวน 10 คนอ่านข้อความภาษาไทยและวิเคราะห์ให้ค่าอารมณ์ที่เหมาะสมกับข้อความนั้นมากที่สุดจากนั้นนำข้อความที่เตรียมไว้มาตัดคำด้วยโปรแกรมการตัดคำภาษาไทย (SWATH) โดยเลือกวิธีการตัดคำแบบการคำนวณเชิงสถิติของการเกิดของคำร่วมแบบ ไบแกรม กับลำดับของหน้าที่ของคำ เพื่อคัดเฉพาะคำที่มีบทบาทต่ออารมณ์ซึ่งได้แก่ คำกริยา คำนาม เท่านั้น ดังนี้

ผู้วิจัยขอแสดงตัวอย่างการทดลองกับชุดข้อความจริง จำนวน 4 ชุดข้อความ ที่ให้ผู้อ่านจำนวน 10 คนให้ค่าอารมณ์ต่าง ๆ แบ่งเป็นอารมณ์มีความสุข 2 ชุดข้อความ และเศร้า 2 ชุดข้อมูล ดังนี้

ชุดข้อความ 1 :

เมื่อเธอได้พบหน้าพ่ออันเป็นที่รักและคิดถึงมานานก็กลั้นน้ำตาไว้ไม่อยู่

อารมณ์จากผู้อ่าน : 8 คน ระบุอารมณ์ มีความสุข และ 2 คน ไม่แสดงความเห็น

การตัดคำ : พบ@VSTA|พ่อ@NCMN|รัก@VACT|คิดถึง@VACT|กลั้น@VACT|น้ำตา@NCMN|

ชุดข้อความ 2 :

วิวาห์หวานชื่นรับต้นปี รักล้นใจของคู่รักตลอดกาล

อารมณ์จากผู้อ่าน : 7 คน ระบุอารมณ์ มีความสุข และ 3 คน ไม่แสดงความเห็น

การตัดคำ : วิวาห์@NPRP|หวาน@VATT|ชื่น@VSTA|รัก@VACT|ล้น@VSTA|ใจ@NCMN|คู่รัก@NCMN

ชุดข้อความ 3 :

ฉันมันโง่เองแม้รักแท้ก็ดูแลไม่ได้

อารมณ์จากผู้อ่าน : 7 คน ระบุอารมณ์ เศร้า และ 3 คน ระบุอารมณ์ โกรธ

การตัดคำ : โง่@VSTA|รัก@VACT|แท้@VSTA|ดูแล@VACT|ไม่ได้@VSTA|

ชุดข้อความ 4 :

ฉันอกหักจากคนที่ฉันรักมากที่สุด ไม่อาจลืมความเจ็บปวด ความเสียใจคราวนี้ง่ายง่าย

อารมณ์จากผู้อ่าน : 8 คน ระบุอารมณ์ เศร้า 2 คน ระบุอารมณ์ โกรธ

การตัดคำ : อกหัก@VSTA|รัก@VACT|ลืม@VSTA|เจ็บปวด@VSTA|เสียใจ@VSTA|ง่าย@VACT|ง่าย@VACT|

ตารางที่ 14 ตารางแสดงความสัมพันธ์ของการปรากฏของคำในแต่ละชุดข้อความเทียบกับอารมณ์มากที่สุดจากผู้อ่าน

อารมณ์มากที่สุดที่ระบุโดยผู้อ่าน	มีความสุข	มีความสุข	เศร้า	เศร้า
	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4
พ่อ	1			
พบ	1			
รัก	1	1	1	1
คิดถึง	1			
กลั้น	1			
น้ำตา	1			
วิวาห์		1		
หวาน		1		
ชื่น		1		
ล้น		1		
ใจ		1		
คู่รัก		1		
โง่			1	
แท้			1	
ดูแล			1	
ไม่ได้			1	
อกหัก				1
ลืม				1
เจ็บปวด				1
เสียใจ				1
ง่าย				2

จากการนับการปรากฏของคำทั้งหมดใน 4 ชุดข้อมูล สามารถสรุปการหาความน่าจะเป็นของแต่ละคำที่จะเกิดขึ้นในแต่ละชุดข้อมูลได้ดังตารางที่ 15

ตารางที่ 15 ตารางค่าความน่าจะเป็นของคำที่ปรากฏทั้งหมดใน 4 ชุดข้อความ

	มีความสุข	เศร้า
พ่อ	1	0
พบ	1	0
รัก	0.5	0.5
คิดถึง	1	0
กลั้น	1	0
น้ำตา	1	0
วิวาห์	1	0
หวาน	1	0
ชื่น	1	0
ล้น	1	0
ใจ	1	0
คู่รัก	1	0
โง่	0	1
แท้	0	1
ดูแล	0	1
ไม่ได้	0	1
อกหัก	0	1
ลืม	0	1
เจ็บปวด	0	1
เสียใจ	0	1
ง่าย	0	1

ขั้นตอนการวิเคราะห์ความหมายแฝงตามอัลกอริทึมการแยกค่าแบบเดี่ยว ได้เป็นระนาบความหมาย ดังตารางที่ 15 และค่าความสัมพันธ์ระหว่างคำและคำบนระนาบความหมาย ดังตารางที่ 16

ตารางที่ 16 ตารางแสดงระนาบความหมายจาก 4 ชุดข้อมูลจริง

	ชุดข้อความ 1	ชุดข้อความ 2	ชุดข้อความ 3	ชุดข้อความ 4
พ่อ	0.514096	0.179822	0.589118	-0.59692
พบ	0.514096	0.179822	0.589118	-0.59692
รัก	0.243413	-0.40641	1.918881	0.305744
คิดถึง	0.514096	0.179822	0.589118	-0.59692
กลั้น	0.514096	0.179822	0.589118	-0.59692
น้ำตา	0.514096	0.179822	0.589118	-0.59692
วิวาห์	-0.18364	0.557736	0.571168	0.573562
หวาน	-0.18364	0.557736	0.571168	0.573562
ชื่น	-0.18364	0.557736	0.571168	0.573562
ล้น	-0.18364	0.557736	0.571168	0.573562
ใจ	-0.18364	0.557736	0.571168	0.573562
คู่รัก	-0.18364	0.557736	0.571168	0.573562
โง่	-0.63436	-0.53829	0.518893	-0.1964
แท้	-0.63436	-0.53829	0.518893	-0.1964
ดูแล	-0.63436	-0.53829	0.518893	-0.1964
ไม่ได้	-0.63436	-0.53829	0.518893	-0.1964
อกหัก	0.547323	-0.60567	0.239702	0.525495
ลืม	0.547323	-0.60567	0.239702	0.525495
เจ็บปวด	0.547323	-0.60567	0.239702	0.525495
เสียใจ	0.547323	-0.60567	0.239702	0.525495
ง่าย	1.094646	-1.21134	0.479404	1.05099

ตารางที่ 17 ตารางค่าความสัมพันธ์ระหว่างคำและคำบนระนาบความหมาย

คำที่ 1	คำที่ 2	ค่าความสัมพันธ์	ค่าความน่าจะเป็นของคำที่ 1 และ คำที่ 2 ที่ทำให้เกิดอารมณ์ใด ๆ	
			มีความสุข	เศร้า
พ่อ	เจ็บปวด	-3.331E-16	0	0
พ่อ	เสียใจ	-3.331E-16	0	0
พบ	รัก	1	0.25	0
รัก	หวาน	1	0.25	0
รัก	แท้	1	0.25	0
คิดถึง	น้ำตา	1	0.5	0
คิดถึง	อกหัก	-7.216E-16	0	0
น้ำตา	อกหัก	-3.886E-16	0	0
น้ำตา	เสียใจ	-3.886E-16	0	0
วิวาห์	คู่รัก	1	0.5	0
วิวาห์	อกหัก	-3.331E-16	0	0
ล้น	ใจ	1	0.5	0
เจ็บปวด	เสียใจ	1	0	0.5
อกหัก	เจ็บปวด	1	0	0.5
รัก	เจ็บปวด	1	0	0.25

เมื่อจับคู่ความหมายบนระนาบกับคลาสอารมณ์ สร้างเป็นชุดข้อมูลสำหรับการฝึกฝนและทดสอบเปรียบเทียบการจำแนกอารมณ์จากข้อความระหว่างตัวแบบที่ 1 และตัวแบบที่ 2 ดังนี้

ตัวแบบที่ 1

เวกเตอร์ตัวแทนของข้อความที่ 1 = { 0.3448896,0.4241862,0.2905709,0.7852934, #มีความสุข }

เวกเตอร์ตัวแทนของข้อความที่ 2 = { -0.3334381,-0.6919544,-0.2196387,0.6014789, #มีความสุข }

เวกเตอร์ตัวแทนของข้อความที่ 3 = { 0.7844740,-0.5614499,0.2309159,-0.1266980, #เศร้า }

เวกเตอร์ตัวแทนของข้อความที่ 4 = { 0.3930276,0.1613664,-0.9022224,0.0740602, #เศร้า }

ตัวแบบที่ 2

เวกเตอร์ตัวแทนของข้อความที่ 1 = { 0.3448896,0.4241862,0.2905709,0.7852934, 0.5, 0.5, 0, 0, #มีความสุข }

เวกเตอร์ตัวแทนของข้อความที่ 2 = { -0.3334381,-0.6919544,-0.2196387,0.6014789, 0.5, 0.5, 0, 0, #มีความสุข }

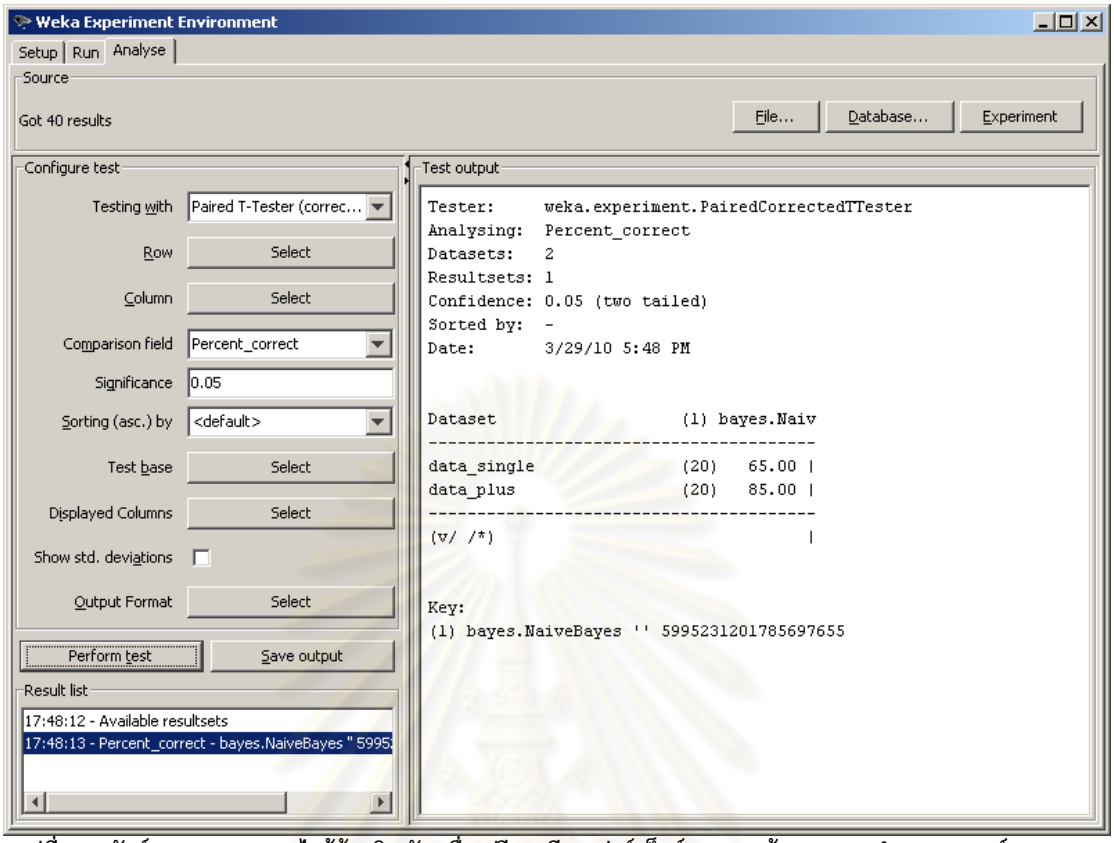
เวกเตอร์ตัวแทนของข้อความที่ 3 = { 0.7844740,-0.5614499,0.2309159,-0.1266980, 0, 0, 0.5, 0.5, #เศร้า }

เวกเตอร์ตัวแทนของข้อความที่ 4 = { 0.3930276,0.1613664,-0.9022224,0.0740602, 0, 0, 0.5, 0.5, #เศร้า }

ทดสอบแบบไขว้ข้ามสิบพับ ด้วยค่านัยสำคัญเป็น 0.05 เพื่อหาเปอร์เซ็นต์ความถูกต้องของการจำแนกอารมณ์ของข้อมูล 4 ชุดข้อความ โดยเปรียบเทียบระหว่างตัวแบบที่ 1 ซึ่งให้ความถูกต้องคิดเป็น 65% และตัวแบบที่ 2 ซึ่งให้ความถูกต้องคิดเป็น 85% แสดงผลลัพธ์จากโปรแกรม เวกา (Weka) ดังรูปที่ 9

4.2 ผลการทดลอง

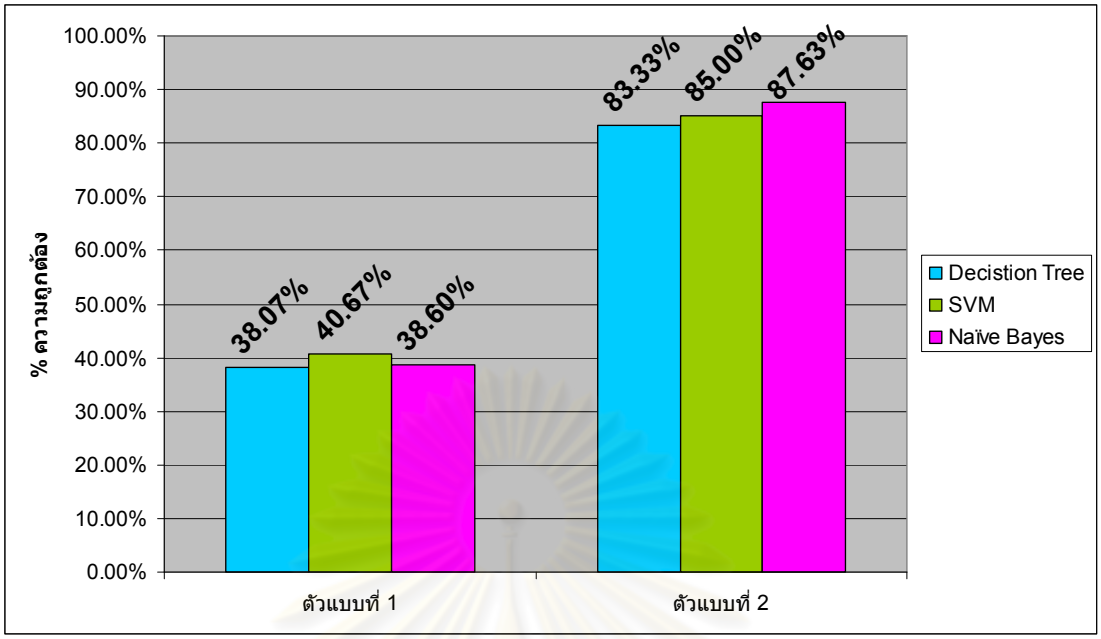
จากการทดลองเปรียบเทียบความถูกต้องของตัวแบบที่ 1 และตัวแบบที่ 2 กับข้อมูลจำนวน 150 ชุดข้อมูล ซึ่งแจกแจงจำนวนคลาสที่ผู้อ่านระบุได้ดังตารางที่ 18 และผลการทดลองแสดงดังรูปที่ 10 พบว่าตัวแบบที่ 2 ให้ผลการจำแนกอารมณ์ได้ดีกว่าตัวแบบที่ 1 ในทุกวิธีการการจำแนก กล่าวคือ การพิจารณาคำคู่ที่มักปรากฏคู่กันมีส่วนช่วยให้การจำแนกอารมณ์จากข้อความภาษาไทยให้เปอร์เซนต์ความถูกต้องสูงขึ้น



รูปที่ 9 ผลลัพธ์การทดสอบแบบไขว้ข้ามสิบพับ เพื่อเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของการจำแนกอารมณ์ของ 4 ชุดข้อความด้วยตัวแบบที่ 1 และตัวแบบที่ 2

ตารางที่ 18 ข้อมูลสำหรับการทดลอง

คลาส	โกรธ	กลัว	มีความสุข	เศร้า	ประหลาดใจ	ขยะแขยง
จำนวนชุดข้อมูล	15	15	35	61	14	10
ค่าความน่าจะเป็น	0.1	0.1	0.233	0.407	0.093	0.067



รูปที่ 10 ผลการทดลองเปรียบเทียบความถูกต้องของตัวแบบที่ 1 และตัวแบบที่ 2

เงื่อนไขการปรับแต่งค่าพารามิเตอร์ของ 3 วิธีการเพื่อใช้ในการฝึกฝนและทดสอบด้วยโปรแกรมเวกา มีดังรูปที่ 11-13

(1) ต้นไม้ตัดสินใจ เลือกตัวจำแนกชื่อ J48 ด้วยเงื่อนไขดังนี้

classifier : weka.classifiers.trees.J48

binarySplits : False

confidenceFactor : 0.25

debug : False

minNumObj : 2

numFolds : 3

reducedErrorPruning : False

saveInstanceData : False

seed : 1

subtreeRaising : True

unpruned : False

useLaplace : False

(2) เครื่องจักรเวกเตอร์สนับสนุน เลือกตัวจำแนกชื่อ SMO เพื่อการจำแนกคลาสที่มากกว่า 2 คลาสข้อมูล ด้วยเงื่อนไขดังนี้

classifier : weka.classifiers.functions.SMO

buildLogisticModels : False

c : 1.0

checksTurnedOff : False

debug : False

epsilon : 1.0E-12

filterType : Normalize training data

kernel : PolyKernel -C 250007 -E 1.0

numFolds : -1

randomSeed : 1

toleranceParameter : 0.0010

(3) นาอีฟเบส์ เลือกตัวจำแนกชื่อ NaiveBayes ด้วยเงื่อนไขดังนี้

classifier : weka.classifiers.bayes.NaiveBayes

debug : False

displayModelInOldFormat : False

useKernelEstimator : False

useSupervisedDiscretization : False

หากทดลองปรับค่า k ระหว่าง 2 ถึง 10 เพื่อหาค่าที่เหมาะสมที่สุดกับการจำแนกอารมณ์จากข้อความภาษาไทยบนตัวแบบที่ทดลอง พบว่า

- ตัวแบบที่ 1 เมื่อทดลองปรับค่า k ที่ต่างกัน จะให้ค่าความถูกต้องที่ต่างกัน ในขณะที่ค่า k มีค่ามากขึ้น ทำให้ความถูกต้องในการจำแนกมีมากขึ้นด้วย เนื่องจาก k มีผลต่อขนาดของระนาบความหมาย ที่จะเป็นตัวกำหนดชี้วัดความหมายที่สะท้อนถึงอารมณ์ของคำได้
- ตัวแบบที่ 2 ตัวแบบที่ 2 ไม่มีผลต่อการปรับค่า k เนื่องจากการประยุกต์ใช้ระนาบความหมายได้ทำหลังจากการปรับค่า k ซึ่งเสร็จสิ้นไปแล้ว

สามารถแสดงผลการทดลองได้ดังตารางที่ 19

ตารางที่ 19 ผลการทดลองเปรียบเทียบความถูกต้องเมื่อปรับค่า k ระหว่าง 2 ถึง 10

ประมาณค่า k	ตัวแบบที่ 1	ตัวแบบที่ 2
2	38.60%	87.63%
3	38.53%	83.75%
4	51.87%	86.80%
5	53.47%	87.47%
6	55.80%	86.54%
7	62.80%	86.79%
8	63.00%	86.72%
9	63.87%	86.86%
10	64.53%	86.60%

เมื่อเปรียบเทียบความถูกต้องของการจำแนกอารมณ์ด้วยตัวแบบที่ 1 และตัวแบบที่ 2 พบว่าคลาสอารมณ์ที่มีผลการจำแนกที่ถูกต้องที่สุด คือ อารมณ์เศร้า ดังแสดงในตารางที่ 20

ตารางที่ 20 ผลการทดลองเปรียบเทียบความถูกต้องในการจำแนกข้อความออกเป็นแต่ละคลาสอารมณ์

คลาสอารมณ์	ตัวแบบที่ 1	ตัวแบบที่ 2
โกรธ	33.33%	40%
ขยะแขยง	20%	20%
กลัว	6.7%	13.33%
มีความสุข	74.30%	71.40%
เศร้า	82%	91.20%
ประหลาดใจ	42.90%	57.10%

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ นำเสนอตัวแบบการจำแนกอารมณ์จากข้อความภาษาไทยเป็น 6 อารมณ์สากล ได้แก่ โกรธ ขยะแขยง กลัว มีความสุข เศร้า และประหลาดใจ โดยอาศัยการวิเคราะห์ความหมายแฝงของคำนามและคำกริยาในประโยค ผู้วิจัยเปรียบเทียบผลลัพธ์ของ 2 ตัวแบบ โดยตัวแบบที่หนึ่งใช้การจำแนกโดยการวิเคราะห์ความหมายแฝงของคำเดี่ยว ส่วนตัวแบบที่สองใช้การประยุกต์การวิเคราะห์ความหมายแฝงของคำคู่ที่มักปรากฏคู่กันร่วมกับระนาบความหมายของคำเดี่ยว และประยุกต์ใช้กับ 3 ระเบียบวิธีได้แก่นาอีฟเบส์, เครื่องจักรเวกเตอร์สนับสนุน และต้นไม้ตัดสินใจ ผลการเปรียบเทียบผลลัพธ์แสดงให้เห็นว่า ตัวแบบที่สองให้ความถูกต้องได้สูงกว่าตัวแบบที่หนึ่ง อ้างอิงจากระเบียบวิธีการจำแนกของนาอีฟเบส์ที่ให้ผลสูงกว่าระเบียบวิธีการอื่น

5.2 ข้อเสนอแนะสำหรับการวิจัยต่อ

งานวิจัยนี้ เป็นการนำเสนอตัวแบบที่ใช้หลักการในการวิเคราะห์ความหมายแฝงของคำในข้อความภาษาไทยเพียง 150 ข้อความเท่านั้น จึงสามารถฝึกฝนเพิ่มเติมประโยคให้กับระบบได้เรียนรู้ประโยคเพิ่มเติมได้ อีกทั้งยังไม่ครอบคลุมถึงการวิเคราะห์คำที่เขียนเหมือนกันแต่มีความหมายต่างกัน (Synonyms) และคำที่มีหน้าที่อื่นในประโยคภาษาไทย เช่น คำขยายคำนาม คำขยายคำกริยา เป็นต้น ทั้งนี้สามารถเป็นประเด็นที่สามารถจะวิจัยต่อไปได้

นอกจากนี้ สามารถนำผลลัพธ์ที่ได้จากระบบนี้ไปพัฒนาร่วมกับระบบอื่นได้อย่างมีประสิทธิภาพ เช่น ระบบการแปลงข้อความเป็นเสียง (Text to Speech) ในการเล่านิทานภาษาไทย เพื่อการปรับโทนเสียงให้สอดคล้องกับอารมณ์ที่เกี่ยวข้องกับข้อความได้

รายการอ้างอิง

[1] Yan, J., Bracewell, D.B., Ren, F. and Kuroiwa, S. The Creation of a Chinese Emotion Ontology Based on HowNet. Journal of the International Association of Engineers 16,1 (February 2008): 166-171.

[2] Gill, A.J., Gergle, D., French, R.M. and Oberlander, J. Emotion Rating from Short Blog. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008), pp.1121-1124. Florence, Italy, 2008.

[3] Jung, Y., Park, H., and Myaeng, S.H. A Hybrid Mood Classification Approach for Blog Text. Proceedings of 9th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2006: Trends in Artificial Intelligence, pp.1099-1103. Guilin, China, 2006.

[4] Danisman, T., Alpkocak, A. Feeler: Emotion Classification of Text Using Vector Space Model. Proceedings of AISB 2008 Convention, Communication, Interaction and Social Intelligence ("Affective Language in Human and Machine"), pp.53-59. Aberdeen, UK, 2008.

[5] Steidl, S., Levit, M., Batliner, A., Nöth, E. and Niemann, H. OF ALL THINGS THE MEASURE IS MAN AUTOMATIC CLASSIFICATION OF EMOTIONS AND INTER-LABELER CONSISTENCY. Proceedings of ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, pp.317-320. Philadelphia, USA, 2005.

[6] Kozareva, Z., Navarro, B., Vazquez, S. and Montoyo A. UA-ZBSA: A Headline Emotion Classification through Web Information. Proceedings of the 4th International Workshop on Semantic Evaluations, pp.334-337. Prague, Czech Republic, 2007.

[7] Deerwester, S., Dumais, S.T., Harshman, R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41 (September 1990): 391-407.

[8] Strang, G. Introduction to linear algebra. Wellesley, MA : Wellesley-Cambridge Press, 1998.

[9] Department of Computer Science, The University of Tennessee. Latent Semantic Indexing [Online]. 2008. Available from : http://www.cs.utk.edu/~berry/lsi++/node5.html [2010, February 22].

[10] Ding, C. A Probabilistic Model for Latent Semantic Indexing. Journal of the American Society for Information Science 56, 6 (April 2005): 597-608.

[11] Cohen, W., Carnegie Mellon University. Text classification [Online]. 2007. Available from : http://videolectures.net/mlas06_cohen_tc/ [2010, February 25].

[12] Thangthai, A. and Jaruskulchai, C. Impact Parameter on LSA Performance for Thai Text Summarization. Kaset Vichakarn'43, Bangkok, Thailand. 2004.

[13] Meknavin, S., Charoenpornsawat, P. and Kijsirikul, B. Feature-based Thai Word Segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997(NLPRS'97), pp.41-48. Phuket, Thailand. 1997.

[14] Strapparava, C., Mihalcea, R. Learning to Identify Emotions in Text. Proceedings of the 2008 ACM Symposium on Applied Computing, Session: Natural Language Processing and Speech Recognition, pp.1556-1560. Fortaleza, Ceara, Brazil, 2008.

[15] Mishne, G. Experiments with Mood Classification in Blog Posts. Proceedings of the 1st Workshop on Stylistic Analysis Of Text For, pp.526-558. Belgium, 2005.

[16] Yang, C., Lin, K.H. and Chen, H. Emotion Classification Using Web Blog Corpora. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.275-278. Washington, DC, USA, 2007.

[17] Strapparava, C., Mihalcea, R. Sem-Eval 2007 Task 14: Affective Text. Proceedings of the 4th International Workshop on Semantic Evaluations, pp.70-74. Prague, Czech Republic, 2007.

[18] Yang, C., Lin, K.H. and Chen, H. Building Emotion Lexicon from Weblog Corpora. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, International Conference on Computational Linguistics, pp.133-136. Prague, Czech Republic, 2007.

[19] Sugimoto, F. and Masahide, Y. A Method for Classifying Emotion of Text Based on Emotional Dictionaries for Emotional Reading. Proceedings of the 24th IASTED

International Conference on Artificial Intelligence and Applications, pp.91-96. Austria, 2006.

[20] Alm, C.O., Roth, D. and Sproat, R. Emotions from text: machine learning for text-based emotion prediction. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp.347-354. Illinois, USA, 2005.

[21] Zhe, X. and Boucouvalas, A. Text to Emotion Engine for Real Time Internet Communication. Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications, pp.164-168. UK, 2006.

[22] Lin, K.H., Yang, C. and Chen, H.H. Emotion Classification of Online News Articles From the Reader's Perspective. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp.220-226. USA, 2008.

[23] Carnegie Mellon University. Text Classification [Online]. 2008. Available from : http://videolectures.net/mlas06_cohen_tc [2010, February 10].

[24] Mitchell, T.M. Machine Learning. Singapore, McGraw Hill, 1997.

[25] Gill, A.J., French, R.M., Gergle, D. and Oberlander, J. The Language of Emotion in Short Blog Texts. Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2008), pp.299-302. San Diego, CA, 2008.

[26] Charoenporn, T., Kruengkrai, C., Sornlertlamvanich, V., Isahara, H. and Theeramunkong, T. Corpus-based Dictionary Development System. Proceedings of the 4th Workshop on Asia Language Resource, IJCNLP-04, pp.47-53. Sanya City, Hainan Island, China. 2004.

[27] Lin, H., Yang, C. and Chen, H. What Emotions do News Articles Trigger in Their Readers. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.733-734. Amsterdam, The Netherlands. 2007.

[28] Leshed, G. and Kaye, J. Understanding How Bloggers Feel-Recognizing Affect in Blog Posts. The CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp.1019-1024. Quebec, Canada. 2006.

[29] Yu, B. and Unsworth, J. An Evaluation of Text Classification Methods for Literary Study. The Literary and Linguistic Computing 23(3), pp.327-343. USA. 2008.

[30] Jones, M.P. and Martin, J.H. Contextual Spelling Correction Using Latent Semantic Analysis. Proceedings of the 5th Conference on Applied Natural Language Processing, pp.166-173. Washington, DC. 1997.

[31] Polpinij, J., Sibunruang, C., Chamchong, R., Chotthanom, A., Puangpronpitag, S. Content-based Probabilistic Text Classifier for Pornographic Web Filtering. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pp.272-284. Taipei, Taiwan. 2006.

[32] Bacan, H., Pandzic, I.S., Gulija, D. Automated News Item Categorization. Proceedings of the 19th Annual Conference of the Japanese Society for Artificial Intelligence, pp.251-256. Faculty of Electrical Engineering and Computing, University of Zagreb. 2005.

[33] Sornlertlamvanich, V., Charoenporn, T. and Isahara, H. ORCHID: Thai Part-Of-Speech Tagged Corpus. Technical Report Orchid TR-NECTEC-1997-001, National Electronics and Computer Technology Center, pp.509-512. Thailand, 1997.

ภาคผนวก

ภาคผนวก ก.
ผลงานตีพิมพ์จากงานวิจัย

บทความทางวิชาการเรื่อง “APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFY EMOTIONS IN THAI TEXT” โดย ปิยธิดา อินทร์รักษ์ และ สุกรี สินธุภิญโญ ในงานประชุมวิชาการระดับนานาชาติ “2010 The 2nd International Conference on Computer Engineering and Technology (ICCET 2010)” ซึ่งจัดขึ้น ณ เมืองเฉิงตู ประเทศสาธารณรัฐประชาชนจีน ระหว่างวันที่ 16-18 เมษายน พ.ศ. 2553

APPLYING LATENT SEMANTIC ANALYSIS TO CLASSIFY EMOTIONS IN THAI TEXT

Piyatida Inrak
Department of Computer Engineering
Faculty of Engineering, Chulalongkorn University
Bangkok, Thailand
E-mail: piyatida.inrak@gmail.com

Sukree Sinthupinyo
Department of Computer Engineering
Faculty of Engineering, Chulalongkorn University
Bangkok, Thailand
E-mail: sukree@cp.eng.chula.ac.th

***Abstract*—**With a rapid growth of the internet communication, many types of text are produced. They can convey the meanings that can contribute to text categorization. Emotion classification also becomes more interesting, but emotion classification in Thai text is still not able to be correctly classified. Thus, this paper proposes a novel approach that takes advantage of bi-words occurrence to classify emotion hidden in a short sentence. In this paper, we classify Thai text into six basic universal emotions including anger, disgust, fear, happiness, sadness, and surprise based on latent semantic analysis approach. We compared the results between two models which construct features from the sentences and applied to three classification methods, i.e. Naïve Bayes, SVM, and Decision Tree. The first feature model uses only single word occurrence in the classification. The second model uses single word combined with bi-words occurrence in the classification. The results show that the second model can yield higher accuracy than the first model based on the Naïve Bayes classification method.

Keywords-Emotions in Text, Latent Semantic Analysis, Affective Computing

I. INTRODUCTION

An emotion is a mental state or a feeling that occurs under subconscious of individual. Emotions are subjective experiences, often associated with mood, temperament, personality, and disposition [6]. Since emotions can naturally occur to a person, they are uncontrollable and difficult to manipulate and measure [1]. However, there are many experiments that pay attention to emotions, whether to create a dictionary of words with emotional property [6, 9, 10] or emotions classification in music, facial movements, and hand gesture [8] which are known to be classified by physical characteristics and that cannot be used to classify emotion in text.

The motivation of this research is that Thai text has never been classified emotion using latent semantic analysis before. There is no emotion dictionary for Thai words. Essentially emotion in Thai text can be beneficial to automatically match with appropriate tones in text to speech system so that the communications convey greater realism.

We introduce latent semantic analysis (LSA) to perform emotion classification on Thai text which can be collected from the internet (emails, blogs, topics, boards). This paper is aimed to focus on the comparison between two ideas to classify emotions in Thai text by considering of single word and bi-words. Therefore basic ideas and assumptions of using bi-words to better indicate the emotion for the sentence is proposed because we believe that bi-words can capture more feeling from the context rather than using single word. The method of getting pair of words is also discussed in this paper.

The rest of the paper is organized as follows. LSA is introduced in section II. Section III describes the basic of emotion classification and the text classification, and how to apply the basic features to the Thai text. In Section IV, we discuss the method to classify emotion in Thai text. Finally, the experiment results and the conclusions are given in the section V, respectively.

II. LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis (LSA) is a powerful method using statistics and linear algebra to extract the meaning of words from text. The texts can be compared to measure how similar they are to the others, resemble human judgments. LSA represents the relationship between words and the text passages in a high dimensional matrix. Then apply to the Singular Value Decomposition (SVD) method to transform the text passage vector into the form of subspace. The result can produce the semantic space, so that we can compute for the semantic similarity of the text.

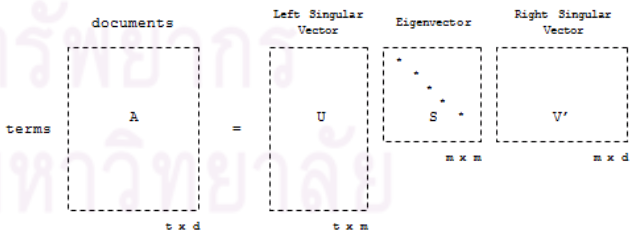


Figure 1. Mathematical representation of matrix A

Given a word-passage matrix A, which represents the correlation between each word to the whole set of text collection; can be written in a form of equation as in (1).

$$A = USV^T \tag{1}$$

Where U and V^T are orthogonal matrices and S is a diagonal matrix [12]. The decomposition of three matrices is called singular value decomposition (SVD):

$$\hat{X} = U_k S_k V_k^T \tag{2}$$

The appropriate k value is selected to reduce the matrix dimension, in order to remove the noise from the semantic space. The first k columns of matrix U are selected as U_k. The first k rows of matrix V^T are selected as V_k^T, and the k factors of the diagonal elements are selected as S_k. The matrix $\hat{X}$ is considered to capture the most important relationship between words and sentences [13].

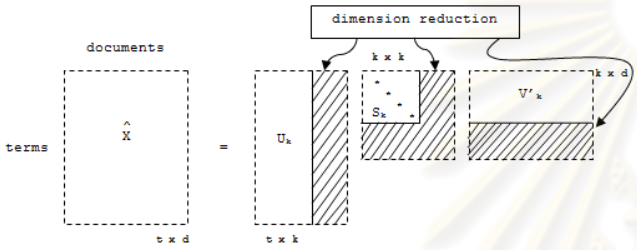


Figure 2. Mathematical representation of the matrix

The semantic space denoted as matrix $\hat{X}$ can be interpreted using the Spearman's Rank Correlation Coefficient approach which is considered as the appropriate method of ranking order with the relative judgments against other similar objects [7]. The result value in the semantic space is between -1 and 1 can be categorized as three relationships: the value closed to -1 is considered to be a negative correlation, the value closed to 0 is considered to be no linear correlation, and the value closed to +1 is considered to be a positive correlation.

In order to create the training features, we compute the semantic vector for each document which can be represented in a vector by using Equation (3). Each document contains the semantic elements feature which can be mapped to the emotion labeled from the readers. [14]

$$\hat{d} = d^T U_k S_k^{-1} \tag{3}$$

Moreover, when a new set of words need to be appended to the existing semantic space, they can represent a weighted sum vectors for documents that words appear as in (4).

$$\hat{t} = t V_k S_k^{-1} \tag{4}$$

Once a new semantic vector $\hat{t}$ of words is folded-in, the semantic space can be recomputed to get the new value of document vector that can be mapped to the emotion labeled as well.

III. Basic Features

To classify emotion in Thai text, there are 3 features need to be considered.

(1) *Thai Text:* A sentence consists of words, while a word can convey the emotion. Therefore to analyze the emotion in text, the sentence is separated into a singular word. But in Thai text, there are no spaces between words; instead spaces in a Thai text indicate the end of a clause or sentence. So it can cause more difficult on getting the appropriate word for analysis. In the experiment, the SWATH (Smart Word Analysis for THai) is used to perform the word segmentation and the part of speech can be tagged in the given sentences [14].

(2) *Term Frequency - Inverse Document Frequency (TF-IDF):* The weight calculation for each word using TF-IDF approach, to evaluate how important a word is to a sentence. The term frequency is the number of times which a given term appears in the sentence.

$$w_{ij} = tf_{ij} \times \log \frac{N}{1 + n_j} \tag{5}$$

, where
wij is the weight of term *Ti* in the sentence *Sj*,
tfij is the frequency of term *Ti* in the sentence *Sj* ,
N is the number of sentences, and
nj is the number of sentences that term *Ti* appears in at least once.

This count is usually normalized to prevent a bias towards longer sentences. The inverse document frequency is to measure the general importance of the term. Therefore a high weight in TF-IDF is reached by a high term frequency and a low document frequency of the term in the whole set of sentences.

(3) *Text Classification:* In our experiment we focus on both nouns and verbs to extract the affective words from the sentences. A large semantic space has been produced from the section of latent semantic analysis. We need to break data into 10 sets and perform the standard 10-fold cross validation. One round of cross validation performs the analysis on each training subset. We train nine sets of semantic spaces, and test one remaining set. Then, the test process is repeated ten times on each subset. Hence, all subsets can be tested by using the remaining nine subsets as the training set.

Even though, Naïve Bayes and SVM are proved to be the most appropriate classifier in English text classification [9]. In our experiment, we compared the result from three methods including Naïve Bayes, SVM, and Decision Tree to find the most appropriate classifier for Thai text. In our experiment, we employ Weka as a tool to train and test different machine learning algorithms [15].

(4) *Emotion Classification:* The emotion detection in texts becomes more interesting and important in the field of computational linguistics. In fact, emotions can be considered in an affective analysis which can be contributed to sentiment analysis, computer assisted creativity, and verbal expressivity in human computer interaction [4]. In this paper, we map the semantic document vector into six basic universal emotions including anger, disgust, fear, happiness, sadness, and surprise [11]. The part of speech is one of the useful factors that help on filtering the unemotional words. Some papers support the idea that nouns can effectively convey the emotion, while Verbs can represent the affective events in the sentences. [5, 6]

IV. METHODOLOGY

In our research, we propose the algorithm to classify emotions in Thai text, which consists of four main steps: (1) Collecting Thai Text (2) Performing LSA (3) Mapping Words and Emotions and (4) Classifying Emotion. Besides, the two different models of using only single word and the single word combined with bi-word have been introduced and can be shown in Fig. 3.

A. Collecting Thai Text

1) At first, we collect 200 Thai texts from the internet including emails, blogs, topics, and boards. The natural language is simply used in the source of text, all emoticons and typographical symbols are ignored.

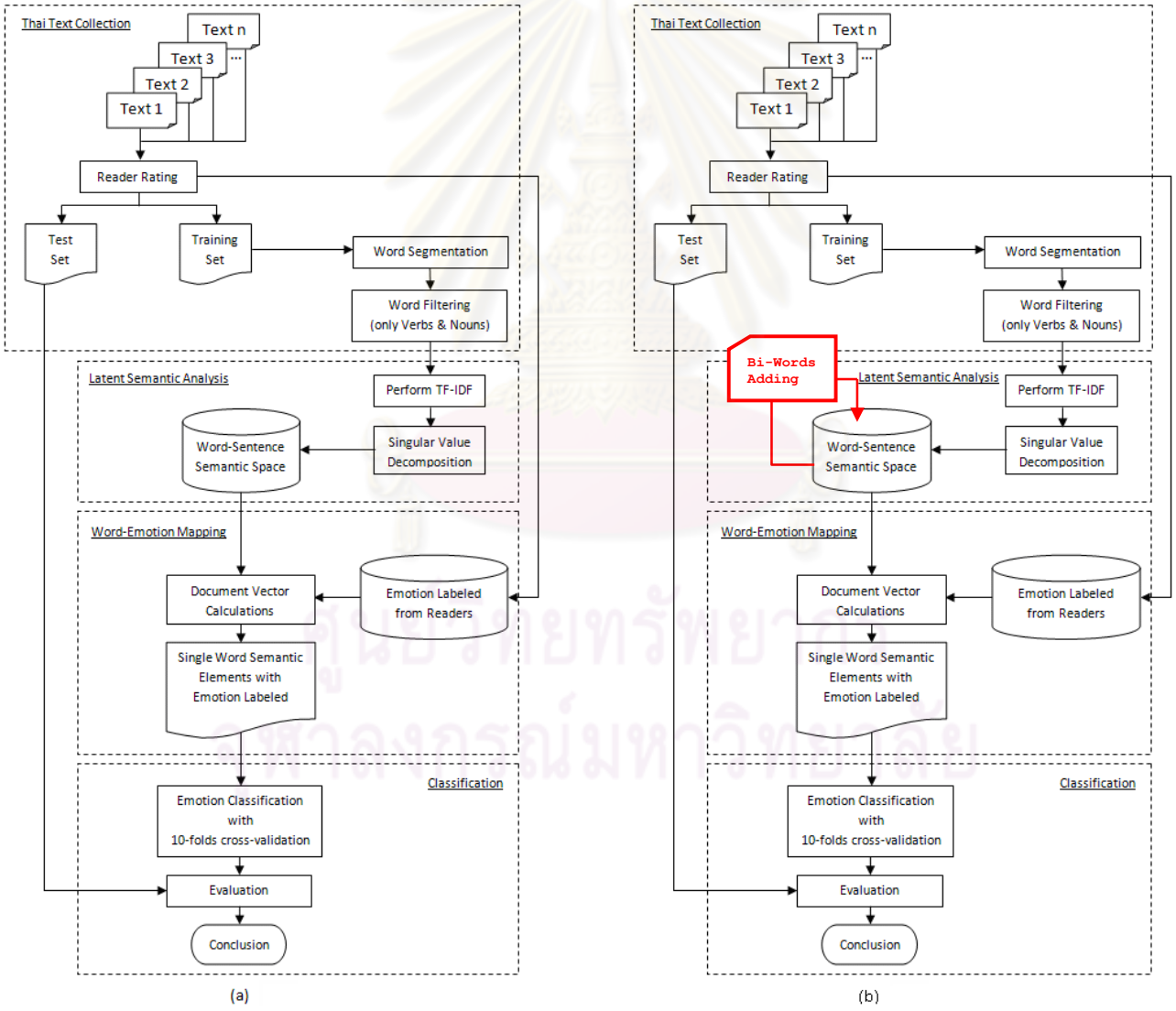


Figure 3. The system overview of two proposed algorithms to classify emotion in Thai Text based on Latent Semantic Analysis. (a) Model #1 Single Word Method (b) Model #2 Single Word combined with Bi-Words Method

2) Then 10 readers are asked to label the best emotion that can be exposed from the text.

3) SWATH is used to separate Thai words and assign the part of speech in Fig. 4, since the Thai writing style is usually written without boundaries.

4) The words filtering is performed by selecting only nouns and verbs indicated in the part of speech tagged that start with N and V [14] in Fig. 4. which have been proved that they contribute the affected words.

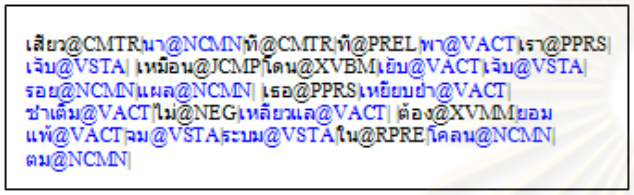


Figure 4. The output of Thai words segmen-tation from program SWATH

B. Performing LSA

The affected words is now prepared in a form of Bag of Words (BOW).

1) We count all the appearances of words and create the matrix between words and sentences.
2) Performing TF-IDF for each word in the matrix.
3) Applying SVD algorithm to the occurrence matrix of words in the sentences. We then get the result of semantic space that represents the semantic value of correlation between words and sentences.

C. Mapping Words-Emotions

1) In the semantic space, we calculate the semantic vector for each sentence using Equation (3).
2) Each sentence can be mapped to the emotion with maximum labeled from the readers.
3) The set of semantic vectors and emotion tagged is used as the training set. This can be the input to the next step of classification method.

D. Classifying Emotion

From the previous step, the training sets contain semantic values of words and emotions are ready. In this experiment, we apply the standard 10-fold cross validation to test the dataset using Naïve Bayes, SVM and Decision Tree algorithm.

E. Two Models Definitions

We introduced two models of classification. The first model uses only single word and the second model uses the single word combined with bi-word.

Model #1: Single Word Method

Each document vector is computed from document representation, dj = [wj1, wj2, …, wj489] which contains the elements of single word semantic values. After mapping emotion labeled to the sentences, the set of semantic values and emotion becomes the training set of the system.

$$\hat{d} = [w_{s1}, w_{s2}, ..., w_{sk}, \#emotion]$$

Model #2: Single Word combined with Bi-Words Method

Bi-word is made from two individual words that appear in the sentence. The sequence of them can be ignored. This simple idea directly exposes bi-word to capture more contextual semantic of both nouns and verbs in the sentences. Bi-words are counted and computed for maximum probability in which emotion that they can contribute to the whole training set as in (6). Let W1 and W2 are bi-word with conditionally independent [17] given emotions E.

$$P(W_1, W_2 \mid e_j) = P(e_j)P(W_1 \mid e_j)P(W_2 \mid e_j) \qquad (6)$$

Here we get the updated semantic space that can represent a weighted sum vectors for documents that bi-words appear in Fig. 5.

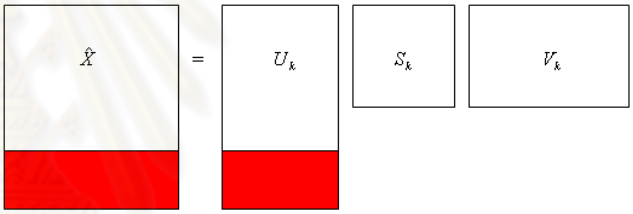


Figure 5. Mathematical representation of folding in bi-words to the existing semantic space

Once a new semantic vector of bi-words is folded-in, the semantic space can be recomputed to get the new document vector *dj = [wj1, wj2, ..., wj5188]* that can be mapped to the emotion labeled from the readers as same as practices with single words.

V. Experimental Results

As described in previous section, our experiment setting presents the comparison between two methods of word-emotion relationship. The first model is to classify text using single word, and the second model is to classify text using single word combined with bi-words. The percentage of accuracy obtained from the proposed models and classifiers are presented in table 1. The result indicates that Naïve Bayes yielded the highest accuracy in emotion classification upon the two models.

TABLE I. Percentage of Accuracy in Emotion Classification

Classifier	Single Word Model	Single Word + Bi Words Model
Decision Tree	48.9038 %	84.6667 %
SVM	50.6667 %	87.3333 %
Naïve Bayes	62.6667 %	90 %

Therefore, we focus on improving the classification method using Naïve Bayes, and continue on different k-rank to find the appropriate model of classification. The results of accuracy against different k-rank are shown in Table II.

TABLE II. PERCENTAGE OF ACCURACY IN DIFFERENT K-RANK

k-rank	Single Word Model	Single Word + Bi Words Model
2	52.6667 %	91.3333 %
5	62.6667 %	90 %
10	33.3333 %	74 %

VI. CONCLUSIONS

The proposed models of emotion classification in Thai text in this paper can classify Thai text into six basic universal emotions including anger, disgust, fear, happiness, sadness, and surprise based on latent semantic analysis of nouns and verbs in the sentences. We compared the results between two models, i.e. single word and single word combined with bi-word. This can extract the emotion from the sentences which are then compared with the result labelled by the readers.

In summary, the experimental results show the second model can yield more accuracy than the first model. Moreoever, Naïve Bayes method is proved in our experiment that it is the best classifier comparing to SVM and Decision Tree for emotion classification in Thai text.

.

REFERENCES

[1] A. J. Gill, D. Gergle, R. M. French, and J. Oberlander. Emotion Rating from Short Blog. CHI 2008 Proceedings. 2008.

[2] A. Thangthai and C. Jaruskulchai. Impact Parameter on LSA Performance for Thai Text Summarization. Kaset Vichakarn'43, Bangkok, Thailand. .2004

[3] B. Yu, J. Unsworth. An Evaluation of Text Classification Methods for Literary Study. Digital Humanities Conference. .2007

[4] C. Strapparava, R. Mihalcea. Learning to Identify Emotions in Text. Proceedings of the 2008 ACM Symposium on Applied computing. 2008.

[5] F. Sugimoto and Y. Masahide. A Method for Classifying Emotion of Text Based on Emotional Dictionaries for Emotional Reading. Proceedings of the ACM International Conference. 2006.

[6] J. Yan, D. B. Bracewell, F. Ren, and S. Kuroiwa. The Creation of a Chinese Emotion Ontology Based on HowNet. Advance online publication.2008.

[7] S. Deerwester, S. T. Dumais, R. Harshman. Indexing by Latent Semantic Analysis. Journal of the Society for Information Science. 1990.

[8] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. OF ALL THINGS THE MEASURE IS MAN AUTOMATIC CLASSIFICATION OF EMOTIONS AND INTER-LABELER CONSISTENCY. ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing. 2005.

[9] T. Danisman, A. Alpkocak. Feeler: Emotion Classification of Text Using Vector Space Model. In AISB 2008 Convention, Communication, Interaction and Social Intelligence, vol.2 ("Affective Language in Human and Machine"). 2008.

[10] Y. Jung, H. Park, and S. H. Myaeng. A Hybrid Mood Classification Approach for Blog Text. Lecture Notes in Computer Science pages 1099-1103, SPRINGER-VERLAG, Germany. 2006.

[11] Dr. Thomas Link, General Psychology (How can we tell emotions apart?), Psychology department at Pierce College, http://www.pierce.ctc.edu/tlink/general/projects/emotions/emotions.html, Jan 17th , 2010.

[12] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. Discourse Processes. 1998.

[13] M.W. Berry, S.T. Dumais, G.W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review 37 (4). 1995.

[14] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara. ORCHID: Thai Part-Of-Speech Tagged Corpus. Technical Report Orchid TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand. Dec 1997.

[15] Weka Version 3.6.0 (c) 1999-2008, The University of Waikato, Hamilton, New Zealand, http://www.cs.waikato.ac.nz/~ml/weka/

[16] B. Yu, Z.B. Xu, and C.H Li. Latent Semantic Analysis for Text Categorization using Neural Network.

[17] T.M. Mitchell, Machine Learning. McGraw Hill International Editions. 1997. pp. 182-186.

ภาคผนวก ข.
รหัสหน้าที่ของคำที่ใช้ในโปรแกรมตัดคำภาษาไทย SWATH

รหัสหน้าที่ของคำที่ใช้ในโปรแกรม SWATH [33] แสดงได้ดังตารางที่ 21

ตารางที่ 21 ตารางความหมายของคำย่อในโปรแกรมการตัดคำภาษาไทย SWATH

No.	POS	Description	Example
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator “ไม่”	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator “ไม่”	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator “ไม่”	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ

25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย,โอ้, เออ, เอ๋, อ๋อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), “, ,, ;

ภาคผนวก ค.
ข้อความภาษาไทยสำหรับการสร้างตัวแบบ

1. ซิ่งชนดับพระบินฑบาตร หมาวัดแสนรู้โดนด้วย ตะเกียกตะกายคลานเข้าหาศพพระ
2. เด็กผู้หญิงจับเอาแมวดำตัวหนึ่งโยนลงไปในน้ำแล้วปล่อยให้มันตะเกียกตะกายปีนตลิ่งหนีไป
3. ปีน ป่าย ตะเกียกตะกาย สู่ลานสนบนภูสอยดาว ผจญภัยไปกันกับทริปโหด แต่ขึ้นไปแล้วความเหน็ดเหนื่อยจะหายเป็นปลิดทิ้ง เมื่อเห็นทิวสน ทุ่งดอกไม้และสายหมอก
4. เด็กบ้านนอกจะเรียนกันสบายๆ มากกว่าเด็กในกรุงเทพฯ เพราะไม่ต้องตะเกียกตะกาย แย่งกันเข้าาโรงเรียนดีๆ
5. วงการสงฆ์ ตะลึง จับเจ้าอาวาสตุ๊ด เจ้าของฉายา "เจ๊ดาว"
6. พระสงฆ์ระดับเจ้าอาวาสวัดชื่อดังใน จ.ลำพูน มีพฤติกรรมไม่เหมาะสม มีการผ่าตัดแปลงเพศเสริมหน้าอก และแต่งกายเป็นหญิง ใช้ชื่อว่า "เจ๊ดาว" ออกตระเวนเที่ยวตามสถานบันเทิงยามค่ำคืน ซื้อบริการทางเพศนักศึกษาชาย จนสุดท้ายต้องถอดจีวรลาสิกขาบท
7. ผงะห้อยศีรษะฝรั่ง! ฆ่าตัดหัว สยอง-พระราม 8
8. คดีที่ชวนขนลุก และกลายเป็นประเด็นที่น่าสนใจ คงหนีไม่พ้นฆ่าตัดหัวฝรั่ง แขวนสะพานพระราม 8 วิธีฆาตกรรมแบบพิศดาร โหดร้าย แต่หากเป็นฆาตกรรมจริง คนลงมือก่อเหตุตัดหัวนำมาแขวนประจานเช่นนี้ถือว่าไม่ธรรมดา
9. คนทั่วโลกช็อค เมื่อ เดวิด คาร์ราดีน อดีตพระเอกฮอลลีวู้ดชื่อดัง-ตัวร้ายจากหนัง "คิล บิล" ผูกคอตายพิสดารในห้องสูทโรงแรมหรูกลางกรุง เข้าข่ายลักษณะ"ออโต้อีโรติก"สำเร็จความใคร่จนไม่รู้สึกตัวถูกเชือกรัดคอ หมกตัวเองในตู้เสื้อผ้าโรงแรม
10. โลกสูญเสียคนดังระดับโลก "ไมเคิล แจ็คสัน" ราชาเพลงป๊อป อำลาโลก ด้วยวัย 50 ปี หลังพบหมดสติในบ้าน คาดเกิดอาการหัวใจวาย ปลุกระแสความคลั่งไคล้ในราชาเพลงป๊อปผู้นี้ แม้จะลาโลก แต่สุดท้ายยังคงตราตรึงในหัวใจทุกคน
11. เปิดตัวนางเอกนัยน์ตาน้ำผึ้ง เพชรา เชาวราษฎร์ อดีตนางเอกตลอดกาล ในโฆษณามิสทีน ในรอบ 33 ปี ความงดงามด้วยท่วงท่า และความคิดถึงที่เหล่าประชาชนแฟนคลับ รอคอยการกลับมาของเธออีกครั้ง
12. หวานชื่น กบ-บรู๊ค วิวาห์รับต้นปีหวานชื่น รักล้นใจของคู่รักตลอดกาลที่กลายเป็นกระแสว่าที่ลูกสะใภ้ กับแม่สามีไม่ลงรอย แต่สุดท้ายต่างชื่นมื่น

13. ไฟไหม้ซานติก้าผับ ฉลองต้อนรับปีใหม่ กับเหตุการณ์สุดสลด เมื่อผับชื่อดังย่านเอกมัย เกิดเพลิงไหม้จนกลายเป็นโศกนาฏกรรมคร่าชีวิตเหล่านักท่องราตรีเกินกว่าครึ่งร้อย จนกลายเป็นข่าวดังใหญ่โตที่อยู่ในกระแส จวบจนการหามาตราการเข้มงวดในการเปิดผับ สถานบันเทิง

14. ค่านิยมในการกำหนดชีวิตลูกนี้กลายเป็นแฟชั่นของสังคมกรุงเทพฯ หัวเมืองใหญ่ไปแล้ว เพราะเมืองไทยไปรับวัฒนธรรมนี้มาจากสังคมตะวันตกซึ่งเป็นสังคมทุนนิยม แข่งขันกันที่ความรู้ความสามารถเพื่อให้ได้งานดี เงินเดือนสูง

15. คุณพ่อคุณแม่จำนวนมากพยายามขวนขวายทำทุกวิธีเพื่อให้ลูกเข้าโรงเรียนดัง ลึกลงไปยังเป็นการยกฐานะทางสังคมให้แก่ครอบครัว

16. การกำหนดชะตาชีวิตของเด็กไปหมดทุกเรื่อง บางครั้งไม่ใช่ผลดีเสมอไป หากสิ่งที่พ่อแม่หยิบยื่นให้ไม่ใช่สิ่งที่ลูกต้องการ อาจไปสร้างแรงกดดัน ความเครียด ดังนั้น ทางที่ดีควรปล่อยให้ลูกเลือกทางเดินของตัวเองบ้าง

17. กระแสความนิยมจัดตารางเรียนให้ลูกตั้งแต่เช้าจรดค่ำของบรรดาคุณพ่อคุณแม่ กำลังขยายวงกว้างขึ้นทุกที และจำนวนมากไม่รู้ตัวว่าที่ให้ลูกตัวเองคร่ำเคร่งอยู่กับตำราเช้ายันดึกได้กระชากรอยยิ้ม ความร่าเริง ความสดใส แววตาแจ่มใส ออกไปจากชีวิตพวกเขาตลอดชีวิต และมองว่า

18. โรฮิงญา จากผู้อพยพพม่า สู่วิกฤตทางการเมือง

19. เมื่อชาวโรฮิงญา แรงงานผิดกฎหมายที่ลักลอบเข้าเมือง หลังพบถูกปล่อยทิ้งในเรือกลางทะเล และถูกทหารพม่าใช้แส้ทุบตีแถมถูกเฆี่ยนหลังเหวอะแผลเน่า จนกลายเป็นปัญหาทางการเมือง เมื่อพม่าไม่ยอมรับ ปัดไม่ใช่คนในประเทศ

20. ไข้หวัดใหญ่สายพันธุ์ใหม่ 2009 คร่าชีวิตคนทั้วโลกไม่น้อยกว่า 12,954 ราย ประเด็นร้อนที่ไม่สามารถละสายตาจากความสนใจได้

21. หลิงปิง ชื่อนี้คงไม่มีใครไม่รู้จัก เพราะตามหน้าหนังสือพิมพ์ หรือแม้กระทั้งวิทยุ โทรทัศน์ สื่อต่าง ๆ ให้ความสนใจในการตั้งครรภ์ของแม่แพนด้า และก่อเกิดลูกแพนด้าน้อย อย่าง หลินปิง จนกลายเป็นขวัญใจของคนไทยทั้งประเทศ

22. เคอิโงะ วิ่งโร่ประกาศตามหาพ่อบังเกิดเกล้าชาวญี่ปุ่น และกว่าทั้งคู่จะได้พบเจอ และสวมกอดด้วยความรัก แทบเป็นระยะเวลาที่ยาวนานยิ่งนัก

23. น้องหม่อง เด็กไร้สัญชาติเข้าแข่งขันพับเครื่องบินกระดาษคว้ารางวัลสร้างชื่อเสียง

24. ความบาดหมางร้าวลึกหลังจากอดีตราชินีลูกทุ่ง พุ่มพวง ดวงจันทร์ ลาโลกไปแล้ว ลูกเพชรลูกชายหัวแก้วหัวแหวน และไกรสร พ่อบังเกิดเกล้า ขัดแย้งจนถึงขั้นหนีออกจากบ้าน สุดท้ายเกิดการฟ้องร้อง เหตุภายในที่สะสางไม่จบไม่สิ้น

25. นาธาน โอมาน อดีตนักร้องชื่อดังจอมลวงโลกประกาศเล่นหนังฮอลลีวู้ดเว่อร์จนนักสืบพันทิพย์ตามสืบ สุดท้ายลวงโลกโกงเงินบังสัญชาติคนทั้งประเทศ

26. เจ๊เบียบ จูงมือ เรือโทหญิงแฉ พล.อ.ชื่อดัง เตรียมทหารรุ่น 10 สังกัดกองทัพไทย ข่มขืนมาราธอน 4 ปี จนวงการสีกากีสั่นคลอน แต่เรื่องยังคงเงียบ

27. เอ อนันต์ ประสบอุบัติเหตุ ขณะขับรถปิกอัพ วีโก้ ไปทำบุญที่จังหวัดเพชรบูรณ์ จนต้องหามส่งโรงพยาบาลอย่างเร่งด่วน เคราะห์ดีหายปกติ

28. นางเอกธิดาวานร หวิดสิ้นชื่อ ขับรถกลับจากกองถ่ายละครที่โรงถ่ายละครดาราวีดีโอเพื่อจะกลับบ้านพักแต่ประสบอุบัติเหตุรถของนางเอกสาวเกิดเสียหลักพุ่งลงข้างถนนและชนกับเสาไฟฟ้าข้างทางจนรถยนต์เก๋งพังยับเยินเป็นเหตุให้นางเอกสาวและเพื่อนชายที่มาด้วยกันได้รับบาดเจ็บ

29. เธอต้องอดตาหลับขับตานอน เพื่อพยาบาลเขาทั้งคืน

30. เมื่อถึงคราวต้องสอบไล่ นักศึกษาต่างก็อดตาหลับขับตานอน เพราะแต่ละคนอ่านหนังสือช่วงใกล้สอบ

31. ปีใหม่กลับบ้านปลอดภัย มอบของขวัญจริงใจไร้แอลกอฮอล์

32. ชาวจีนซึ่งเป็นชนชาติใหญ่ที่สุดของโลก ได้ออกเสียงทางอินเตอร์เน็ตให้สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี “เป็นมิตรที่ดีที่สุดในโลกอันดับ 2 ของชาวจีน”

33. ในหัวข้อ “ประเทศที่เป็นมิตรกับชาวต่างชาติมากที่สุด” ไทยติดลำดับต้น ๆ ที่นักท่องเที่ยวมาเมืองไทยต้องชมเรื่องนี้กันแทบทุกราย โดยเฉพาะ “ยิ้มสยาม” และน้ำใจไมตราจากคนไทย แม้คนไทยส่วนใหญ่พูดภาษาอังกฤษไม่ได้ แต่ก็สื่อกันได้ด้วยรอยยิ้ม

34. ไทยถูกจัดให้ติดลำดับต้น ๆ ของประเทศที่มี “การแพทย์โดดเด่น” ถือกันว่า ไทยเป็น “ฮับการแพทย์” ในภูมิภาคนี้ก็ได้ ซ้ำยังราคาไม่แพงนักเมื่อเทียบกับมาตรฐานสากล

35. เธอต้องอดมื้อกินมื้อ และพบแต่ความอดอยากหิวโหยนับตั้งแต่สามีเธอจากไป

36. พม่าเปิดใช้โครงข่ายโทรศัพท์มือถือ 3G ทำให้เกิดคำถามขึ้นทันทีว่า 3G ของเมืองไทยเมื่อไรจะมา

37. ถึงแม้ยอดขายของ Atom จะร้อนแรง แต่ผู้บริหารอินเทลกลับกล่าวว่าเน็ตบุ๊กอาจไม่มีอนาคต

38. แอลจีประกาศผลิตจอแอลซีดีแบบ 3 มิติในระดับ Full HD เป็นตัวแรกของโลก ซึ่งมีขนาดใหญ่ถึง 23นิ้วและมีรูปร่างที่ดูดีแถมยังบางเฉียบ

39. แอลจีลั่นจอแอลซีดีขนาด 42 นิ้ว ที่บางที่สุดในโลก ในการแข่งขันในเรื่อง ”ความบางที่สุดในโลก” ทำให้จอแอลซีดีเกิดความร้อนระอุขึ้นอีก

40. เธอนอนหมดสติไปเพราะสูญเสียเลือดมาก

41. ลดกระหน่ำ กับสินค้าไอทีพร้อมราคาที่ดึงดูดใจ ในงานคอมมาร์ท

42. เมื่อลูกสาวที่รอคอยการกลับมาของพ่อได้พบหน้าพ่ออันเป็นที่รักและคิดถึงมานานก็กลั้นน้ำตาไว้ไม่อยู่

43. เมื่อศาลเกาหลีใต้ต้องตัดสินคดีผู้ชายข่มขืนกะเทยที่ผ่านการผ่าตัดแปลงเพศมาแล้วมากกว่า 70000 คนในประเทศ

44. เมื่อหญิงสาวคิดถึงความสูญเสียที่เรียกคืนไม่ได้ก็กรีดร้องสุดเสียงก่อนหมดสติไป

45. เรื่องขนลุกที่นักศึกษาหญิงถูกฆ่าข่มขืนแขวนคอที่ตึกคณะนิเทศศาสตร์เมื่อ 10 ปีก่อนถูกนำมาเล่าอีกครั้งเมื่อยามกะดึกคนหนึ่งเห็นผีคอขาด

46. เหน็ดเหนื่อยหนักหนากับการไขว่คว้าอย่างเอาเป็นเอาตาย

47. โจรใต้กระหน่ำยิงและเผารถบรรทุกรถจักรยานยนต์ เสียชีวิตทันที 2 ศพ

48. กระหน่ำยิง6นัดคนขับรถทัวร์ดับคาสายใต้ใหม่

49. กระหน่ำยิงกลางกรุง นักศึกษาเทคนิคกรุงเทพดับ 2

50. ขนลุก เฆ่าออสเตรียจับลูกสาวขังข่มขืนมาราธอน 24 ปีจนมีลูก 7 คน

51. ฆาตกรโหดฆ่าแขวนคอ5ศพได้รับการลดโทษ

52. ฉันไม่เคยเสียใจที่มีชีวิตรันทดแต่งดงาม

53. ตะลึง..อดีตนางงามแปลงเพศเปลี่ยนเป็นชาย

54. ตะลึงสแปมเมล์เป็นสาเหตุของภาวะโลกร้อน เมื่อพลังไฟฟ้าที่สูญเสียไปกับการส่งต่อปีคือ 33 พันล้านกิโลวัตต์ชั่วโมงหรือเทียบเท่ากับการใช้ไฟฟ้าตามบ้านกว่า 2.4 ล้านหลัง

55. ทุกคนต้องพยามยามเบียดตัวหลบฝนที่กระหน่ำลงมา และตะเกียกตะกายหนีเอาชีวิตรอดจากอาคารหลังนั้นแทบไม่ทัน

56. ผงะ สาวมะกันก่อคดีสยองโลก ฆ่าลูกน้อยทารกวัย 3 สัปดาห์ จัดการตัดหัว แล้วเปิดกะโหลกควักกินสมองสดๆ รวมทั้งตัดนิ้วเท้า 3 นิ้วมากิน

57. พ่อข่มขืนลูกจนต้องหนีออกจากบ้านมาเร่ร่อนแล้วก็โดนข่มขืนอีก

58. มีข่าวดีให้หายคิดถึงกับการกลับมาเมืองไทยอีกครั้งของหนุ่มๆกลุ่มซูเปอร์จูเนียร์

59. สภาพที่อู่ฮุ่ต้องตะเกียกตะกายไปตามเนินทรายหรือลื่นไถลไปตามสันที่ลาดเอียง

60. สาวผูกคอประชดแฟน ถ่ายคลิปมือถือลาตาย

61. ตะลึง พบตัวตืดในข้าวหลาม

62. ความรักราบรื่นไร้ปัญหา

63. จัดวิวาห์ ผาสุก ทุกโมงยาม รักชื่นบาน ดังบุพเพ เสกสรรมา

64. ครูชายแดนใต้เปิดใจ ความรุนแรง ไม่อาจบั่นทอนจิตวิญญาณเพื่อชาติ

65. โครตเสียความรู้สึกเลย

66. อย่าร้องไห้เสียใจเพราะเขาไม่รักเรา แต่จงดีใจว่า เราเคยรักใครสักคนมากมายขนาดนี้

67. เมื่อเธอนั้นทิ้งฉันไปมีคนใหม่ ฉันร้องไห้เสียใจไปเท่านั้น เมื่อคนที่ฉันรักมาทิ้งกัน ในเมื่อมันไม่เห็นค่าของฉันเลย อย่าร้องไห้ให้กับคนที่ลวงโลก คนที่พูดโกหกหลอกลวงฉัน ร้องไห้ไปเขาก็ไม่สนใจกัน จงปล่อยมันออกไปจากใจเรา

68. เสี้ยวนาทีที่พาเราเจ็บ เหมือนโดนเย็บเจ็บรอยแผล เธอเหยียบย่ำซ้ำเติมไม่เหลียวแล ต้องยอมแพ้จมระบมในโคลนตม

69. รักคือการให้อย่างผ่อนคลาย แม้อกหักไม่ตายยังหายใจอยู่

70. เธอไปอยู่ที่ไหน ห่างหายกันไปนาน ไม่เป็นเหมือนวันวาน งานยุ่งหรืออย่างไร กลัวเธอจะลืมกัน ว่ามีฉันรออยู่ตรงนี้ อย่าลืมฉันนะคนดี กลัวเธอจะไปมีใคร

71. ถ้าลูกเจ็บ ลูกเสียใจแค่ไหน พ่อแม่จะรู้สึกมากกว่านั้นหลายร้อยเท่า

72. อยากได้ยินว่ารักกัน

73. น้ำตามันเอ่อออกมาตลอดเวลา แค่นึกขึ้นมาว่าเค้าได้พูดอะไรกับเราก่อนทุกอย่างจะจบลง มันก็ไหลออกมาเป็นทางหยุดไม่ได้

74. คนเราบางคนเลือกยอมทน เจ็บปวดและเสียน้ำตามากมาย เพียงเพื่อแลกกับความสุขจากการได้มีความรักนิดเดียว

75. เมื่อถึงวันที่เราต่างต้องก้มหน้ายอมรับกับตัวเองว่าช่วงเวลาแห่งความผูกพันที่เลยผ่าน มันคงไม่ได้หอมหวานเหมือนดังเก่า และ ความรักในโลกแห่งความเป็นจริงก็คงไม่สามารถสวยงามได้ตลอดเวลา

76. พ่อยิ้มแล้วดึงฉันเข้าไปนั่งใกล้ๆ ฉันกลั้นน้ำตาไม่อยู่แล้ว พรั่งพรูความเจ็บช้ำ เสียใจทุกอย่าง พ่อโอบฉันไว้นิ่งๆ จนฉันหยุด

77. ฉันอกหัก เพิ่งแยกทางกับผู้ชายที่ฉันรักมากที่สุด ฉันจะไม่มีทางลืมความเจ็บปวด ความเสียใจคราวนี้ง่ายๆ

78. มีชีวิตเพื่ออะไร มันหมดความหมายเมื่อไร้เธอ ต่อไปคงได้แค่เพ้อ ถึงวันเวลาที่เลยผ่าน รู้ดีว่ามันจะเกิด แต่ก็เตรียมใจไม่ทัน ที่สุดความรักในความต่างกันก็เลิกรา

79. เมื่อชีวิตไม่ใช่ฝัน เธอเหนื่อยกับฉันมามากพอ ไม่ผิดที่เธอจะท้อ ขอไปเจอคนที่ดีกว่า แต่มีบางคนที่เจ็บ จนไม่อาจหยุดน้ำตา ไม่อาจทำใจกับการจากลาในวันนี้

80. ต่อให้มันรู้ คนอย่างฉันไม่ควรคู่เธอ ต่อให้รู้สักวันต้องเจอ ก็ยังปวดร้าว เมื่อคำว่าเรากลายเป็นความหลัง

81. มันไม่เหลือเรี่ยวแรงก้าวเดิน เมื่อมันไม่มีเธอร่วมทาง ที่ได้พบได้เจอวันนี้ มันเกินกว่าสิ่งที่ฉันกลัว

82. เคยเธอเคยบ้างไหม ต้องเป็นทุกข์ทุรนทุรายเพราะใครสักคน ร้อนรนอย่างนี้

83. เคยเธอเคยหรือยัง ต้องผิดหวังเสียใจให้ใครที่ใจไม่จริง กับสิ่งลวงลวงที่เจอ

84. ต้องเจ็บต้องช้ำที่เรานั้นมันโง่ไป อยากหลีกไปไกล ยังหลีกไม่พ้น

85. นั่นแหล่ะคือความกดดันที่ฉันต้องทนต้องเจอ ทุ่มเทใจให้เธอ มากมายจนเกินตัดใจ

86. เพียงเพราะแค่อยากได้ยินว่ารัก ต่อให้เจ็บปวดมากมายนักก็คงพอรับไหว ทุ่มสุดตัวจนรู้สึกว่ามันเหนื่อยเกินไป ยอมจนคิดว่าหัวใจเจ็บแล้วจนชาชิน

87. ทรมานเมื่อสิ่งที่ทำไปดูไร้ประโยชน์ ไม่อยากให้เธอโกรธจนคิดว่าฉันนั้นบ้าบิ่น แค่คำว่ารักเท่านั้นที่ฝันอยากจะได้ยิน อีกกี่หยาดน้ำตาโรยริน หัวใจเธอที่แข็งเหมือนหิน จะพูดมันสักที

88. รักแท้อย่าทิ้ง รักจริงอย่าจาก รักมากอย่าพราก รักแต่ปาก อย่ารักเลย

89. ขอสาปแช่งคนที่มันด่าว่าฉัน ขอให้มันพบแต่ความฉิบหาย ขอสาปแช่งให้มันนั้นวอดวาย ขอให้ตายตกนรกอเวจี

90. ขอสาปแช่งให้มันนั้นพินาจ ขออาฆาตให้มันตายกลายเป็นผี ขอสาปแช่งให้มันไม่ได้ดี ขอให้มีแต่ความเศร้าทุกข์ระทม

91. ขอให้ม้วยดับสูญและขื่นขม ขอสาปแช่งให้มันนั้นล่มจม ขอให้ตรอมตรมกับสิ่งที่มันทำ

92. ขอสาปแช่งให้มันนั้นดับดิ้น ขอให้สูญสิ้นมอดม้วยและตกต่ำ ขอสาปแช่งให้ลูกหลานไม่จดจำ ขอให้กรรมที่มันก่อสนองคืน

93. ขอให้สำลักความสุข ขอให้ทุกข์กระเด็น ขอให้เห็นรอยยิ้ม ขอให้อิ่มความรัก ขอให้หนักเงินทอง

94. อันว่าสันดานผู้ชายนับว่าแย่ ข่มเพศแม่ของมึงเหมือนหมูหมา ทั้งตบทั้งตียิ่งกว่าควายไถนา ทั้งด่าทั้งว่าไม่รู้จะสาธยายพฤติกรรมมันยังไง

95. เมื่อคืนฝันสยองตีสองกว่า ฝันไปว่านอนซบกับศพผี นอนขึ้นอืดค้างตายมาหลายปี เหมือนปิศาจอเวจีที่น่ากลัว

96. เหม็นซากศพ อบอวลไปทั่วห้อง ไม่กล้ามองต้องคู้คดนอนหดหัว ตกใจตื่นขนลุกซู่ดูรอบตัว เห็นชัวร์ชัวร์เมียนอนตดสลดใจ

97. อันนารีมีมากเหมือนฝูงลิง จะจีบทิ้งจีบขว้างก็ยังไหว มันไม่รักช่างมันไม่เป็นไร อย่าเสียใจเพื่อนเอ๋ยกะเทยมี

98. หนึ่งสมองสองมือต้องไขว่คว้า เมื่อมีปัญหาขึ้นมาต้องแก้ไข เหนื่อยและท้อขอพักหน่อยไม่เป็นไร ให้มีแรงลุกขึ้นใหม่ได้อีกครา

99. คนเราก็เป็นแบบนี้กันทั้งนั้น มีสุขสันต์โศกเศร้าเท่ากันแหละหนา ความสำเร็จแต่ละครั้งกว่าจะได้มา ต้องแลกด้วยเหงื่อและน้ำตาแทบทุกคน

100. ขอเป็นอีกหนึ่งแรงใจให้คนสู้ ให้เธอรู้ว่าเธอยังมีหวัง ขอส่งแรงใจให้เธอมีแรงพลัง ไปยังฝั่งฝันที่เธอวาดหวังและตั้งใจ

101. นั่งมองดาวบนฟ้า ส่งความห่วงหาไปให้ ข้ามน้ำทะเลเขาที่ยาวไกล ให้ใจอีกใจที่ไกลกัน

102. คิดถึงใครคนหนึ่งตรึงตราจิต พรหมลิขิตขีดทางให้ห่างเหิน ต่างคนต่างหนทางที่ต้องเผชิญ ย่างก้าวเดินบนเส้นทางที่ต่างกัน

103. ไม่ว่าจะอยู่ที่ไหน เธอก็อยู่ใกล้ใกล้ฉัน อยู่ในความคิด คิดถึงทุกวัน ไม่ว่าฝันหรือความจริง

104. วันเกิดปีนี้ฉันไม่มีอะไรจะให้ มีก็แต่ความจริงใจที่ยังมีให้สม่ำเสมอ พร้อมทั้งความรักความห่วงใยเมื่อได้เจอ ทั้งหมดนี้มีให้เธอสม่ำเสมอและตลอดไป

105. วันนี้วันเกิดคนที่รัก แต่ไม่อาจอวยพรได้ ไม่แม้แต่จะอยู่ใกล้ ด้วยเหตุผลมากมายนับพัน

106. รักคุณแล้วรู้ไว้โปรดได้ทราบ จะให้กราบงามงามยอมตามสั่ง สมัครรักหน้าจอจะรอฟัง มีหรือยังคนในหัวใจคุณ

107. เธอคงไม่มีวันกลับมา จึงปล่อยฉันให้อ่อนล้าอยู่ตรงนี้ เจ็บปวดกับน้ำตาที่มี ให้ฉันพร่ำเพ้อทุกนาทีโดยไม่เหลือใคร

108. จุดประทัดสวัสดีรับปีใหม่ พร้อมพลุไฟใส่ฟ้าเฮฮาแสน เฉลิมฉลองก้องดังทั้งต่างแดน เหมือนเป็นแผ่นดินเดียวรักเกี่ยวกัน

109. ดุจโลกยิ้มอิ่มเอมเกษมสุข ประหนึ่งทุกข์ซุกซ่อนได้ผ่อนผัน เสียงชโยโห่ตรึมกระหึ่มมัน แต้มสีสันวันใหม่ด้วยใจคอย

110. พอหน้าท้อง ของแม่ เริ่มป่องออก พ่อก็บอก อยากให้ลูก เป็นดอกเตอร์ ย่าก็อยาก ให้หลานเป็น นายอำเภอ แต่ต้องเก้อ หมอบอกแม่ แค่ลงพุง

111. มีแฟนคนหนึ่งที่น่ารัก ฉันก็มักโทรหาเขาเสมอ แต่ว่าเขาก็ไม่โทรมาเลยเธอ นอนละเมอถึงเขาเศร้าอุรา

112. อันตัวเราแสนเหงา เศร้าใจนัก ไม่สมรักกับแก้วตา ยอดยาหยี เหมือนสุนัขเห่าเครื่องบิน สิ้นฤดี โอ้ยาหยีทำให้เรา ต้องเศร้าใจ

113. อย่าบังคับให้ฉันรู้สึกดี ยิ่งเธอทำแบบนี้รู้มั้ยยิ่งเจ็บเหลือเกิน ความรู้สึกมันเปลี่ยนกันไม่ได้ ไม่ต้องมาเจอหน้ากันอีกนะ ก็แค่เธอไม่รักมันเจ็บอยู่แล้วเข้าใจมั้ย ก็ฉันเป็นเพื่อนกับเธอไม่ได้จริงๆ

114. เพราะว่าใจกลัว กลัวว่าเธอจะทิ้งกันจากไป ลืมคนที่เคยบอกรักกัน

115. เขาอดตาหลับขับตานอนข้ามวันข้ามคืน เร่งอัดอ่านหนังสือเตรียมสอบ

116. ยามไข้ไร้สุขแม่ก็ทุกข์อาทร อดตาหลับขับตานอน พยาบาลรักษา ครั้นเมื่อมดหมอร้องขอค่ายา ก็ดิ้นรนเสาะหา จนหน้าแดงหน้าดำ ทรัพย์สินเงินสดก็หมดก็สิ้น บางครั้งจะกินก็ไม่มีอีกซ้ำ ต้องเอาไร่นา ไปเที่ยวเร่จำนำ หวังให้ลูกงามของแม่รอดตาย

117. ภาพพระอาจารย์ที่อดตาหลับขับตานอนเฝ้าดูแลอยู่เคียงข้างกายเขาในวาระสุดท้ายคือความอบอุ่นใจ คือที่พึ่งสุดท้ายที่พวกเขาได้เห็นก่อนลมหายใจรวยรินจะขาดห้วง

118. ได้เวลาคร่ำเคร่งกับกองชีทตำรากองโต ได้เวลาของการอดตาหลับขับตานอน ได้เวลาของมาม่าอาหารด่วนจี๋มื้อดึกกะกาแฟแก้วโต ได้เวลาที่ต้องการกำลังใจมากมาย

119. น่าสงสารดารานักแสดง ที่ในบั้นปลายของชีวิตมักมีเรื่องรันทด จนทำให้ต้องสลดใจหลายต่อหลายคนอยู่เสมอๆ

120. นาธาน โอมาน ฉายา "ลับ ลวง แหล" ก่อวีรกรรมมาหลายเรื่อง หลายประเด็นที่ทำให้สังคมช่วยกันขุดคุ้ย

121. เซ็นทรัลเวิลด์ ผุดลานสเกตน้ำแข็ง "ไอซ์ เวิลด์ แอท เซ็นทรัลเวิลด์" ลานสเกตน้ำแข็งกลางแจ้งแห่งแรกในไทย หวังเป็นแม่เหล็กดึงลูกค้า คาดคืนทุนภายใน2ปี

122. ทองแพงบ้าเลือด ทองคำตลาดนิวยอร์กปิดพุ่งขึ้นแตะระดับสูงสุดเป็นประวัติการณ์ และทำสถิติปิดบวกติดต่อกันยาวนานที่สุดในรอบ 27 ปี ฉุดราคาทองไทยพุ่งกระฉูดบาทละ 19,200..

123. ศาลให้ประกันตัว"นาธาน โอมาน" วงเงิน 1 แสนบาท แม่เลี้ยงใช้โฉนดที่ดิน 2 ไร่ วางประกันเจ้าตัวลิงโลดยิ้มแฉ่ง พร้อมกลับลำประกาศสู้คดีจนถึงที่สุด

124. เจ้าชายวิลเลียม แห่งราชวงศ์อังกฤษ ประทับค้างคืนริมถนน กรุงลอนดอน ในสภาพอากาศหนาว ติดลบ 4 องศาเซลเซียส เพื่อให้เข้าใจหัวอกคนไร้บ้าน ซึ่งเป็นหนึ่งในงานการกุศล ขององค์กรเซ็นเตอร์พอยท์ ที่ช่วยเหลือคนไร้บ้าน

125. พบสารเคมีในอากาศและน้ำที่แหลมฉบัง จำนวนผู้ป่วยเพิ่มขึ้น

126. ผู้ชายมีปัญหาเรื่องการ นอน กรน เสียงดังมาก จนฝ่ายภรรยาถึงกับทนไม่ได้ เพราะเสียง กรน ปลุกเธอตื่นแทบทุกคืน

127. คน กรน เอง ไม่รู้ตัว และไม่สามารถบังคับตัวเองให้ กรน เสียงเบาลงได้ ส่งผลให้ทั้งสอง ตกลงที่จะแยกห้องกัน เพื่อการนอนที่เป็นสุขของฝ่ายภรรยา

128. ผลสำรวจสุขภาพวัยรุ่นเมืองกรุง นอนดึก ชอบกินอาหารฟาสท์ฟู้ด ใช้ช้อนและหลอดดูดน้ำร่วมกัน ดื่มเครื่องดื่มแอลกอฮอล์ เครียด ขณะที่สิ่งกังวลมากที่สุดสำหรับวัยรุ่นคือกลัวเป็นโรคมะเร็ง

129. ยามไข้ไร้สุขแม่ก็ทุกข์อาทร อดตาหลับขับตานอน พยาบาลรักษา

130. ครั้นเมื่อหมอร้องขอค่ายา ก็ดิ้นรนเสาะหา จนหน้าแดงหน้าดำ

131. ทรัพย์สินเงินสดก็หมดก็สิ้น บางครั้งจะกินก็ไม่มีอีกซ้ำ ต้องเอาไร่นา ไปเที่ยวเร่จำนำ หวังให้ลูกงามของแม่รอดตาย

132. ได้เวลาคร่ำเคร่งกับกองชีทตำรากองโต ได้เวลาของการอดตาหลับขับตานอน

133. สายเกินไป ปวดร้าวคิดอยากย้อนเรื่องราวแค่ไหน ได้แต่ฝัน

134. ได้เวลาของมาม่าอาหารด่วนจี๋มื้อดึกกะกาแฟแก้วโต

135. ได้เวลาที่ต้องการกำลังใจมากมาย

136. ผวา หวัดใหญ่ 2009 ผู้นำสหรัฐ ประกาศภาวะฉุกเฉินแห่งชาติรับมือการระบาดของโรคไข้หวัดใหญ่ 2009

137. ราชองครักษ์กำลังหัวเราะอย่างเอาเป็นเอาตายอยู่ไม่ไกลนัก

138. ร่างสูงที่นั่งคร่ำเคร่งอยู่กับโต๊ะเอกสารตัวใหญ่ในห้องเงยหน้าขึ้น เมื่อเห็นเงาคนก้าวผ่านเข้ามา

139. เอาผู้ชายของฉันคืนมา หน้าไม่อายจริงๆ อายุก็ป่านนี้แล้ว ยังมาหลอกเด็ก

140. พม่าเปิดใช้โครงข่ายโทรศัพท์มือถือ 3G ทำให้เกิดคำถามขึ้นทันทีว่า 3G ของเมืองไทยเมื่อไรจะมา

141. ทรุดนั่งลงบนพื้นศิลาของห้อง ก้มหน้าด้วยความอ่อนล้า สิ้นหวัง เมื่อไม่พบส่วนสำคัญของชีวิต

142. เขาเก็บตัวฝึกฝนร้องเพลงอยู่อย่างขมักเขม้น

143. ทำไมเหรอ พระเอกคนไหนเป็นเกย์ หรือว่าดาราคนไหนท้องก่อนแต่ง ถึงได้ดูตื่นเต้นเหลือเกิน

144. ฉันทั้งเกลียดทั้งขยะแขยงคนนิสัยอย่างนี้

145. การที่เธอได้ครองหัวใจของผู้ชายหลายคนแบบข้ามคืน เพราะผู้ชายมักเบื่อกับการคร่ำเคร่งพิธีรีตอง เอาแต่ใจตัวจัดของผู้หญิงทั่วไป

146. เหตุรถไฟวิ่งมาถึงบริเวณสถานีรถไฟ เกิดฝนตกลงมาอย่างหนัก ทำให้รถไฟตกราง

147. รถไฟตกราง สร้างความเสียหาย การดำเนินการล่าช้า

148. รัฐบาลไม่เอาใจใส่ประชาชน

149. คิดถึงและอยากเห็นหน้าหลานเร็วเร็วจังเลย

150. ฉันมันคนโง่เหนือใครใคร มีรักแท้อยู่ดูแลไม่ได้

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวปิยธิดา อินทร์รักษ์ เกิดวันที่ 31 สิงหาคม พ.ศ. 2525 ที่จังหวัดนครศรีธรรมราช สำเร็จการศึกษาระดับมัธยมศึกษาปีที่ 5 จากโรงเรียนสาธิต มหาวิทยาลัยสงขลานครินทร์ อำเภอเมือง จังหวัดปัตตานี และได้รับทุน AFS ไปศึกษาต่อในระดับชั้นมัธยมศึกษาปีที่ 6 ณ ประเทศออสเตรเลีย และสำเร็จการศึกษา Senior Certificate จาก Mackay North State High School จากนั้นได้เข้าศึกษาที่คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ และสำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต เกียรตินิยมอันดับสอง สาขาวิศวกรรมศาสตร์คอมพิวเตอร์ ในปี พ.ศ. 2548 ต่อมาได้เข้าทำงานที่บริษัทซีเกท เทคโนโลยี (ประเทศไทย) จำกัด และได้รับทุนในการศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิทยาศาสตร์คอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย