

การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์

กานดา แผ้ววัฒนากุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (บริหารเทคโนโลยีสารสนเทศ)

คณะสถิติประยุกต์


สถาบันบัณฑิตพัฒนบริหารศาสตร์

2555

การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์

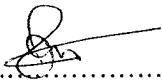
กานดา แผ้ววัฒนากุล

คณะสถิติประยุกต์

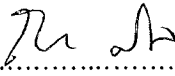
อาจารย์..........อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ดร.ปราโมทย์ ลือนาม)

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาแล้วเห็นสมควรอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต (บริหารเทคโนโลยีสารสนเทศ)

รองศาสตราจารย์..........ประธานกรรมการ


(ดร.กฤษณะ ไวยมัย)

ผู้ช่วยศาสตราจารย์..........กรรมการ

(ดร. โอม ศรีนิต)

ผู้ช่วยศาสตราจารย์..........กรรมการ

(ดร. ปรีชา วิจิตรธรรมรส)

อาจารย์..........กรรมการ

(ดร.ปราโมทย์ ลือนาม)

รองศาสตราจารย์..........รักษาราชการแทนคณบดีคณะสถิติประยุกต์

(ดร.ระวีวรรณ เอื้อพันธ์วิริยะกุล)

เมษายน 2556

บทคัดย่อ

ชื่อวิทยานิพนธ์	การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์
ชื่อผู้เขียน	นางสาว กานดา แพ้วฒนากุล
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (บริหารเทคโนโลยีสารสนเทศ)
ปีการศึกษา	2555

ข้อเสนอแนะของผู้บริโภคช่วยบ่งชี้ว่าธุรกิจควรปรับปรุงในทิศทางใด แต่เนื่องจากอินเทอร์เน็ตมีบทวิจารณ์จำนวนมาก ทั้งข้อเท็จจริง ข้อคิดเห็น และข้อเสนอแนะปะปนกัน อีกทั้งโครงสร้างประโยคที่ไม่แน่นอนทำให้ยากต่อการตีความ การจำแนกประเภทข้อมูลจะช่วยให้ประมวลผลได้ดีขึ้น บทความวิจัยนี้จึงนำเสนอกระบวนการแก้ปัญหาดังกล่าว ได้แก่ (1) กระบวนการจำแนกข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น โดยเปรียบเทียบผลลัพธ์ของอัลกอริทึมต้นไม้ตัดสินใจ นาอูฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน เพื่อหาอัลกอริทึมที่เหมาะสมที่สุด (2) กระบวนการจำแนกประเภทข้อเสนอแนะ ออกเป็น 4 ประเภท ได้แก่ ข้อเสนอแนะทางตรง ข้อเสนอแนะเชิงขอร้อง ข้อเสนอแนะเชิงคำถาม และข้อเสนอแนะเชิงเงื่อนไข

การทดลองใช้บทวิจารณ์ทั้งสิ้น 2,561 ประโยค พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนแบบเคอร์เนลโพลิโนเมียล ที่มีอินพุตเวกเตอร์ประกอบด้วย คำ กับการกำกับคำที่เกิดขึ้นร่วมกันบ่งชี้ได้ผลลัพธ์การจำแนกข้อเสนอแนะดีที่สุด มีค่าความแม่นยำ 85.75% ค่าความระลึกลับ 93.62% และค่าถ่วงดุล 89.51% จากนั้นจำแนกประเภทข้อเสนอแนะและวัดประสิทธิภาพด้วยค่าเฉลี่ยแบบให้น้ำหนักทุกประเภทเท่ากัน (Micro averaging) ได้ค่าความแม่นยำ 94.94% และความระลึกลับ 94.94%

กระบวนการที่นำเสนอถือว่ามีความถูกต้องสูงสำหรับข้อเสนอแนะที่ไม่มีความกำกวม ช่วยลดระยะเวลาการอ่านบทวิจารณ์และข้อเสนอแนะลงได้

ABSTRACT

Title of Thesis	Suggestion Mining from reviewers' reviews of television programs
Author	Miss Kanda Phawattanakul
Degree	Master of Science (Information Technology Management)
Year	2012

Suggestions are important pieces of information which is an indicator to the way for business improvement; however, the customer reviews are an enormous information, including with facts, opinions and suggestions, moreover are expressed in unstructured text which it difficult for business to handle. One reason that we are able to solve these problems is that we categorize them. Thus, our purposes are study the characteristics of suggestion and proposed 2 methodologies to solve the problems: (1) Suggestions classification (2) Suggestion types classification.

Our experiment is collect customer reviews 2,561 sentences. We compare suggestions classifications performances of Decision tree, Naïve Bayes and Support Vector Machine. Our experimental results shown SVM with polynomial kernel is the best classifier by analysis term along with frequency of couple terms tagged in vector, with the Precision, Recall and F-Measure are equal to 85.75%, 93.62% and 89.51% respectively. For Suggestion type classification, we use macro averaging to measurement the performance with the Precision and Recall are equal to 94.94% and 94.94% respectively.

Results show that our suggestion mining framework has good performance for disambiguation suggestions sentences and can reduce time consumption to read all customer reviews.

กิตติกรรมประกาศ

วิทยานิพนธ์เรื่องการวิเคราะห์เหมืองข้อมูลเสนอแนะสำเร็จได้ เนื่องมาจากบุคคลหลายท่านให้ความกรุณาทั้งคำแนะนำ ความช่วยเหลือ และกำลังใจ ผู้เขียนจึงขอขอบพระคุณบุคคลดังต่อไปนี้

อาจารย์ปราโมทย์ ลีমনาม อาจารย์ประจำภาควิชาบริหารเทคโนโลยีสารสนเทศ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์นี้ ได้ให้คำแนะนำปรึกษาและชี้แนะแนวทางในการดำเนินงานที่เป็นประโยชน์ให้ลุล่วงได้ด้วยดี

ผู้ช่วยศาสตราจารย์โอม ศรีนิล และผู้ช่วยศาสตราจารย์ปรีชา วิจิตรธรรมรส คณะกรรมการสอบวิทยานิพนธ์ ที่ได้ชี้แนะแนวทางจนสำเร็จผลของการวิเคราะห์เหมืองข้อมูลเสนอแนะ

รองศาสตราจารย์กฤษณะ ไวยมัย อาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ บางเขน ให้ความอนุเคราะห์ในการเป็นประธานคณะกรรมการสอบวิทยานิพนธ์ และได้ชี้แนะแนวทางจนสำเร็จผลของการวิเคราะห์เหมืองข้อมูลเสนอแนะ

อาจารย์หัชทัย ชาญเลขา อาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ บางเขน ที่ให้ความรู้แก่นักศึกษาในวิชาการประมวลผลภาษาธรรมชาติ

และสุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และคุณภัสกร ที่วัดชานนท์ที่ได้ช่วยเหลือ ส่งเสริม สนับสนุน และเป็นกำลังใจตลอดช่วงเวลาการทำวิทยานิพนธ์

กานดา แผ้ววัฒนากุล

เมษายน 2556

สารบัญ

	หน้า
บทคัดย่อ	(3)
ABSTRACT	(4)
กิตติกรรมประกาศ	(5)
สารบัญ	(6)
สารบัญตาราง	(8)
สารบัญภาพ	(10)
บทที่ 1 บทนำ	1
1.1 ความสำคัญและปัญหา	1
1.2 นิยามความหมาย	3
1.3 ปัญหาของงานวิจัย	7
1.4 วัตถุประสงค์	8
1.5 ประโยชน์ที่คาดว่าจะได้รับ	8
1.6 ขอบเขตการดำเนินงานวิจัย	9
1.7 ภาพรวมเอกสารงานวิจัย	9
บทที่ 2 ทฤษฎีและทบทวนวรรณกรรม	10
2.1 ทฤษฎีที่เกี่ยวข้อง	10
2.1.1 การแทนข้อความ	13
2.1.2 การประมวลผลภาษาธรรมชาติ	17
2.1.3 การจำแนกประเภทข้อความ	22
2.1.4 การประเมินประสิทธิภาพการจำแนกประเภทข้อความ	30
2.2 ทบทวนวรรณกรรม	32

บทที่ 3	กรอบการดำเนินงานวิจัย	35
3.1	การนิยามปัญหา	35
3.2	กรอบการวิเคราะห์เหมืองข้อเสนอนะ	39
3.2.1	กระบวนการสร้างฐานความรู้ทางภาษา	39
3.2.2	กระบวนการเตรียมข้อมูล	47
3.2.3	กระบวนการจำแนกข้อความ	52
บทที่ 4	การทดลองและวัดประสิทธิภาพของกระบวนการ	59
4.1	ข้อมูลและเครื่องมือที่ใช้ในการวิจัย	59
4.2	กระบวนการทดสอบการวิเคราะห์เหมืองข้อเสนอนะ	61
4.3	การวัดประสิทธิภาพของกระบวนการ	69
บทที่ 5	การทดลองและวัดประสิทธิภาพของกระบวนการ	83
5.1	สรุปผลการวิจัย	83
5.2	การอภิปรายผล	84
5.3	สิ่งที่ได้จากงานวิจัย	93
5.4	ปัญหาและอุปสรรค	95
5.5	ข้อเสนอแนะ	95
บรรณานุกรม		97
ภาคผนวก		101
	ภาคผนวก ก ตัวอย่างคำที่ไม่มีนัยสำคัญ (Stop words)	102
	ภาคผนวก ข ตัวอย่างคำที่เกิดขึ้นร่วมกันบ่อย (Association Wordlist)	103
	ภาคผนวก ค ตัวอย่างคำเฉพาะเจาะจงที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (Domain Wordlist)	104
ประวัติผู้เขียน		105

สารบัญตาราง

ตารางที่	หน้า	
1.1	ประเภท ความหมาย และคำบ่งชี้ข้อเสนอแนะ	7
2.1	สัญลักษณ์ที่ใช้ในการกำกับหน้าที่ของคำและความหมายของสัญลักษณ์	20
2.2	เปรียบเทียบประสิทธิภาพการค้นคืนข้อมูล	30
3.1	ประเภทและความหมายของข้อเสนอแนะ	37
3.2	รูปแบบของประโยคข้อเสนอแนะแบ่งตามประเภท	38
3.3	ตัวอย่างคำระบุนามตามหัวข้อที่สนใจ	42
3.4	คำบ่งชี้ข้อเสนอแนะ	42
3.5	ตัวอย่างคำกริยาแบบเฉพาะเจาะจง	42
3.6	ตัวอย่างคู่ของคำนามและคำกริยา (AW)	45
3.7	ตัวอย่างคำเฉพาะเจาะจงที่เกิดขึ้นบ่อย (DW)	46
3.8	ตัวอย่างการแปลงข้อความให้เป็นพีเจอร์เวกเตอร์ด้วยค่า TF-IDF	52
4.1	ตัวอย่างข้อความที่ผ่านการเลือกคุณลักษณะในเบื้องต้น	63
4.2	ตัวอย่างการแทนข้อความด้วยฐานความรู้ทางภาษา	64
4.3	การแทนข้อความด้วยค่า TF-IDF ของคำ (t) และหน้าที่ของคำ (p)	67
4.4	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ	70
4.5	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ ด้วยวิธีการทดสอบการแทนข้อความ 5 วิธี	72
4.6	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนแบบเส้นตรง เมื่อมีการปรับค่าพารามิเตอร์ C ที่แตกต่างกัน	74
4.7	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนแบบฟังก์ชันเคอร์เนล Radial basic เมื่อมีการปรับค่าพารามิเตอร์ C และ γ ที่แตกต่างกัน	74

4.8	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอนั้นด้วยอัลกอริทึมซัพพอร์ต เวกเตอร์แมชชีนแบบฟังก์ชันเคอร์เนล Polynomial เมื่อมีการปรับ ค่าพารามิเตอร์ C , γ และ degree ที่แตกต่างกัน	75
4.9	ตารางแสดงประสิทธิภาพการสกัดข้อเสนอนั้นด้วยอัลกอริทึมซัพพอร์ต เวกเตอร์แมชชีน แบบเคอร์เนล Polynomial ที่ $C = 1, \gamma = 1$ และ $\text{degree} = 1$	75
4.10	ประสิทธิภาพการจำแนกประเภทข้อเสนอนั้น	76
4.11	รูปแบบของประโยคข้อเสนอนั้นแบ่งตามประเภท	77
4.12	แสดงตัวอย่างประโยคข้อเสนอนั้นและวลีข้อเสนอนั้นที่สกัดได้จาก รูปแบบของประโยคข้อเสนอนั้นที่แบ่งตามประเภท	78
5.1	ตัวอย่างคู่ของคำนามและคำกริยา (DW)	84
5.2	ตัวอย่างส่วนประกอบของประโยคข้อเสนอนั้นและประโยคที่มี ค่าความถูกต้องเชิงบวก	88
5.3	ตัวอย่างส่วนประกอบของประโยคข้อเสนอนั้นและประโยคที่มี ค่าความผิดพลาดเชิงบวก	89
5.4	ตัวอย่างส่วนประกอบของประโยคข้อเสนอนั้นและประโยคที่มี ค่าความผิดพลาดเชิงลบ	91
5.5	ตัวอย่างส่วนประกอบของประโยคข้อเสนอนั้นและประโยคที่มี ค่าความถูกต้องเชิงลบ	92

สารบัญภาพ

ภาพที่	หน้า
1.1	4
1.2	6
2.1	11
2.2	16
2.3	23
2.4	25
2.5	25
2.6	27
2.7	27
2.8	28
Polynomial	
2.9	28
Radial basic	
3.1	40
3.2	43
3.3	44
3.4	44
ด้วยโปรแกรม Rapid miner	
3.5	46
3.6	48
3.7	53
3.8	53
3.9	54
3.10	57

3.11	กระบวนการทดสอบแบบจำลองการทำเหมืองข้อเสนอแนะ 3 กระบวนการ	58
4.1	เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะด้วยจำนวนชุดข้อมูลเรียนรู้ที่แตกต่างกัน	60
4.2	กระบวนการหลักสำหรับการวิเคราะห์เหมืองข้อเสนอแนะ	62
4.3	การแทนข้อความด้วยค่า TF-IDF ด้วยโปรแกรม Rapid miner	66
4.4	กระบวนการวัดประสิทธิภาพงานวิจัย	70
4.5	ภาพเปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ	73
5.1	การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อเสนอแนะ	86
5.2	การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความถูกต้องเชิงบวก	87
5.3	การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความผิดพลาดเชิงบวก	89
5.4	การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความผิดพลาดเชิงลบ	90
5.5	การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความถูกต้องเชิงลบ	92

บทที่ 1

บทนำ

1.1 ความสำคัญและปัญหา

การวิเคราะห์เหมืองข้อความ (Text mining) ได้รับความนิยมอย่างมากในปัจจุบัน เนื่องจาก 90% ของปริมาณข้อมูลมหาศาลที่อยู่บนอินเทอร์เน็ตเป็นข้อมูลที่มีโครงสร้างไม่แน่นอน (ชูชาติ หฤไชยะศักดิ์, 2554) ซึ่งหมายถึงข้อความหรือภาษาธรรมชาติที่มนุษย์ใช้สื่อสารและแลกเปลี่ยนประสบการณ์ร่วมกัน ข้อความปริมาณมากเหล่านั้นมักมีข้อมูลที่เป็นประโยชน์ซ่อนอยู่ อาทิเช่น บทวิจารณ์ของผู้บริโภค (Customer reviews) ที่ปรากฏทางบล็อก สมุดเยี่ยมชมเว็บไซต์ หรือสื่อสังคมออนไลน์ต่าง ๆ เป็นต้น

บทวิจารณ์ของผู้บริโภคประกอบด้วยข้อความประเภทข้อเท็จจริง (Facts) ความคิดเห็น (Opinions) และข้อเสนอแนะ (Suggestions) ซึ่งมักมีข้อมูลสำคัญที่ธุรกิจจำเป็นต้องค้นหาความต้องการของผู้บริโภคที่ซ่อนอยู่ และตอบสนองต่อความต้องการนั้น โดยเฉพาะอย่างยิ่งข้อเสนอแนะที่จะช่วยบ่งชี้ว่าผู้บริโภคให้ความสำคัญกับเรื่องใด และธุรกิจควรปรับปรุงไปในทิศทางใด แต่ทว่าข้อเสนอแนะถูกปะปนอยู่กับบทวิจารณ์ประเภทอื่น ทำให้ยากต่อการนำไปใช้ประโยชน์ อีกทั้งภาษาที่ใช้แสดงความคิดเห็นและข้อเสนอแนะเป็นภาษาธรรมชาติที่มีโครงสร้างไม่แน่นอน ทำให้ต้องใช้เวลานานในการค้นหาข้อเสนอแนะที่ซ่อนบนข้อมูลปริมาณมากและยากต่อการตีความเพื่อนำไปใช้ประโยชน์

เทคนิคการวิเคราะห์เหมืองข้อความ และการประมวลภาษาธรรมชาติ (Natural Language Processing: NLP) ถูกนำมาประยุกต์ใช้ในกระบวนการวิเคราะห์บทวิจารณ์ ซึ่งงานวิจัยส่วนใหญ่มุ่งเน้นการวิเคราะห์ความรู้สึก (Sentiment analysis) หรือเหมืองความคิดเห็น (Opinion mining) ที่เป็นความคิดเห็นทางตรงเท่านั้น (Explicit opinions) โดยวิเคราะห์เพียงว่าผู้บริโภคมีความคิดเห็นเชิงบวก ลบ หรือกลาง และจึงสรุปว่าผู้บริโภคชอบหรือไม่ชอบสินค้าหรือคุณลักษณะของสินค้า ซึ่งความคิดเห็นที่นำมาวิเคราะห์ต้องมีคำแสดงชี้แจงแสดงความคิดเห็น (Polar word) ที่ชัดเจน (Peter, 2002; Mingqing Hu and Bing Liu, 2004a; Bo Pang and Lillian Lee, 2008; Alisa Kongthon,

Niran Angkawattanawit, Chatchawal Sangkeetrakarn, Pompimon Palingoon and Choochart Haruechaiyasak, 2010) แต่สำหรับข้อเสนอแนะที่ไม่มีข้อความคิดเห็นที่ชัดเจนได้ถูกละเลยไป ตัวอย่างเช่น “รายการนี้มีแต่พิธีกรเก่ง ๆ แต่คิดว่าควรปรับปรุงเรื่องการใช้ภาษาให้ถูกต้องอีกนิด” ประโยคที่มีคำแสดงข้อความคิดเห็นที่ชัดเจนคือ “พิธีกรเก่ง” แต่ข้อเสนอแนะเรื่อง “ควรปรับปรุงการใช้ภาษา” กลับไม่ถูกนำมาวิเคราะห์ ซึ่งในงานวิจัยของ Amar Viswanathan, Prasanna Venkatesh, Bintu Vasudevan, Rajesh Balakrishnan and Lokendra Shastri (2011) กล่าวว่า 20-30% ของบทวิจารณ์จะมีข้อเสนอแนะซ่อนอยู่

ดังนั้นการวิเคราะห์ข้อเสนอแนะจำเป็นต้องค้นหาบทวิจารณ์ประเภทข้อเสนอแนะที่ปะปนอยู่กับบทวิจารณ์ประเภทอื่นที่ไม่เกี่ยวข้องกับข้อเสนอแนะออกมาให้ได้ก่อน แต่เนื่องจากจำนวนบทวิจารณ์ที่มีอยู่มากบนอินเทอร์เน็ต อีกทั้งรูปแบบของภาษาที่ใช้แสดงข้อเสนอแนะมีโครงสร้างที่ไม่แน่นอน การใช้คนในการค้นหาและอ่านเพื่อวิเคราะห์และตีความข้อเสนอแนะจะต้องใช้เวลาและทรัพยากรจำนวนมาก จะทำอย่างไรให้สามารถจำแนกข้อเสนอแนะเพื่อช่วยลดระยะเวลาการค้นหา การอ่านเพื่อวิเคราะห์และการตีความทำได้อย่างรวดเร็วมากยิ่งขึ้น

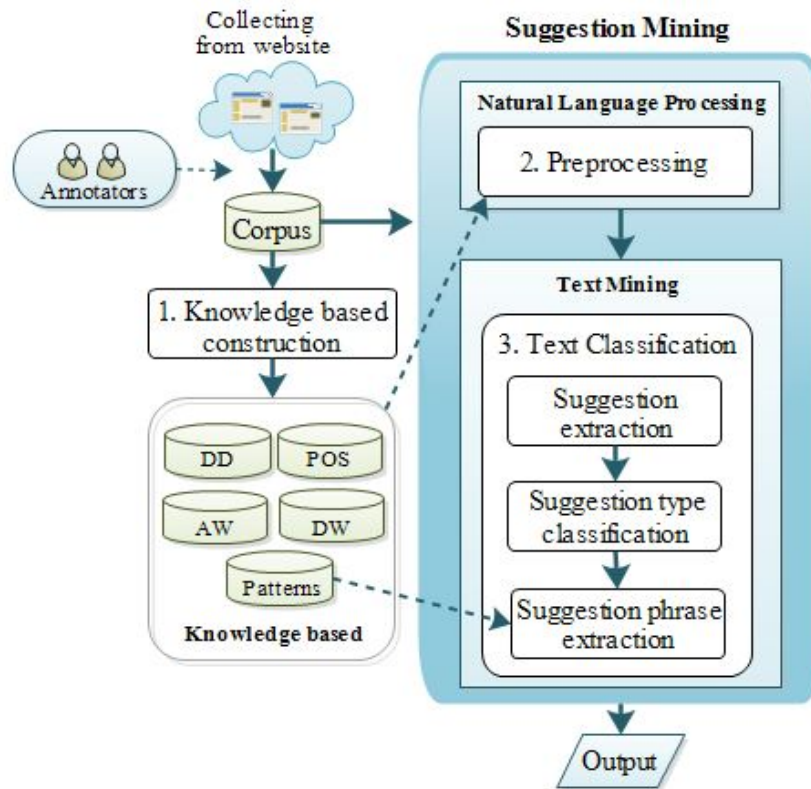
งานวิจัยนี้จึงศึกษาข้อเสนอแนะและนำเสนอกระบวนการแก้ปัญหาแบ่งเป็น 3 กระบวนการหลักได้แก่ (1) กระบวนการสกัดข้อเสนอแนะ (Suggestion extraction) (2) กระบวนการจำแนกประเภทข้อเสนอแนะ (Suggestion type classification) และ (3) กระบวนการสกัดวลีข้อเสนอแนะ สำหรับกระบวนการแรกคือกระบวนการสกัดข้อเสนอแนะเป็นกระบวนการจำแนกประโยคข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น (ข้อเท็จจริงและความคิดเห็นทั่วไป) กระบวนการนี้จะช่วยให้สามารถจำแนกข้อเสนอแนะออกจากข้อความปริมาณมากบนอินเทอร์เน็ตได้ โดยไม่จำเป็นต้องใช้ทรัพยากรด้านคน เวลาและค่าใช้จ่ายจำนวนมาก กระบวนการดังกล่าวใกล้เคียงกับงานวิจัย Vishwanath and Aishwarya (2011) ที่มีเป้าหมายในการจำแนกข้อเสนอแนะออกจากความคิดเห็นทั่วไป โดยใช้วิธีวิศวกรรมองค์ความรู้ (Knowledge engineering approach) ด้วยเทคนิคการใช้ผู้เชี่ยวชาญสร้างกฎการตัดสินใจจำแนกเอกสารหรือข้อความ (Document/Text classification) ซึ่งวิธีนี้ถือว่ามีความถูกต้องสูง เนื่องจากใช้คนในการสร้างกฎการจำแนก แต่มีข้อเสียคือเมื่อมีปริมาณข้อมูลมากขึ้นต้องใช้ระยะเวลาในการสร้างกฎ และเมื่อโดเมนข้อมูลเปลี่ยนไปจำเป็นต้องสร้างกฎการตัดสินใจใหม่ทุกครั้ง ดังนั้นในงานวิจัยนี้ได้นำเสนอวิธีการจำแนกข้อเสนอแนะด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning: ML) โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อเสนอแนะของอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree: DT), นาอิวเบย์ (Naïve Bayes: NB) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เพื่อหาอัลกอริทึมที่เหมาะสมกับการจำแนกข้อเสนอแนะมากที่สุด พร้อมกับนำเสนอวิธีการสร้างฐานความรู้ทางคำศัพท์

(Construction knowledge based) และการแทนข้อความ (Text representation) ที่นอกเหนือจากการวิเคราะห์เวกเตอร์ของ “คำ” เพียงอย่างเดียว ซึ่งจะช่วยให้ประสิทธิภาพการจำแนกข้อเสนอแนะดียิ่งขึ้น และ (2) กระบวนการจำแนกประเภทข้อเสนอแนะ (Suggestion type classification) เป็นกระบวนการจำแนกประโยคข้อเสนอแนะออกตามรูปแบบการใช้ภาษาและเจตนาการส่งสารงานวิจัยนี้ได้จำแนกประเภทของข้อเสนอแนะออกเป็น 3 ประเภท ได้แก่ ข้อเสนอแนะทางตรง (Explicit suggestion: S_e) หมายถึงข้อเสนอแนะที่มีความต้องการให้ปฏิบัติตามอย่างชัดเจน ข้อเสนอแนะเชิงคำถาม (Query suggestion: S_q) หมายถึงการตั้งคำถามถึงสิ่งที่ต้องการให้ปฏิบัติ และข้อเสนอแนะเชิงเงื่อนไข (Condition suggestion: S_c) หมายถึงการแสดงข้อเสนอแนะแบบมีเงื่อนไขการปฏิบัติ ซึ่งกระบวนการจำแนกประเภทข้อเสนอแนะนี้มีวัตถุประสงค์เพื่อจะช่วยให้การสกัดหัวข้อข้อเสนอแนะที่ซ่อนอยู่ภายในประโยคมีประสิทธิภาพมากยิ่งขึ้น ดังงานวิจัยของ Amar Viswanathan et al. (2011) ที่เสนอกระบวนการวิเคราะห์เหมืองข้อเสนอแนะ (Suggestion mining) เพื่อสกัดหัวข้อข้อเสนอแนะที่ซ่อนอยู่ภายในประโยคข้อเสนอแนะ ด้วยวิธีการสร้างกฎหรือรูปแบบ (Patterns) ของวลีข้อเสนอแนะสำหรับภาษาอังกฤษ แต่ภาษาไทยมีลักษณะภาษาแบบใช้คำซ้ำซ้อนบางประโยคประกอบด้วยคำบางซึ่งข้อเสนอแนะมากกว่า 1 คำ มีโอกาสให้กระบวนการสกัดหัวข้อข้อเสนอแนะผิดพลาดได้สูง ผู้วิจัยจึงได้นำเสนอกระบวนการจำแนกประเภทของข้อเสนอแนะเพื่อช่วยเพิ่มประสิทธิภาพการสกัดหัวข้อข้อเสนอแนะ และกระบวนการสุดท้ายคือกระบวนการสกัดวลีข้อเสนอแนะ เป็นกระบวนการสกัดวลีข้อเสนอแนะจากรูปแบบประโยคข้อเสนอแนะที่เกิดขึ้นบ่อยสำหรับภาษาไทย โดยมีวัตถุประสงค์เพื่อให้ได้ข้อมูลที่อ่านง่าย และสะดวกต่อการนำไปใช้งาน

กรอบงานการวิเคราะห์เหมืองข้อเสนอแนะสามารถอธิบายได้ดังภาพที่ 1.1 ซึ่งประกอบด้วย 3 กระบวนการ ได้แก่ (1) การสร้างฐานความรู้ทางคำศัพท์ (Knowledge based construction) (2) การเตรียมข้อมูล (Preprocessing) (3) การจำแนกข้อความ (Text classification) ซึ่งกระบวนการจำแนกข้อความประกอบด้วย 3 กระบวนการหลัก ได้แก่ (3.1) กระบวนการสกัดข้อเสนอแนะ (3.2) การจำแนกประเภทข้อเสนอแนะ และ (3.3) การสกัดวลีข้อเสนอแนะ

1.2 นิยามความหมาย

ข้อเสนอแนะเป็นส่วนหนึ่งของบทวิจารณ์ของผู้บริโภคโดยที่บทวิจารณ์หมายถึงการเขียนที่ประกอบด้วยข้อมูลอันเป็นข้อเท็จจริง ความคิดเห็นและข้อเสนอแนะที่แสดงอารมณ์ ความรู้สึก



ภาพที่ 1.1 กรอบงานวิจัยเหมืองข้อเสนอแนะ

ความคิด และข้อสันนิษฐานของผู้เขียนต่อเรื่องใดเรื่องหนึ่ง มักปรากฏในรูปของบทความตามสื่อสิ่งพิมพ์หรืออินเทอร์เน็ต เช่น หนังสือพิมพ์ วารสาร นิตยสาร บล็อก ฟอรัม สมุดเยี่ยมชมเว็บไซต์ และสื่อสังคมออนไลน์ต่างๆ เป็นต้น ซึ่งบทวิจารณ์ประกอบด้วยข้อมูล 3 ประเภท ได้แก่ ข้อเท็จจริง ความคิดเห็น และข้อเสนอแนะ ดังภาพที่ 1.2 ความหมายของประเภทบทวิจารณ์ มีดังนี้

1. ข้อเท็จจริง

พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ได้ให้ความหมายไว้ว่า หมายถึง ข้อความหรือเหตุการณ์ที่เป็นมาหรือที่เป็นอยู่ตามจริง

2. ความคิดเห็น

พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ได้ให้ความหมายไว้ว่า หมายถึง ความเห็น ข้อวินิจฉัยหรือความเชื่อที่แสดงออกตามความเห็น ู้ หรือ คิด

การเรียนรู้ภาษาไทย (2555) ได้ให้ความหมายไว้ว่า หมายถึง ความรู้สึกต่อสิ่งใดสิ่งหนึ่ง

การแสดงความคิดเห็นแบ่งเป็นความคิดเห็นในเชิงสนับสนุน เพื่อสนับสนุนความคิดเห็นของผู้อื่น ซึ่งผู้พูดอาจจะพิจารณาแล้วว่า ความคิดเห็นที่ตนสนับสนุนมีสาระและประโยชน์ต่อหน่วยงานและส่วนรวม และความคิดเห็นในเชิงขัดแย้ง เป็นการแสดงความคิดเห็นในกรณีที่มีความคิดไม่ตรงกันหรือเสนอความคิดที่ไม่ตรงกับผู้อื่น (การเรียนรู้ภาษาไทย, 2555) ซึ่งทั้งความคิดเห็นเชิงสนับสนุนและความคิดเห็นเชิงขัดแย้งนี้จะมีข้อแสดงความคิดเห็นที่ชัดเจน

3. ข้อเสนอแนะ

พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ได้ให้ความหมายไว้ว่า หมายถึง ข้อคิดเห็นเชิงแนะนำที่เสนอเพื่อพิจารณา ชี้แจงให้ทำหรือปฏิบัติตาม

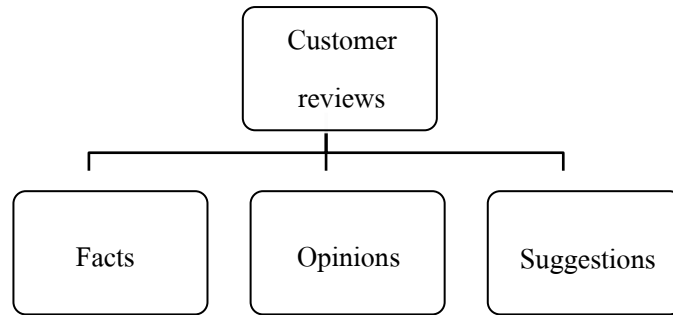
Brown Dictionary ได้นิยามความหมายข้อเสนอแนะไว้ว่า (1) The act of suggesting or state of being suggested (2) Something suggested, as a piece of advice และ (3) the calling up in the mind of one idea by another by virtue of some association or of some natural connection between the ideas

Collins English Dictionary ได้นิยามความหมายข้อเสนอแนะว่า (1) something that is suggested (2) a hint or indication และ (3) the process whereby the mere presentation of an idea to a receptive individual leads to the acceptance of that idea.

ในสารานุกรมวิกิพีเดีย ให้คำนิยามไว้ว่า in psychology, process of leading a person to respond uncritically, as in belief or action.

สามารถสรุปความหมายของข้อเสนอแนะได้ว่า หมายถึง การนำเสนอความคิดเห็นใหม่ของตนเองที่คิดว่าจะเป็นประโยชน์ต่อส่วนรวม เพื่อเป็นแนวทางให้ปฏิบัติ หรือการชี้ให้เห็นข้อบกพร่องพร้อมทั้งเสนอแนวทางแก้ไข มักแสดงความคิดเห็นในกรณีที่ไม่เห็นด้วยกับความคิดเห็นหรือการกระทำของผู้อื่น ไม่มีข้อแสดงความคิดเห็นที่ชัดเจน

ดังนั้นการวิเคราะห์ข้อเสนอแนะจึงเป็นการวิเคราะห์เพื่อทำให้ทราบว่าผู้บริโภครู้สึกถึงความสำคัญกับเรื่องใด มีข้อเสนอแนะหรือความต้องการให้ธุรกิจปรับปรุงในส่วนใด ซึ่งเป็นประโยชน์อย่างยิ่งต่อธุรกิจ



ภาพที่ 1.2 ประเภทของบทวิจารณ์

ตัวอย่างประโยคบทวิจารณ์รายการโทรทัศน์

ตัวอย่างที่ 1 : “เมื่อคืนได้ดูรายการพื้นที่ชีวิต ไม่ชอบพิธีกรเลย อยากให้ปรับปรุงเรื่องการใช้ภาษาหน่อย”

จากประโยคตัวอย่างที่ 1 สามารถจำแนกประเภทของบทวิจารณ์ได้ ดังนี้

ข้อเท็จจริง : “เมื่อคืนได้ดูรายการพื้นที่ชีวิต”

ความคิดเห็น : “ไม่ชอบพิธีกรเลย”

ข้อเสนอแนะ : “อยากให้ปรับปรุงเรื่องการใช้ภาษาหน่อย”

ตัวอย่างประโยคข้อเสนอแนะ

ตัวอย่างที่ 2 : “อยากให้เพิ่มเวลารายการ กินอยู่คือ เป็น 1 ชม.ค่ะ”

ตัวอย่างที่ 3 : “ถ้าเป็นไปได้อยากให้เปลี่ยนเวลาริรัน ASEAN FOCUS เป็นตอนกลางวัน ๗ ใต้ไหมคะ”

ตัวอย่างที่ 4 : “รายการดี ๆ แบบ นังพาไป ทำไมถอดออกล่ะ”

ประเภทของข้อเสนอแนะจำแนกออกเป็น 3 ประเภท ตามรูปแบบการใช้ภาษาและเจตนาของการเสนอแนะได้แก่ ข้อเสนอแนะทางตรง ข้อเสนอแนะเชิงคำถาม และข้อเสนอแนะเชิงเงื่อนไข ดังตารางที่ 1.1

ตารางที่ 1.1 ประเภท ความหมาย และคำบ่งชี้ข้อเสนอแนะ

สัญลักษณ์	ความหมาย	คำบ่งชี้ข้อเสนอแนะ
S_c	ข้อเสนอแนะทางตรง หมายถึงข้อเสนอแนะที่มีคำบ่งชี้ข้อเสนอแนะที่ชัดเจน ในการแสดงความคิดเห็นในกรณีที่ไม่เห็นด้วยกับความคิดเห็นหรือการกระทำของผู้อื่น และนำเสนอความคิดเห็นใหม่ของตนเองเพื่อเป็นแนวทางให้ปฏิบัติ รูปแบบประโยคประกอบด้วยคำบ่งชี้ข้อเสนอแนะที่ชัดเจน และมีเจตนาให้ปฏิบัติตามข้อเสนอแนะดังกล่าว	อยาก, ขอเสนอแนะ, ดี, น่าจะ, ครอบคลุม, ควร, พิจารณา
S_q	ข้อเสนอแนะเชิงคำถาม หมายถึงข้อเสนอแนะที่มีรูปแบบประโยคที่ประกอบด้วยคำบ่งชี้ข้อเสนอแนะในเชิงคำถามอยู่ในประโยค และมีเจตนาการตั้งคำถามถึงสิ่งที่ต้องการให้ปฏิบัติ	ทำไม, ได้ไหม
S_c	ข้อเสนอแนะเชิงเงื่อนไข หมายถึงข้อเสนอแนะที่มีรูปแบบประโยคที่ประกอบด้วยคำบ่งชี้ข้อเสนอแนะในเชิงเงื่อนไขอยู่ในประโยค และมีเจตนาในการแสดงข้อเสนอแนะแบบมีเงื่อนไขของการปฏิบัติ	ถ้า, หาก

1.3 ปัญหาของงานวิจัย

1. ข้อเสนอแนะที่ถูกปะปนอยู่กับข้อมูลที่ไม่เกี่ยวข้อง สามารถลดระยะเวลาการค้นหา และจำแนกออกมาได้อย่างไร
2. ข้อเสนอแนะที่เป็นภาษาธรรมชาติที่มีโครงสร้างไม่แน่นอน จะทำอย่างไรเพื่อช่วยในการอ่านเพื่อวิเคราะห์และตีความทำได้อย่างรวดเร็ว

1.4 วัตถุประสงค์

1. เพื่อศึกษาข้อเสนอแนะสำหรับภาษาไทยและนำเสนอกระบวนการทำเหมืองข้อเสนอแนะด้วยเทคนิคการทำเหมืองข้อความ ร่วมกับการประมวลผลภาษาธรรมชาติ
2. เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการจำแนกข้อเสนอแนะ ได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ นาอูฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธีการเปรียบเทียบค่าความแม่นยำ (Precision) ค่าระลึก (Recall) และประสิทธิภาพโดยรวมของระบบ (F-measure) ของผลลัพธ์ที่ได้จากแต่ละอัลกอริทึม
3. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแทนข้อความด้วยฐานความรู้ทางภาษาของแต่ละวิธี เพื่อจำแนกประเภทข้อเสนอแนะ ซึ่งได้นำเสนอ 5 วิธี ได้แก่ (1) วิธีการเลือกคุณลักษณะสำคัญจากคำเพียงอย่างเดียว (2) วิธีการเลือกคุณลักษณะสำคัญจากคำและหน้าที่ของคำ (3) วิธีการเลือกคุณลักษณะสำคัญจากคำ หน้าที่ของคำและคำกริยาแบบเฉพาะเจาะจง (4) วิธีการเลือกคุณลักษณะสำคัญจากคำและคู่ของคำที่เกิดขึ้นร่วมกันบ่อย ด้วยเทคนิคกฎความสัมพันธ์ (Association rules) และ (5) วิธีการเลือกคุณลักษณะสำคัญจากคำที่เกิดขึ้นบ่อยภายใต้โดเมน (Domain) ที่ใกล้เคียงกัน และเปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วยค่าเฉลี่ยแบบให้น้ำหนัก (Averaging) ของผลลัพธ์ที่ได้จากแต่ละวิธี

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. กระบวนการวิเคราะห์เหมืองข้อเสนอแนะ ที่ช่วยลดระยะเวลาการค้นหาและวิเคราะห์ข้อเสนอแนะจากข้อมูลปริมาณมากได้
2. อัลกอริทึมที่เหมาะสมสำหรับการวิเคราะห์ข้อเสนอแนะ
3. วิธีการแทนข้อความด้วยฐานความรู้ทางภาษาที่เหมาะสมสำหรับกระบวนการวิเคราะห์ข้อเสนอแนะ

1.6 ขอบเขตการดำเนินงาน

1. ศึกษารูปแบบการแสดงผลข้อเสนอแนะที่เขียนเป็นภาษาไทยบนเว็บไซต์ เพื่อนำเสนอกระบวนการจำแนกข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น ๆ และกระบวนการจำแนกข้อเสนอแนะตามประเภท เพื่อให้ได้ข้อมูลสารสนเทศที่อยู่ในรูปแบบที่ง่ายต่อการพิจารณา

2. ขอบเขตของข้อมูลที่ศึกษาคือ บทวิจารณ์ที่กล่าวถึงรายการโทรทัศน์ ซึ่งนำมาจากสมุดเยี่ยมชมเว็บไซต์ สื่อสังคมออนไลน์เฟสบุ๊คและทวิตเตอร์ของสถานีโทรทัศน์ ซึ่งเป็นข้อมูลตั้งแต่ช่วงเดือนมิถุนายน พ.ศ.2554 ถึงเดือนกุมภาพันธ์ พ.ศ.2555 รวมถึงเว็บบอร์ดต่าง ๆ ที่มีการแสดงบทวิจารณ์เกี่ยวกับรายการโทรทัศน์อื่น ๆ รวมทั้งสิ้น 1,105 เอกสาร แบ่งออกเป็น 2,561 ประโยค ประกอบด้วยข้อเสนอแนะ 787 ประโยค และบทวิจารณ์ประเภทอื่นที่ไม่เป็นข้อเสนอแนะ 1,774 ประโยค

3. การทำเหมืองข้อเสนอแนะอ้างอิงผลการตัดคำภาษาไทยจาก Java API ชื่อ BreakIterator และใช้การกำกับหน้าที่ของคำจากคลังคำไทยเล็กซิตรอน (Lexitron) และคลังคำออร์คิด (Orchid) ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ผลลัพธ์ที่ได้จากการตัดคำและการกำกับคำจะให้ผู้เชี่ยวชาญตรวจสอบและแก้ไขความถูกต้องก่อนนำเข้าสู่กระบวนการวิเคราะห์เหมืองข้อเสนอแนะ

4. ผลลัพธ์ที่ได้จากการทดลองนำไปเปรียบเทียบความถูกต้องกับผลการวิเคราะห์โดยผู้เชี่ยวชาญ

1.7 ภาพรวมเอกสารงานวิจัย

บทที่ 1 อธิบายความสำคัญและปัญหาของงานวิจัยเหมืองข้อเสนอแนะ รวมถึงการนิยามคำศัพท์ที่เกี่ยวข้องกับงานวิจัย บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง บทที่ 3 อธิบายกรอบการดำเนินงานวิจัย บทที่ 4 การทดลองและวัดประสิทธิภาพของกระบวนการทำเหมืองข้อเสนอแนะ และบทที่ 5 สรุปผลการทดลองและข้อเสนอแนะ

บทที่ 2

ทฤษฎีและทบทวนวรรณกรรม

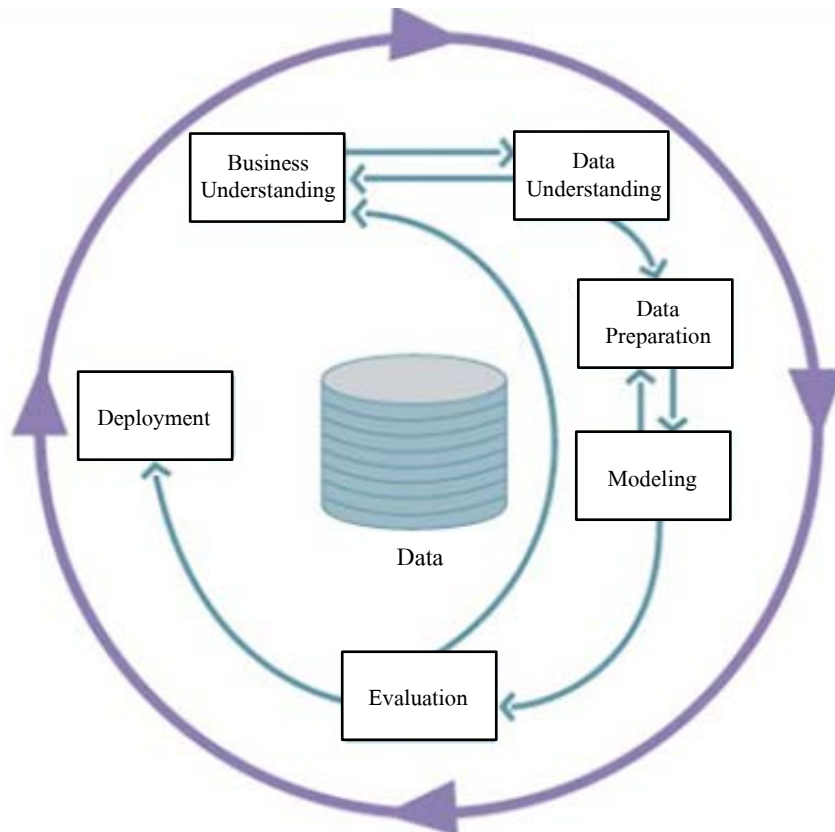
บทนี้กล่าวถึงทฤษฎีที่เกี่ยวข้องที่นำมาประยุกต์ใช้ในกระบวนการวิเคราะห์เหมืองข้อมูล ซึ่งประกอบด้วย 4 ทฤษฎี ได้แก่ (1) การแทนข้อความ (2) การประมวลผลภาษาธรรมชาติ (3) การจำแนกประเภทข้อความและ (4) การประเมินประสิทธิภาพการจำแนกประเภทข้อความ และหัวข้อสุดท้ายของบทนี้ได้กล่าวถึงวรรณกรรมที่เกี่ยวข้องกับงานวิจัย

2.1 ทฤษฎีที่เกี่ยวข้อง

ข้อเสนอแนะที่แสดงอยู่ในบทวิจารณ์มีรูปแบบเป็นข้อความหรือภาษาที่มีโครงสร้างไม่แน่นอน มีความแตกต่างเฉพาะตัวที่เป็นไปตามธรรมชาติของการเรียนรู้ในสมองมนุษย์แต่ละคน (กนกวรรณ เขียววรรณ, 2555) การวิเคราะห์ข้อความที่คอมพิวเตอร์พยายามทำความเข้าใจกับภาษาธรรมชาติของมนุษย์ ได้นำองค์ความรู้ในด้านมาประยุกต์ใช้ เช่น การค้นหาลักษณะแฝงของข้อมูล (Knowledge discovery) หรือการทำเหมืองข้อมูล (Data mining) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เป็นต้น (Andreas, Gerhard, Fraunhofer and Sankt, 2005) สำหรับการวิเคราะห์ข้อความเป็นกระบวนการค้นหาและสกัดความรู้จากฐานข้อมูลขนาดใหญ่ (Large textual information) เพื่อให้ได้สารสนเทศที่มีประโยชน์ (Useful textual information) โดยข้อมูลที่น่ามาวิเคราะห์เป็นข้อมูลที่มีลักษณะเป็นข้อความ (Text data sets) หรือภาษาธรรมชาติ (Natural language) จึงเรียกว่าการวิเคราะห์เหมืองข้อความ (Text mining) ซึ่งได้นำองค์ความรู้ในด้านการค้นหาลักษณะแฝงของข้อมูลมาประยุกต์ใช้ในการวิเคราะห์

การค้นหาลักษณะแฝงของข้อมูล (Knowledge Discovery Data: KDD) หรือการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Large database) เป็นกระบวนการค้นหารูปแบบ โครงสร้าง ความสัมพันธ์ หรือการเปลี่ยนแปลงที่แฝงอยู่ของข้อมูล ซึ่งข้อมูลที่น่ามาวิเคราะห์เป็นได้ทั้งฐานข้อมูล ข้อความ หรือแม้กระทั่งรูปภาพ การวิเคราะห์เหมืองข้อมูลเป็นกระบวนการหลักในการ

ค้นหาลักษณะแฝงของข้อมูล (Knowledge Discovery) หรือบางครั้งการวิเคราะห์เหมืองข้อมูลอาจหมายถึงกระบวนการค้นหาลักษณะแฝงของข้อมูลเลยก็ได้ (Andreas et al., 2005) โดยหน่วยงาน Cross Industry Standard Process for Data Mining (Crisp DM) ได้นำเสนอกระบวนการวิเคราะห์ข้อมูลประกอบด้วยขั้นตอนดังต่อไปนี้ (1) การทำความเข้าใจกับธุรกิจและระบุปัญหาของงาน (Business understanding) (2) การรวบรวมข้อมูลและพิจารณาความถูกต้องเหมาะสมของข้อมูล (Data understanding) (3) การเตรียมข้อมูล (Data preparation) (4) การสร้างแบบจำลองการวิเคราะห์ข้อมูล (Modeling) (5) การประเมินหรือวัดประสิทธิภาพของแบบจำลองการวิเคราะห์ข้อมูล (Evaluation) และ (6) การนำผลลัพธ์ที่ได้จากการวิเคราะห์ไปใช้งานจริง (Deployment) ดังภาพที่ 2.1



ภาพที่ 2.1 กระบวนการค้นหาลักษณะแฝงของข้อมูล (Crisp DM)

แหล่งที่มา: IBM Software Business Analytics, 2012.

ลักษณะของข้อมูลที่น่ามาวิเคราะห์เพื่อหาลักษณะแฝง แบ่งเป็น 2 ประเภทคือ (1) ข้อมูลที่เป็นโครงสร้าง (Structured data) การประมวลผลข้อมูลที่เป็นโครงสร้าง เรียกว่าการวิเคราะห์เหมือนข้อมูล และ (2) ข้อมูลที่ไม่เป็นโครงสร้างหรือไม่มีโครงสร้างที่แน่นอน (Unstructured or implicit structured data) ซึ่งส่วนใหญ่มักอยู่ในรูปแบบของข้อความหรือภาษาธรรมชาติ เรียกว่าการวิเคราะห์เหมือนข้อความ หรือกระบวนการค้นหาลักษณะแฝงของข้อความ (Knowledge Discovery from Text: KDT)

สำหรับการวิเคราะห์ข้อความมีกระบวนการแตกต่างจากการวิเคราะห์เหมือนข้อมูลเล็กน้อยคือข้อความ ซึ่งมีลักษณะข้อมูลแบบไม่เป็นโครงสร้างหรือไม่มีโครงสร้างที่แน่นอน คือกระบวนการวิเคราะห์ข้อความจำเป็นต้องแปลงรูปแบบของข้อความที่ไม่มีโครงสร้างให้เป็นโครงสร้างก่อน (Ronen and James, 2007) เรียกว่าเป็นกระบวนการเตรียมข้อมูล (Data preparation) ขั้นตอนการเตรียมข้อมูลในกระบวนการวิเคราะห์ข้อความถือเป็นขั้นตอนที่สำคัญและต้องใช้เวลาาน ซึ่งแตกต่างจากการวิเคราะห์เหมือนข้อมูลทั่วไป เพื่อให้ได้ข้อมูลที่มีโครงสร้างที่เหมาะสมสำหรับขั้นตอนวิธี (Algorithm) ที่จะใช้ในการวิเคราะห์ กระบวนการเตรียมข้อมูลเกี่ยวข้องกับเทคนิคการประมวลผลภาษาธรรมชาติ ที่นำมาช่วยในการเตรียมข้อความให้คอมพิวเตอร์สามารถทำความเข้าใจกับคำและประโยคในภาษาธรรมชาติได้มากขึ้น และผลลัพธ์ที่ได้จากการค้นหาลักษณะแฝงของข้อมูลคือองค์ความรู้ (Knowledge) หรือสารสนเทศที่เป็นประโยชน์

ดังนั้นกระบวนการวิเคราะห์เหมือนข้อเสนอนี้ จึงประกอบด้วยการนำทฤษฎีการประมวลผลภาษาธรรมชาติ และการวิเคราะห์เหมือนข้อมูล มาประยุกต์ใช้เป็นทฤษฎีหลักในการวิเคราะห์เหมือนข้อความ ซึ่งมีขั้นตอนวิธี ดังนี้

การเตรียมข้อมูลในการค้นหาลักษณะแฝงของข้อมูลได้แบ่งเป็น 3 กระบวนการ (IBM Software Business Analytics, 2013) ได้แก่ การคัดเลือกข้อมูล (Data selection), การกลั่นกรองข้อมูล (Data cleaning), การสร้างข้อมูลที่เหมาะสม (Data constructing) เช่นการ Derived attributes หรือ การ Generated records เป็นต้น, การรวมข้อมูล (Integrate Data) และสุดท้ายคือการแปลงรูปข้อมูล (Data transformation) ให้อยู่ในรูปแบบที่มีโครงสร้างที่เหมาะสม

สำหรับการวิเคราะห์เหมือนข้อความในงานวิจัยนี้ได้จำแนกกระบวนการเตรียมข้อมูลออกเป็น 3 กระบวนการหลัก ได้แก่

- 1.1 การเลือกคุณลักษณะ (Feature selection)
- 1.2 การกลั่นกรองข้อความ (Text cleaning)
- 1.3 การแทนข้อความ (Text representation)

ซึ่งกระบวนการเหล่านี้ได้อ้างอิงตามหลักการเตรียมข้อมูลของเทคนิคการวิเคราะห์เหมืองข้อมูล แต่สำหรับการวิเคราะห์เหมืองข้อความต้องอาศัยทฤษฎีการประมวลผลภาษาธรรมชาติในการวิเคราะห์ร่วมด้วย

การสร้างแบบจำลองการวิเคราะห์ข้อมูล เป็นกระบวนการสร้างแบบจำลองจากข้อมูลที่มีอยู่เพื่ออธิบายรูปแบบของข้อมูลหรือทำนายรูปแบบของข้อมูลที่ยังไม่เกิดขึ้น แบ่งเป็น 2 เทคนิคหลักคือ (1) กระบวนการเรียนรู้แบบมีผู้สอน (Supervised learning) เช่น การจำแนกข้อมูล (Classification) สำหรับการวิเคราะห์ข้อความนิยมเรียกกระบวนการดังกล่าวว่า การจำแนกข้อความ (Text classification หรือ Text categorization) และ (2) กระบวนการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) เช่น การจัดกลุ่มข้อมูล (Clustering)

การประเมินประสิทธิภาพของแบบจำลองการวิเคราะห์ข้อมูล ใช้เพื่อเป็นเครื่องมือวัดความน่าเชื่อถือของแบบจำลอง

จากกระบวนการวิเคราะห์เหมืองข้อความข้างต้น เมื่อนำมาประยุกต์ใช้กับงานวิจัยเรื่องการวิเคราะห์เหมืองข้อเสนอแนะนี้ จำเป็นต้องอาศัยความรู้ 4 ทฤษฎีหลักคือ (1) การแทนข้อความ (2) การประมวลผลภาษาธรรมชาติ (3) การจำแนกประเภทข้อความ และ (4) การประเมินประสิทธิภาพการจำแนกประเภทข้อความ มีรายละเอียดดังนี้

2.1.1 การแทนข้อความ

การวิเคราะห์ข้อความที่มีโครงสร้างไม่แน่นอนจำเป็นต้องแปลงให้อยู่ในรูปแบบที่มีโครงสร้างก่อน เพื่อให้คอมพิวเตอร์สามารถนำไปประมวลผลได้ การแทนข้อความให้อยู่ในรูปแบบเวกเตอร์สเปซโมเดล (Vector Space Model: VSM) เป็นวิธีการหนึ่งในการแทนข้อความให้มีโครงสร้างแบบพีเจอร์เวกเตอร์ (Feature vector) สามารถเลือกคุณลักษณะของข้อความมาแทนได้หลายวิธี เช่น คำ วลี หรือหน้าที่ของคำ เป็นต้น แต่โดยปกติแล้วการวิเคราะห์ข้อความจะอาศัยการวิเคราะห์จากคำเป็นหลัก เรียกว่าการแทนข้อความด้วยถุงคำ (Bag-of-word) เป็นการแทนคำทุกคำในเอกสารด้วยเวกเตอร์ที่ประกอบด้วยสมาชิกคือ คำที่เป็นตัวแทนของคำภายในข้อความนั้น มีวิธีที่ใช้ในการคำนวณค่าเพื่อกำหนดให้เป็นตัวแทนข้อความ (Atom Nuntiyagul, 2006) เช่น

- การแทนคำด้วยค่าการเกิดขึ้นหรือไม่เกิดขึ้นของคำ (Binary weighting) ดังสมการ (1)

$$\text{binary weighting} = \begin{cases} 1, & \text{for term present in the document} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

- การแทนด้วยค่าความถี่ของคำ (Term Frequency: TF)
- ค่าความถี่คำ-ค่าส่วนกลับความถี่เอกสารที่เกิดคำ (Term Frequency – Inverse Document Frequency: TF-IDF) ซึ่งวิธีการแทนค่าข้อความด้วยค่า TF-IDF เป็นวิธีที่ได้รับความนิยมมากที่สุด เนื่องจากเป็นวิธีการคำนวณที่ง่ายและมีประสิทธิภาพสูง

2.1.1.1 Term Frequency – Inverse Document Frequency (TF-IDF)

แนวความคิดของการแทนข้อความด้วยค่า TF-IDF เกิดจากแนวความคิดว่าการแทนข้อความด้วยค่าความถี่ของคำเพียงอย่างเดียว ไม่สามารถจำแนกเอกสาร (หรือข้อความ) ได้ดีพอ เนื่องจากค่าความถี่ของคำสูง หมายถึงคำนั้นมีโอกาสเกิดขึ้นในหลายเอกสารพร้อมกัน จึงไม่มีประโยชน์ต่อการจำแนกเอกสาร ดังนั้น Salton and Buckley (1988) จึงนำเสนอวิธีการแทนค่าเอกสารด้วยค่า TF-IDF คือคำนึงถึงความถี่ของคำในเอกสารด้วย โดยค่านำหนักของคำที่ได้ คือ w_d เกิดจากการคูณค่า TF ด้วย IDF ดังสมการ (2)

$$w_d = f_{w,d} * \log\left(\frac{|D|}{f_{w,d}}\right) \quad (2)$$

เมื่อ $f_{w,d}$ หมายถึง ความถี่ของคำ (Term Frequency: TF) เป็นค่าที่เกิดจากการคำนวณค่าความถี่ของคำ (w) ที่พบภายในเอกสาร (d) และค่า Logarithmic scale ของสัมประสิทธิ์ของจำนวนเอกสารทั้งหมดที่นำมาวิเคราะห์ (D)หารด้วยค่าความถี่เอกสารที่พบคำ (w) ซึ่งหมายถึงค่าส่วนกลับของเอกสาร (Inverse Document Frequency: IDF) หรือค่าส่วนกลับความถี่ของทุกเอกสาร (D) ที่ปรากฏคำ (w) (Juan, 1999)

สำหรับการวิเคราะห์ข้อความในระดับประโยคจะแทนเอกสารด้วยประโยค ดังนั้นค่า TF หมายถึงค่าที่เกิดจากการคำนวณค่าความถี่ของคำ (w) ที่พบในประโยค (s) และค่า IDF หมายถึงค่า Logarithmic scale ของสัมประสิทธิ์ของจำนวนประโยคทั้งหมดที่นำมาวิเคราะห์ (D) หารด้วยค่าความถี่ของเอกสารที่พบคำ (w)

อย่างไรก็ตาม การแทนข้อความด้วยวิธีการคำนวณค่าจากคำทั้งหมดอาจไม่เหมาะสมนัก เนื่องจากมีคำบางคำที่ไม่มีประโยชน์ต่อการจำแนกข้อความ จึงควรเลือกคุณลักษณะของข้อความที่สามารถใช้เป็นตัวแทนเอกสารหรือข้อความที่ดี และตัดคำที่ไม่สามารถใช้ตัวแทนเอกสารหรือข้อความที่ดีได้ออก การตัดคำไม่มีนัยสำคัญและการทำรากศัพท์ถือเป็นวิธีการเลือกคุณลักษณะเบื้องต้นที่ดีและได้รับความนิยมเพื่อจะช่วยเหลือเพิ่มประสิทธิภาพการจำแนกข้อความ

นอกจากนั้นแล้วยังสามารถลดขนาดของเนื้อที่เก็บข้อมูลได้มากถึง 30-50% (การวิเคราะห์ข้อความ, 2555)

การตัดคำที่ไม่มีนัยสำคัญ (Stop word removal) หมายถึงคำที่เกิดขึ้นบ่อยในหลายเอกสารและไม่เป็นประโยชน์ต่อการจำแนกประเภทเอกสาร หากตัดคำเหล่านั้นออกจะไม่ทำให้ความหมายของประโยคเปลี่ยนไป เช่น คำเชื่อม (Conjunction) คำบุพบท (Preposition) คำหยุด (Ending words) เป็นต้น

การทำรากศัพท์ (Word stemming) หมายถึงคำที่มีรากศัพท์คำเดียวกันแต่มีการแปลงรูปไป เช่น “run”, “ran” และ “runs” มีรากศัพท์คำเดียวกันคือ “run” แต่สำหรับภาษาไทยไม่มีคำในลักษณะดังกล่าว มีเพียงคำที่มีความหมายคล้ายคลึงกันแต่ใช้คำต่างกัน (Synonym word) เช่น คำว่า “กิน” กับ “รับประทาน” เป็นต้น

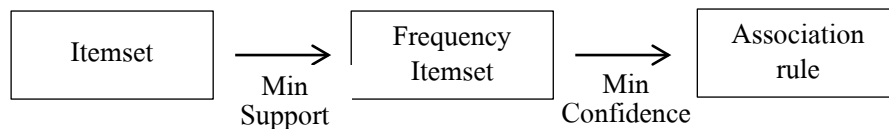
อย่างไรก็ตามกระบวนการแทนข้อความด้วยวิธี Vector Space Model ที่แทนข้อความด้วยคำเพียงคำเดียว ไม่ได้วิเคราะห์จากกลุ่มคำ หรือลำดับของคำ จึงมีงานวิจัยที่ได้นำเสนอกระบวนการในการเลือกคุณลักษณะที่พิจารณาการเกิดขึ้นร่วมกันของคำ ได้แก่ การเลือกคุณลักษณะสำคัญด้วยค่าการเพิ่มของข้อมูล (Information Gain: IG), ค่าข้อมูลร่วม (Mutual Information: MI), ค่าสถิติ Chi-square ซึ่งในงานวิจัย Jan and Yiming Yang (1997) ได้นำแต่ละวิธีมาเปรียบเทียบประสิทธิภาพการจำแนกเอกสาร ซึ่งได้แก่ การเพิ่มของข้อมูล ค่าข้อมูลร่วม, ค่าสถิติ Chi-square และค่าความถี่ของเอกสารที่เกิดคำ (Document Frequency: DF) พบว่า IG และ Chi-square ได้ผลลัพธ์การจำแนกเอกสารดีที่สุด ส่วน DF มีวิธีการคำนวณที่ง่ายกว่าวิธีอื่นและมีประสิทธิภาพการจำแนกที่ใกล้เคียงกับวิธี IG และ Chi-square

นอกจากนี้ในงานวิจัยของ Maria and Osmar (2002) และ Minqing Hu and Bing Liu (2004b) ได้นำเสนอกระบวนการเลือกคุณลักษณะสำคัญ ด้วยเทคนิคการหากฎความสัมพันธ์ (Association rules mining) เพื่อวิเคราะห์หาความสัมพันธ์ระหว่างคำหรือฟีเจอร์ (Feature) ตั้งแต่ 2 คำขึ้นไปที่เกิดขึ้นร่วมกันบ่อย

2.1.1.2 กฎความสัมพันธ์ของข้อมูล (Association rules mining)

การสร้างกฎความสัมพันธ์ของข้อมูลเป็นกระบวนการหนึ่งในการทำเหมืองข้อมูลเพื่อหาความสัมพันธ์ซึ่งกันและกันของข้อมูลตั้งแต่ 2 ชุดขึ้นไปภายในกลุ่มข้อมูลขนาดใหญ่ ซึ่งถูกนำเสนอโดย Rakesh Agrawal and Ramakrishnan Srikant (1994) ตัวอย่างหนึ่งของกฎความสัมพันธ์ที่นิยมใช้คือ การวิเคราะห์การซื้อสินค้าของลูกค้า (Market-basket analysis) ซึ่งหมายถึงการวิเคราะห์หาความสัมพันธ์ของสินค้าที่มีแนวโน้มในการซื้อร่วมกันบ่อยภายในรายการเดียวกัน

กระบวนการสร้างกฎความสัมพันธ์ของข้อมูล แบ่งเป็น 2 กระบวนการ คือ (1) การหาความสัมพันธ์ของข้อมูลที่เกิดขึ้นร่วมกัน (Frequency itemset) และวัดผลความสัมพันธ์กันระหว่างข้อมูลด้วยค่าสนับสนุน (Support) ความสัมพันธ์ของชุดข้อมูลใดที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำที่กำหนดไว้ (Minimum support) จะเรียกความสัมพันธ์ของชุดข้อมูลนั้นว่า “Frequency patterns” หรือ “Frequency itemset” และกระบวนการถัดไปคือ (2) การสร้างกฎความสัมพันธ์ (Association rules) เป็นกระบวนการนำชุดข้อมูลที่มีความสัมพันธ์กันภายใต้ค่าสนับสนุนขั้นต่ำที่กำหนด มาสร้างเป็นกฎความสัมพันธ์ ความสัมพันธ์ของชุดข้อมูลใดที่มีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence) ที่กำหนดไว้ จะเรียกความสัมพันธ์ของชุดข้อมูลนั้นว่า “Association rule” ดัง ภาพที่ 2.2



ภาพที่ 2.2 ขั้นตอนการสร้างกฎความสัมพันธ์

ค่าสนับสนุน หมายถึง เปอร์เซ็นต์ของจำนวนข้อมูลที่มีสมาชิกสอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด ยกตัวอย่างเช่น ชุดข้อมูล A กับ B จะถูกกำหนดเป็นกฎความสัมพันธ์ $A \Rightarrow B$ ก็ต่อเมื่อมีค่าสนับสนุนมากกว่าค่าเปอร์เซ็นต์ต่ำสุดที่กำหนดไว้ สมการการคำนวณค่าสนับสนุนดังสมการที่ (3) (Kenneth and Narciso, 2007)

$$support(AB) = \frac{\text{transactions contain } AB}{\text{total transactions}} \quad (3)$$

ค่าความเชื่อมั่น หมายถึง เปอร์เซ็นต์ของจำนวนข้อมูลที่สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่มีสมาชิกตามกฎฝั่งซ้ายมือ โดยที่กฎความสัมพันธ์ $A \Rightarrow B$ ต้องมีค่าความเชื่อมั่นมากกว่าค่าเปอร์เซ็นต์ต่ำสุดที่กำหนดไว้ ดังสมการที่ (4) (Kenneth et al., 2007)

$$confidence(A \rightarrow B) = \frac{support(AB)}{support(A)} \quad (4)$$

นอกจากการแทนข้อความด้วยค่าความถี่ของคำคำเดียว หรือกลุ่มคำแล้ว ยังสามารถใช้หน้าที่ของคำ เป็นตัวแทนข้อความได้เช่นเดียวกัน สำหรับงานวิจัยนี้ได้นำเสนอวิธีการแทนข้อความด้วยคำ หน้าที่ของคำและกลุ่มคำ ซึ่งการแทนข้อความด้วยกลุ่มคำได้ใช้เทคนิคกฎความสัมพันธ์ของข้อมูลมาหาความสัมพันธ์ของคำที่เกิดขึ้นร่วมกันบ่อย และคำนวณค่าตัวแทนข้อความด้วยค่า TF-IDF

2.1.2 การประมวลผลภาษาธรรมชาติ

ภาษาเป็นเครื่องมือที่ในการสื่อสารทั้งกับมนุษย์ด้วยกันเองหรือแม้กระทั่งสื่อสารกับคอมพิวเตอร์ แต่ภาษาที่ใช้มีรูปแบบที่แตกต่างกันไป ภาษาที่มนุษย์ใช้สื่อสารกับคอมพิวเตอร์ เป็นภาษาที่มีโครงสร้างแน่นอน คอมพิวเตอร์สามารถนำไปประมวลผลได้ทันที เช่น php, java หรือ C++ เป็นต้น แต่สำหรับภาษาที่มนุษย์ใช้สื่อสารกันเองเป็นภาษาที่ไม่มีโครงสร้างหรือรูปแบบเฉพาะตัว และเป็นไปตามธรรมชาติของการเรียนรู้ในสมองมนุษย์แต่ละคนซึ่งมีลักษณะที่แตกต่างกันไป เรียกว่า “ภาษาธรรมชาติ” เป็นภาษาไม่มีโครงสร้างที่แน่นอน การที่คอมพิวเตอร์พยายามทำความเข้าใจกับภาษาธรรมชาติที่มนุษย์ใช้สื่อสารกันนั้น สามารถทำได้ด้วยวิธีการแทนความรู้ การสร้างกฎเกณฑ์ และการประเมินค่าเพื่อหาความหมายของภาษา (กนกวรรณ เขียววรรณ, 2555)

ดังนั้นการที่คอมพิวเตอร์จะเข้าใจภาษาธรรมชาติได้ดีเพียงไรนั้นขึ้นอยู่กับ 2 แนวทาง หนึ่งคือพัฒนาการทางด้านปัญญาประดิษฐ์ซึ่งเป็นวิธีการแทนความรู้ (Knowledge representation) และอีกแนวทางหนึ่งคือการศึกษาศาสตร์และเข้าใจโครงสร้างทางภาษาศาสตร์แบบมีโครงสร้าง (ฮิน ฮูววรรณ, 2535) ซึ่งทั้งสองกระบวนการดังกล่าวถูกเรียกว่าการประมวลผลภาษาธรรมชาติ โดยระบบประมวลผลภาษาธรรมชาติจะรับข้อมูลอินพุตเป็นข้อความและแทนค่าข้อความด้วยแนวทางการวิเคราะห์ต่าง ๆ เช่น ความรู้ หรือโครงสร้าง เป็นต้น

การประมวลผลภาษาธรรมชาติได้แบ่งระดับขั้นการวิเคราะห์ ดังนี้ (Christopher and Hinrich, 1999)

1. การวิเคราะห์ระดับวากยสัมพันธ์ (Morphological analysis) เป็นการวิเคราะห์ระดับคำ
2. การวิเคราะห์ระดับวากยสัมพันธ์ (Syntactic analysis) เป็นการวิเคราะห์คำตามหน้าที่ของคำ (Part-of-Speech) เพื่อเป็นข้อมูลพื้นฐานในการตรวจสอบโครงสร้างทางไวยากรณ์เกี่ยวกับการวางตำแหน่งของคำ กลุ่มคำประเภทต่าง ๆ ที่รวมกันเป็นประโยค
3. การวิเคราะห์ระดับความหมาย (Semantic analysis) เป็นการวิเคราะห์เพื่อให้ทราบความหมายของคำแต่ละคำในประโยค

4. วิเคราะห์ระดับวจนิจพจน์ (Discourse integration) เป็นการพิจารณาความหมายของประโยคโดยดูจากประโยคข้างเคียงร่วมด้วย

5. การวิเคราะห์ระดับปฏิบัติ (Pragmatic analysis) เป็นการแปลความหมายของประโยคถึงสิ่งที่ผู้พูดต้องการสื่อความหมายถึง

กระบวนการวิเคราะห์ภาษาจะเริ่มต้นที่ระดับต่ำสุดก่อนคือการวิเคราะห์ระดับวจิวิภาคหรือคำ ไปจนถึงระดับวากยสัมพันธ์ ที่สามารถอธิบายได้ด้วยโครงสร้างของภาษาที่ประกอบด้วยหน่วยต่าง ๆ ดังต่อไปนี้

1. คำ (Word) คือ หน่วยหนึ่งที่เปล่งเป็นเสียงออกมาจะเป็นอิสระหรือไม่ก็ได้
2. หน่วยคำ (Morpheme) คือ ส่วนประกอบที่มีนัยสำคัญที่แท้จริงทางภาษา คือ คำและกลุ่มคำ หรือหมายถึงหน่วยคำที่เล็กที่สุดที่มีความหมาย ความแตกต่างระหว่าง หน่วยคำ กับ คำ คือ หน่วยคำอาจเกิดขึ้นอิสระหรือไม่อิสระก็ได้ แต่คำต้องเป็นหน่วยอิสระเสมอ เช่น นักเรียน “นัก” เป็นหน่วยคำ และ “นักเรียน” จึงถือเป็นคำ
3. วลี (Phrase) คือกลุ่มของคำหรือคำ ๆ เดียวก็ได้ ที่เป็นส่วนประกอบของประโยค (วลีเป็นส่วนประกอบของประโยค แต่คำไม่ใช่ส่วนประกอบของประโยค) ดังนั้นไม่ว่าจะเป็นคำ ๆ เดียวหรือกลุ่มที่ประกอบด้วยหลาย ๆ คำ เมื่อเป็นส่วนประกอบของประโยคเราเรียกเป็น วลี
4. นามวลี (Noun phrase) คือวลีที่ทำหน้าที่เป็นส่วนของประโยคที่เป็นหน่วยประธาน หน่วยกรรม นามวลี
5. กริยาวลี (Verb phrase) คือวลีที่ทำหน้าที่เป็นภาคแสดงของประโยค
6. ประโยค (Sentence) คือคำหลายคำเรียงกันในการพูดหรือเขียน เพื่อเป็นการแสดงความคิด 1 ความคิดอย่างสมบูรณ์ทางไวยากรณ์ โดยปกติมักประกอบไปด้วยประธานและภาคแสดง
7. ไวยากรณ์ (Grammar) คือ ภาษาที่มีกฎเกณฑ์ และเป็นไปตามระเบียบวิธีการประกอบรูปคำให้เป็นประโยค

การวิเคราะห์ในระดับที่สูงขึ้นกว่าระดับวากยสัมพันธ์หรือโครงสร้างของภาษา จำเป็นต้องอาศัยฐานความรู้ที่สูงขึ้น เช่น ฐานความรู้เครือข่ายคำ (WordNet) สำหรับกระบวนการวิเคราะห์ภาษารวมชาติในภาษาไทยส่วนใหญ่ เป็นการวิเคราะห์ภาษาในระดับวจิวิภาคและระดับวากยสัมพันธ์เท่านั้น เนื่องจากเครือข่ายคำไทย (Thai WordNet) ก่อนข้างมีจำนวนจำกัดและยังอยู่ในขั้นพัฒนา (Alisa et al., 2010)

การวิเคราะห์ภาษาในระดับจิตวภาคและวากยสัมพันธ์สำหรับภาษาไทย มีความยุ่งยากกว่าภาษาอังกฤษเล็กน้อย เนื่องจากภาษาไทยเป็นภาษาที่ไม่มีการแบ่งวรรคตอนระหว่างประโยคหรือระหว่างคำที่แน่นอน มักเขียนติดกันไปทั้งย่อหน้า ดังนั้นการวิเคราะห์ในระดับต่ำสุดหรือคำ จึงจำเป็นต้องแบ่งเอกสารตั้งแต่หน่วยใหญ่สุดคือประโยค ให้มีหน่วยเล็กสุดที่มีความหมายก่อน ซึ่งหมายถึงคำ แล้วจึงนำคำแต่ละคำมาประกอบเป็นโครงสร้างประโยค

ดังนั้นการวิเคราะห์ภาษาสำหรับภาษาไทยต้องอาศัย 3 กระบวนการพื้นฐานคือ (1) กระบวนการตัดคำ (Tokenization) เพื่อให้มีความสามารถในการวิเคราะห์ระดับจิตวภาคได้ (2) กระบวนการกำกับคำตามหน้าที่ของคำ (Part-of-Speech tagging) และ (3) กระบวนการวิเคราะห์โครงสร้างภาษา (Syntactic analysis) เพื่อให้สามารถนำคำแต่ละคำมาประกอบเป็นโครงสร้างประโยคได้

2.1.2.1 กระบวนการตัดคำ

ในงานวิจัยด้านการตัดคำได้แบ่งการตัดคำออกเป็น 3 เทคนิคหลัก ได้แก่ (1) การตัดคำโดยใช้กฎ (Rule based approach) เป็นวิธีการพิจารณาการตัดคำจากพยัญชนะ สระ วรรณยุกต์ ตัดสะกด การันต์ วิธีนี้เป็นวิธีที่ง่ายที่สุด และทำงานได้เร็วที่สุด แต่ไม่สามารถแก้ปัญหาความกำกวมของกฎได้ เช่น พยัญชนะบางตัวสามารถเป็นได้ทั้งพยัญชนะต้นและตัวสะกด (2) การตัดคำโดยใช้ฐานความรู้จากพจนานุกรม (Dictionary approach) ดังตัวอย่างวิธีการเลือกตัดคำจากคำที่พบในพจนานุกรมและมีความยาวมากที่สุด (Longest matching) หรือเลือกตัดคำจากความเหมือนมากที่สุด (Maximal matching) วิธีการตัดคำด้วยพจนานุกรมนี้มีความถูกต้องมากกว่าการตัดคำโดยใช้กฎ แต่ยังไม่สามารถแก้ปัญหาคำกำกวมทั้งหมดได้ และ (3) การตัดคำโดยใช้ค่าสถิติจากคลังเอกสาร (Corpus based approach) วิธีการนี้นำเอาค่าสถิติการเกิดคำและหน้าที่ของคำเข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อเลือกแบบที่มีโอกาสการเกิดมากที่สุด วิธีการนี้มีความถูกต้องมากกว่า 2 วิธีการแรก แต่มีข้อจำกัดคือต้องมีฐานข้อมูลที่มีการตัดคำไว้อย่างถูกต้อง ฐานข้อมูลต้องมีขนาดใหญ่มากพอ เพื่อให้ได้สถิติข้อมูลที่มีความน่าเชื่อถือ

สำหรับงานวิจัยนี้ได้เลือกใช้การตัดคำจาก Java API ชื่อ BreakIterator ซึ่งอาศัยเทคนิคการตัดคำแบบใช้ฐานความรู้จากพจนานุกรม แบบเลือกตัดคำจากคำที่พบในพจนานุกรมและมีความยาวมากที่สุด ซึ่งมีความเร็วในการประมวลผลและมีความถูกต้องในระดับหนึ่ง แต่ปัญหาหลักของการประมวลผลภาษาธรรมชาติคือ ความกำกวมของภาษา เนื่องจากในภาษาธรรมชาติมีกฎเกณฑ์และข้อยกเว้นมากมาย คำแต่ละคำอาจตีความได้แตกต่างกัน ถ้าอยู่ในบริบทที่แตกต่างกัน (กนกวรรณ เขียววรรณ, 2555) ผู้วิจัยจึงตรวจสอบและแก้ไขผลลัพธ์ที่ได้จากกระบวนการตัดคำให้มีความถูกต้องก่อนนำเข้าสู่กระบวนการวิเคราะห์

2.1.2.2 กระบวนการกำกับคำตามหน้าที่ของคำ

การกำกับคำตามหน้าที่ของคำและการวิเคราะห์ในระดับวากยสัมพันธ์จะช่วยให้เข้าใจประโยคได้ถูกต้องยิ่งขึ้น ซึ่งการกำกับคำตามหน้าที่ของคำเป็นวิธีการบ่งบอกว่าคำแต่ละคำในประโยคเป็นคำชนิดใด จึงทำให้การวิเคราะห์ไวยากรณ์โครงสร้างประโยคและการแปลความถูกต้องยิ่งขึ้น

หน้าที่หลักของคำ แบ่งได้ดังนี้ คำนาม คำสรรพนาม คำกริยา คำกริยาช่วย คำวิเศษณ์ คำบุพบท คำสันธาน คำนามชี้เฉพาะ คำนามบอกลักษณะ คำปฏิเสธ และคำหยุด สัญลักษณ์ที่ใช้ในการกำกับคำจะใช้คำย่อเป็นภาษาอังกฤษเพื่อเป็นสัญลักษณ์บอกชนิดของคำ ดังตารางที่ 2.1

ตารางที่ 2.1 สัญลักษณ์ที่ใช้ในการกำกับหน้าที่ของคำและความหมายของสัญลักษณ์

สัญลักษณ์	คำอธิบายหน้าที่ของคำ
N	คำนามใช้เรียกคน สัตว์ สิ่งของ
PRON	คำสรรพนามที่ใช้แทนคำนาม
V	คำกริยาแสดงอาการหรือการกระทำของนามและสรรพนาม
AUX	คำที่เติมหน้าคำกริยาหลักในประโยคเพื่อช่วยขยายความหมายของคำกริยาให้ได้ใจความชัดเจนยิ่งขึ้น
ADJ	คำคุณศัพท์ที่ใช้ขยายได้คำนามและคำสรรพนาม
ADV	คำกริยาวิเศษณ์ ใช้ขยายคำกริยาและคำวิเศษณ์เอง
PREP	คำบุพบททำหน้าที่เชื่อมคำหรือกลุ่มคำ
CONJ	คำที่ใช้เชื่อมประโยคกับประโยค
DET	คำนามชี้เฉพาะ
CLAS	คำนามบอกลักษณะ ขนาดหรือปริมาณ
NEG	คำปฏิเสธ
END	คำหยุด

2.1.2.3 การวิเคราะห์กฎไวยากรณ์โครงสร้างประโยค

กฎไวยากรณ์โครงสร้างประโยค (Syntax) หรือการแจงประโยค (Parsing) เป็นกระบวนการอธิบายโครงสร้างประโยคด้วยสูตรไวยากรณ์ (Grammar formalism) เพื่ออธิบาย

รูปแบบของคำที่ประกอบกันเป็นประโยค ซึ่งจะถูกรับด้วยโครงสร้างในลักษณะแผนภูมิต้นไม้ (Tree diagram) และ โครงสร้างในลักษณะวงเล็บ (Labeled bracketing) สูตรไวยากรณ์ที่นิยมใช้มากที่สุดตัวหนึ่งได้แก่ไวยากรณ์แบบไม่พึ่งบริบท (Context Free Grammar: CFG)

การอธิบายโครงสร้างภาษาด้วยสูตรไวยากรณ์แบบไม่พึ่งบริบทจัดเป็นส่วนหนึ่งของการอธิบายโครงสร้างภาษาแบบไวยากรณ์วลี (Phrase Structure grammars: PS) ซึ่งไม่พิจารณาถึงความหมายของคำ แต่จะอาศัยหน้าที่ของคำและอธิบายตามกลุ่มคำนามและคำกริยา ในการแบ่งประโยคออกเป็น ส่วน และพิจารณาลำดับของชนิดคำจากซ้ายไปขวาและแตกย่อยออกไปเรื่อย ๆ ด้วยเซตของกฎที่ใช้อธิบายความสัมพันธ์ของคำในประโยคประกอบด้วย 2 สัญลักษณ์ คือ (1) สัญลักษณ์ไม่จบท้ายหรือหมายถึงสัญลักษณ์ที่สามารถแตกต่อไปได้อีก (Non-terminal symbols) เช่น หน้าที่ของคำ (Part-of-Speech: POS) หรือกลุ่มคำ (Chunks) และ (2) สัญลักษณ์จบท้ายหรือไม่สามารถแตกต่อไปได้อีก (Terminal symbols) คือคำศัพท์ที่อยู่ในพจนานุกรม สูตรไวยากรณ์แบบไม่พึ่งบริบท ดังตัวอย่าง

$$S = NP + VP$$

$$NP = N \mid N + (ADJ) + (ADV) + (PP) \mid PRON$$

$$PP = PERP + NP \mid PERP + VP \mid PERP + NP + VP$$

$$VP = V \mid V + (ADV) \mid AUX + V \mid VP + NP$$

เทคนิคการอธิบายโครงสร้างภาษาหรือการแจงประโยค มี 2 เทคนิค ได้แก่

1) Top-down parsing

เริ่มต้นด้วยสัญลักษณ์ S แล้วเขียนใหม่ด้วยสัญลักษณ์ทางซ้ายมือ แจงประโยคจนกว่าจะพบสัญลักษณ์สิ้นสุด

2) Bottom-up parsing

เริ่มจากคำศัพท์หรือสัญลักษณ์สิ้นสุด แทนคำด้วยหน้าที่ของคำ จากนั้นใช้สัญลักษณ์ที่อยู่ทางซ้ายของกฎแทนด้วยกลุ่มของสัญลักษณ์ไม่จบท้ายทำไปจนกว่าจะพบสัญลักษณ์ S

งานวิจัยของ สมนึก สิริบุปวน (2546) ได้นำเสนอโครงสร้างวลีอย่างละเอียดสำหรับการวิเคราะห์โครงสร้างวลีภาษาไทย แต่เนื่องจากภาษาธรรมชาติเป็นภาษาที่ซับซ้อนและไม่เป็นไปตามกฎเสมอไป มีทั้งการรวมประโยคและประโยคเชิงซ้อน หรือละประธาน กรรม นักภาษาศาสตร์จึงแบ่งแนวทางการวิเคราะห์ภาษาออกเป็น 2 แนวทางหลัก คือ

1) การอธิบายโครงสร้างภาษาตามหลักไวยากรณ์ (Rules-based หรือ Prescriptive rules) เป็นการอธิบายและวิเคราะห์โครงสร้างภาษาตามหลักไวยากรณ์ทางภาษาศาสตร์ ซึ่งวิธีการนี้มีความยุ่งยากและเสียค่าใช้จ่ายสูง

2) การอธิบายโครงสร้างภาษาตามการใช้งานที่เกิดขึ้นจริง (Corpus-based หรือ Descriptive rules) ซึ่งการใช้ภาษาจะไม่เป็นไปตามหลักไวยากรณ์ทางภาษาศาสตร์เสมอไป การวิเคราะห์โครงสร้างภาษาที่อาศัยความรู้ที่ได้จากคลังข้อมูลการใช้ภาษาที่เกิดขึ้นจริง หรือเป็นการเพิ่มคุณสมบัติทางข้อมูลสถิติเข้าไปในส่วนของไวยากรณ์ ทำให้ไวยากรณ์มีลักษณะที่เป็นแบบสถิติ (Stochastic)

2.1.3 การจำแนกประเภทข้อความ

การจำแนกข้อความ (Text categorization หรือ Text classification: TC) มีเป้าหมายเพื่อสร้างแบบจำลองจากชุดข้อมูลเรียนรู้ที่รู้ผลเฉลยแล้ว (Train set) และนำแบบจำลองดังกล่าวไปจัดกลุ่มให้กับชุดข้อมูลทดสอบหรือชุดข้อมูลที่ยังไม่รู้ผลเฉลย (Test set) ให้อยู่ในกลุ่มที่กำหนดไว้ ซึ่งวิธีการจัดกลุ่มข้อความจะอาศัยการวิเคราะห์จากคำภายในข้อความเป็นหลัก โดยแบบจำลองการเลือกกลุ่มที่ดีที่สุดเกิดขึ้นจากการเรียนรู้จากชุดข้อมูลเรียนรู้ที่มีการจัดกลุ่มไว้แล้วโดยผู้เชี่ยวชาญ เรียกว่าการเรียนรู้แบบมีผลเฉลย วิธีการจำแนกข้อความแบ่งออกเป็น 2 วิธีหลักได้แก่

2.1.3.1 วิธีวิศวกรรมองค์ความรู้ (Knowledge engineering approach)

วิธีวิศวกรรมองค์ความรู้เป็นวิธีการสร้างกฎการจำแนกข้อความแบบ “ถ้า-แล้ว” ด้วยผู้เชี่ยวชาญ โดยวิธีการระบุคุณลักษณะสำคัญของข้อความและกำหนดกลุ่มที่เหมาะสมให้กับข้อความ ข้อดีคือ ได้กฎที่มีความถูกต้องแม่นยำสูง สามารถแก้ไขและจัดการง่ายหากมีจำนวนกฎน้อย ข้อเสียคืออาจเกิดความขัดแย้งของกฎหากมีผู้เชี่ยวชาญสร้างกฎมากกว่าหนึ่งคน และต้องสร้างกฎใหม่ทุกครั้งที่เปลี่ยน โดเมนของชุดข้อมูล ทำให้ไม่เหมาะกับการจำแนกข้อความที่มีปริมาณมาก

2.1.3.2 วิธีการเรียนรู้ของเครื่อง (Machine learning approach)

การจำแนกเอกสารหรือข้อความ เป็นเทคนิคหนึ่งในการเรียนรู้ของเครื่อง ซึ่งงานวิจัยนี้ได้ใช้วิธีการเรียนรู้จากชุดข้อมูลเรียนรู้ เพื่อนำมาสร้างเป็นแบบจำลองแบบอัตโนมัติในการจำแนกข้อความ โดยข้อมูลที่นำมาเรียนรู้นั้นจะต้องมีการกำหนดผลเฉลยไว้แล้ว เพื่อให้คอมพิวเตอร์สามารถเรียนรู้รูปแบบของข้อมูล และสร้างแบบจำลองเพื่อไว้ใช้ทำนายหรือจัดกลุ่มของชุดข้อมูลทดสอบได้

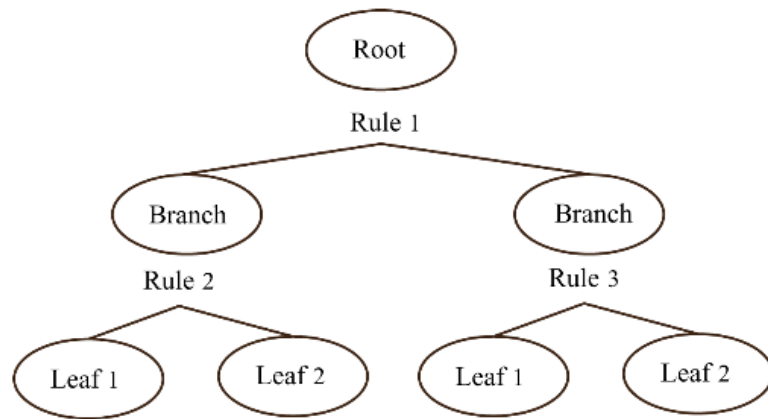
ยังมีปริมาณชุดข้อมูลเรียนรู้มาก จะยิ่งทำให้แบบจำลองการจำแนกข้อความที่มีความถูกต้องสูง แต่ระยะเวลาที่เครื่องใช้ในการสร้างแบบจำลองก็มากตามไปด้วย อัลกอริทึมสำหรับการจำแนกข้อความ เช่น ต้นไม้ตัดสินใจ นาอ็พเบย์ และซัพพอร์ตเวกเตอร์แมชชีน เป็นต้น

1) ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นอัลกอริทึมจำแนกข้อมูลที่มีลักษณะการตัดสินใจเลือกแบบโครงสร้างต้นไม้ ประกอบด้วยโหนดราก (Root node) โหนดกิ่ง (Branch node) และโหนดใบ ดังภาพที่ 2.3 โดยใช้คุณลักษณะของข้อมูล (Attributes) ในชุดข้อมูลเรียนรู้ มาสร้างโหนดการตัดสินใจแบบ “ถ้า-แล้ว” (IF-THEN) เช่น

“IF Income = High and Married = No THEN Risk = Poor”

“IF Income = High and Married = Yes THEN Risk = Good”



ภาพที่ 2.3 โครงสร้างต้นไม้ตัดสินใจ

ซึ่งเกณฑ์การเลือกเงื่อนไข (คุณลักษณะ) เพื่อสร้างเป็นโหนดของต้นไม้ นั้น นิยมใช้ค่าสารสนเทศ (Information Gain: IG) หรือค่าดัชนี Gini (Gini index) ดังสมการ

สมการ Entropy

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t) \quad (5)$$

สมการ Information Gain

$$Gain_{split} = Entropy(p) - \sum_{i=1}^j \frac{n_i}{n} Entropy(i) \quad (6)$$

สมการ Gini index

$$Gini(t) = 1 - \sum [p(j|t)]^2 \quad (7)$$

สมการ Gini split

$$Gini_{split} = 1 - \sum_{t=1}^k \frac{n_i}{n} Gini(i) \quad (8)$$

กระบวนการสร้างต้นไม้ตัดสินใจเริ่มจากคำนวณค่าเกณฑ์การเลือกจากทุกคุณลักษณะในชุดข้อมูล แล้วพิจารณาเลือกคุณลักษณะจากค่าเกนสารสนเทศสูง (หรือค่าเอ็นโทรที่ต่ำ) หรือค่าดัชนี Gini ต่ำก่อน (เพราะถือว่าคุณลักษณะดังกล่าวมีความสามารถในการจำแนกหมวดหมู่สูง) จากนั้นนำคุณลักษณะดังกล่าวมาสร้างเป็นโหนดราก ชุดข้อมูลที่ผ่านโหนดรากจะถูกแบ่งกลุ่มตามค่าในคุณลักษณะที่เป็นไปได้ (ค่าเงื่อนไขการตัดสินใจ) จากนั้นสร้างโหนดกิ่ง โดยคำนวณค่าเกณฑ์การเลือกจากคุณลักษณะที่เหลือ ทำซ้ำกระบวนการเดิมเพื่อสร้างโหนดกิ่งไปจนกว่าชุดข้อมูลที่ถูกแบ่งกลุ่มตามเงื่อนไข จะอยู่ในคลาสเดียวกันทั้งหมด หรือมีค่าเอ็นโทรที่เท่ากับศูนย์ หมายถึงไม่มีการเปลี่ยนแปลงของคำตอบ ทุกข้อมูลให้ค่าความจริงเดียวกัน ตัวอย่างอัลกอริทึมต้นไม้ตัดสินใจ เช่น CART, ID3, C4.5 และ CHAID

ข้อดีของอัลกอริทึมต้นไม้ตัดสินใจคือ ผู้ใช้สามารถทำความเข้าใจแบบจำลองการตัดสินใจของต้นไม้ได้ แต่ไม่รองรับข้อมูลแบบต่อเนื่อง (Continuous data) หากข้อมูลมีลักษณะดังกล่าว ต้องแบ่งข้อมูลให้เป็นแบบไม่ต่อเนื่องหรือแบบช่วงข้อมูล (Discrete data) ก่อน นอกจากนี้ประสิทธิภาพการจำแนกข้อความอาจดีกว่าวิธีอื่นด้วย จึงมักใช้เป็นพื้นฐานสำหรับเปรียบเทียบผลลัพธ์กับอัลกอริทึมอื่นเท่านั้น (Ronen et al., 2006)

2) นาอ็ฟเบย์

นาอ็ฟเบย์เป็นวิธีการเรียนรู้เพื่อสร้างแบบจำลองการจำแนกเอกสาร ที่มีพื้นฐานมาจากทฤษฎีความน่าจะเป็นของเบย์ (Bay theorem) แบบจำลองที่ได้จะอยู่ในรูปแบบของความน่าจะเป็น ซึ่งอาศัยค่าความน่าจะเป็นจากชุดข้อมูลเรียนรู้ หรือเรียกว่าความรู้ก่อนหน้า (Prior probability) มาทำนายผลของชุดข้อมูลทดสอบ ด้วยวิธีคำนวณค่าความน่าจะเป็นของชุดข้อมูลที่จะอยู่ในคลาส C จำนวนจนครบทุกคลาส ความน่าจะเป็นของคลาสใดที่มากที่สุดจะถูกเลือกเป็นคำตอบ

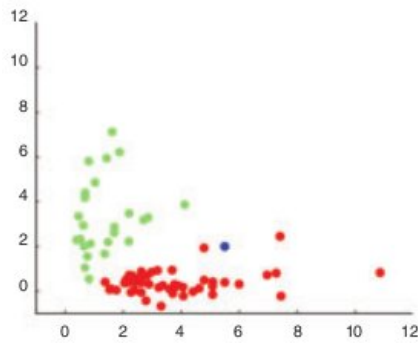
การจำแนกข้อความด้วยวิธีนาอ็ฟเบย์จะแทนข้อความด้วยเวกเตอร์ที่ประกอบด้วยค่า w_i ; $\vec{x}_i = (w_1, w_2, w_3, \dots, w_m)$ ดังนั้นการคำนวณหาความน่าจะเป็นของเอกสาร d ที่อยู่ในคลาส C เกิดจากผลรวมของความน่าจะเป็นของค่า w_i ที่พบในคลาสนั้น ดังสมการที่ 9

$$P(d|c) = \prod_{i=1}^{\text{length}(d)} P(w_i|c) \quad (9)$$

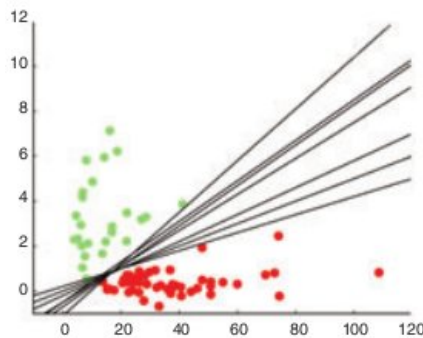
นาอ็พเบย์ถือเป็นวิธีที่มีประสิทธิภาพสำหรับการจำแนกข้อความวิธีหนึ่ง
และมีการคำนวณไม่ซับซ้อน (Ronen et al., 2006)

3) ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการแบ่งกลุ่มข้อมูลออกเป็น 2 กลุ่มด้วย
เส้นระนาบแบ่งข้อมูล (Hyperplane) จากชุดข้อมูลในภาพที่ 2.4 สังเกตว่าข้อมูลสามารถถูกแบ่งได้
ด้วยเส้นไฮเปอร์เพลนมากกว่า 2 เส้น ดังภาพที่ 2.5 ดังนั้นวิธีในการหาเส้นไฮเปอร์เพลนที่เหมาะสม
ที่สุดคือเส้นไฮเปอร์เพลนที่ทำให้ชุดข้อมูลทั้งสองกลุ่มมีระยะห่างระหว่างกันมากที่สุด (Maximum
Margin Hyperplane: MMH)



ภาพที่ 2.4 ตัวอย่างกลุ่มข้อมูล



ภาพที่ 2.5 เส้นไฮเปอร์เพลนแบ่งกลุ่มข้อมูล

กำหนดให้ชุดข้อมูลเรียนรู้ $D = \{ (\vec{x}_i, y_i) \}$ โดยที่ $\vec{x}_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{im})$ เป็นอินพุตเวกเตอร์ ตัวแทนของข้อความ และแต่ละ x_i ถูกกำหนดคลาสไว้ด้วยคลาส y_i เมื่อ y_i เป็นค่าจำนวนจริงตั้งแต่ -1 ถึง +1 ดังสมการที่ 10

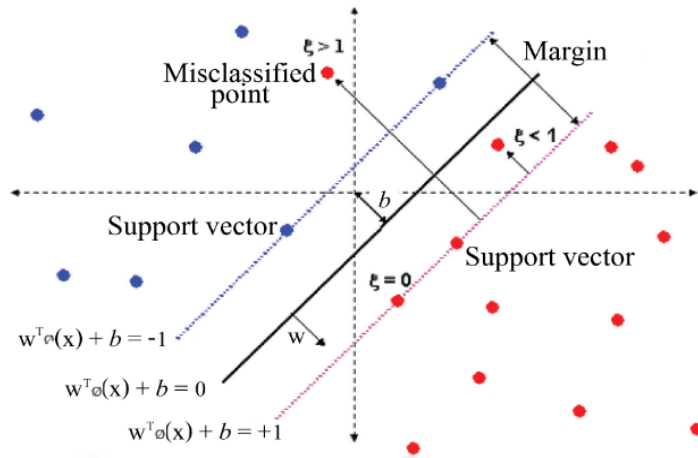
$$y = \begin{cases} +1, \vec{w} * \vec{x} + b > 0 \\ -1, \vec{w} * \vec{x} + b < 0 \end{cases} \quad (10)$$

เส้นไฮเปอร์เพลนคือเส้นที่ทำให้สมการ y_i มีค่าเท่ากับ 0 โดยมีเวกเตอร์ \vec{w} คือเวกเตอร์ที่ตั้งฉากกับเส้นไฮเปอร์เพลน, \vec{x}_i คือเวกเตอร์ข้อมูล และ b คือค่าโน้มเอียง (Bias)

เมื่อมีข้อมูลเข้ามาใหม่ที่ไม่รู้ผลเฉลย, \vec{x} เมื่อต้องการทำนายหาคلاس y ของข้อมูล \vec{x} จะทำการเปรียบเทียบ \vec{x} ว่าสอดคล้องหรือใกล้เคียงกับค่า \vec{x}_i ใด ๆ ที่มีการกำหนดคลาส y_i ไว้แล้ว กล่าวคือทำนายค่าคลาส y จากชุดข้อมูลเรียนรู้ (\vec{x}_i, y_i) ที่คล้ายกันมากที่สุด ทำให้ซัพพอร์ทเวกเตอร์แมชชีนสามารถทำนายข้อมูลลักษณะที่เป็นเส้นตรงได้ดี

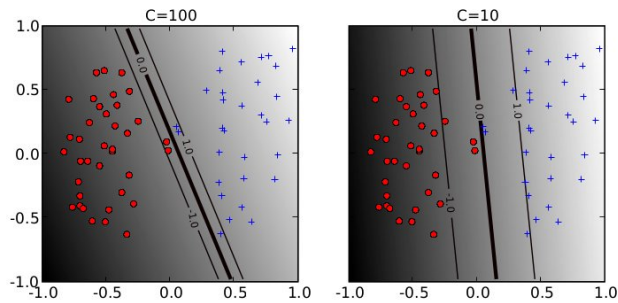
สมมติฐานสำหรับเส้นไฮเปอร์เพลนคือเส้นที่สามารถแบ่งกลุ่มข้อมูล 2 กลุ่มออกจากกันได้ทั้งหมด (Zero error) โดยสามารถเพิ่มประสิทธิภาพการจำแนกเอกสารได้ โดยวิธีเพิ่มระยะห่างระหว่างกลุ่มให้มากที่สุด ด้วยการปรับค่า \vec{w} และค่า b ให้เหมาะสม แต่ในความเป็นจริงเส้นไฮเปอร์เพลนไม่สามารถแบ่งกลุ่มข้อมูลได้ทั้งหมดเนื่องจากมีข้อมูลบางอย่างที่แตกต่างออกไปจากกลุ่ม (Misclassified point) ดังภาพที่ 2.6 การปรับเส้นไฮเปอร์เพลนเพื่อให้เกิด Zero error นั้นไม่เหมาะสมนัก จึงมีการกำหนดค่า Soft margin ให้สามารถยอมรับความผิดพลาดของการจัดกลุ่มได้ ด้วยค่าพารามิเตอร์ C (Cost parameter) กล่าวคือค่าพารามิเตอร์ C เป็นตัวแปรที่ใช้สำหรับการพิจารณาถึงความเหมาะสมระหว่างการยอมรับความผิดพลาดที่เกิดขึ้นกับเส้นแบ่งกลุ่มข้อมูลที่มีระยะห่างระหว่างกลุ่มมากที่สุด พารามิเตอร์ C ที่มีค่าสูง หมายถึงกำหนดให้ยอมรับความผิดพลาดน้อย บางครั้งการกำหนดค่าพารามิเตอร์ C ที่สูงเกินไป จะทำให้เกิดปัญหา Overfitting ได้ ดังภาพที่ 2.7 ด้านซ้าย การกำหนดค่าพารามิเตอร์ C ต่ำ ๆ ก็คือค่ายอมรับความผิดพลาดที่เกิดขึ้นได้ ดังภาพที่ 2.7 ด้านขวา

สำหรับชุดข้อมูลที่ไม่สามารถแบ่งกลุ่มข้อมูลได้ด้วยเส้นตรง ซัพพอร์ทเวกเตอร์แมชชีนมีฟังก์ชันเคอร์เนล ϕ (Kernel function) ในการแปลงข้อมูลนำเข้า (Input space) ให้เป็นฟีเจอร์สเปซ (Feature space) (R. Feldman and J. Sanger, 2006) ดังสมการที่ 11 และ 12 เพื่อให้สามารถแบ่งกลุ่มข้อมูลที่ไม่สามารถแบ่งได้ด้วยเส้นตรงได้



ภาพที่ 2.6 ซัพพอร์ตเวกเตอร์แมชชีน

แหล่งที่มา: The Standard SVM Formulation, 2012.



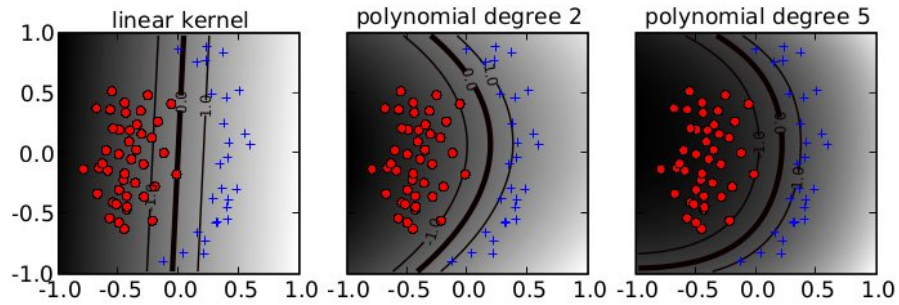
ภาพที่ 2.7 การเส้นไฮเปอร์เพลนเมื่อมีการปรับค่าพารามิเตอร์ C

$$(x, x') \Rightarrow k(x, x') \tag{11}$$

$$k(x, x') = (x \cdot x') = \phi(x) \cdot \phi(x') \tag{12}$$

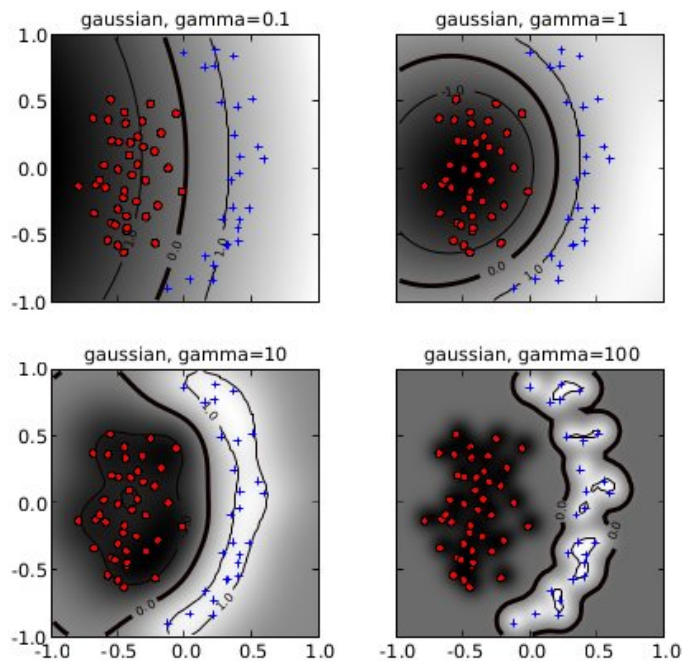
สามารถเลือกใช้ฟังก์ชันเคอร์เนลในการแปลงข้อมูลได้อย่างอิสระเพื่อให้รูปแบบข้อมูลหรือฟีเจอร์สเปซที่เหมาะสมสำหรับการวิเคราะห์ เช่น

ฟังก์ชันเคอร์เนล Polynomial ที่มีพารามิเตอร์คือค่าดีกรี (Degree) สำหรับการปรับค่าความโค้งของเส้นไฮเปอร์เพลนให้เหมาะสมกับข้อมูล ดังภาพที่ 2.8



ภาพที่ 2.8 การแบ่งกลุ่มข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีนด้วยฟังก์ชันเคอร์เนล Polynomial

ฟังก์ชันเคอร์เนล Radial basic (RBF) คือการปรับค่าความกว้างด้วยพารามิเตอร์แกมมา (γ) ให้เส้นไฮเปอร์เพลนเหมาะสมกับข้อมูล ดังภาพที่ 2.9



ภาพที่ 2.9 การแบ่งกลุ่มข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนล Radial basic

ข้อดีของซัพพอร์ตเวกเตอร์แมชชีนคือสามารถแบ่งกลุ่มข้อมูลได้ทั้งรูปแบบข้อมูลที่แบ่งกลุ่มได้ด้วยเส้นตรงและไม่เป็นเส้นตรง นอกจากนี้ยังสามารถรองรับคุณลักษณะจำนวนมากได้ (มากกว่า 10,000 คุณลักษณะ) เนื่องจากการแทนข้อมูลแบบเวกเตอร์

และพิจารณาเส้นแบ่งกลุ่มข้อมูลจากเวกเตอร์ซัพพอร์ต (Support vector) แต่ข้อเสียคือต้องทดลองเพื่อปรับค่าพารามิเตอร์ให้เหมาะสมสำหรับแต่ละเคอร์เนลที่เลือกใช้

ในงานวิจัย (Thorsten, 1998) ได้ทดสอบเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการจำแนกข้อความด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์, Rocchio, C4.5 และเพื่อนบ้านที่ใกล้ที่สุด (k-nearest neighbor: k-NN) ซึ่งพบว่าซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพการจำแนกข้อความดีกว่าวิธีอื่น เนื่องด้วยคุณสมบัติของข้อความ ดังนี้ (Thorsten, 1998; Ian and Eibe, 2000)

- มิติของข้อมูลจำนวนมาก (High dimensional) บางข้อความอาจมีคุณลักษณะมากถึง 10,000 คุณลักษณะ แต่ซัพพอร์ตเวกเตอร์แมชชีนมีการแบ่งกลุ่มข้อมูลด้วยเส้นไฮเปอร์เพลนที่ไม่ขึ้นอยู่กับมิติของข้อมูล แต่ขึ้นอยู่กับเวกเตอร์ซัพพอร์ต (Support vector) หรือข้อมูลที่อยู่บนเส้นไฮเปอร์เพลนเท่านั้น ซึ่งทำให้วิเคราะห์ข้อมูลขนาดใหญ่ได้ (แก้ปัญหา Infeasible computational complexity)

- มีคุณลักษณะที่ไม่เกี่ยวข้องน้อย (Few irrelevant feature) การลดมิติของข้อมูลเพื่อประสิทธิภาพการวิเคราะห์ บางครั้งอาจทำให้สูญเสียคุณลักษณะสำคัญของข้อมูลบางส่วนไป

- เวกเตอร์ของข้อความมีลักษณะกระจาย คือมีเวกเตอร์ที่มีค่าเป็น 0 จำนวนมาก การคำนวณค่าโน้มเอียงในซัพพอร์ตเวกเตอร์แมชชีนจะเหมาะกับการแก้ปัญหาดังกล่าว

- ส่วนใหญ่การจำแนกข้อความมักแยกได้ด้วยเส้นตรง

ประสิทธิภาพการจำแนกข้อความด้วยซัพพอร์ตเวกเตอร์แมชชีน สอดคล้องกับผลการทดลองในงานวิจัยของ Yiming Yang and Xin Liu (1999) และ Basu, Watters and Shepherd (2002) แต่สำหรับการจำแนกเอกสารความคิดเห็นในภาษาไทยในงานวิจัยของ Khampol Sukhum, Supot Nitsuwat and Choochart Haruechaiyasak (2011) ที่ได้นำเสนอการจำแนกเอกสารความคิดเห็นออกจากข้อเท็จจริงในบทความข่าว โดยเปรียบเทียบความถูกต้องของการจำแนกด้วยอัลกอริทึม นาอ์ฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีนและเพื่อนบ้านใกล้เคียง พบว่า นาอ์ฟเบย์ได้ผลดีที่สุด ซึ่งได้ผลลัพธ์ที่แตกต่างจากงานวิจัยที่ได้กล่าวมาข้างต้น (Thorsten, 1998; Yiming Yang et al., 1999; Basu et al., 2002) แต่ได้นำเสนอเพิ่มเติมว่าสามารถเพิ่มประสิทธิภาพการจำแนกข้อความได้ด้วยวิธีการเพิ่มคุณลักษณะให้กับข้อความ เช่น หน้าทีของคำ วลี หรืออื่น ๆ ที่

นอกเหนือจากการวิเคราะห์โดยใช้ “คำ” เพียงอย่างเดียว สำหรับกระบวนการวิเคราะห์ Association Rule Discovery จึงสรุปไว้ว่า ไม่มีวิธีใดที่ทำงานได้ดีที่สุดสำหรับข้อมูลทุกประเภท ดังนั้นงานวิจัยนี้จึงทดลองเพื่อหาอัลกอริทึมที่สามารถจำแนกข้อเสนอแนะได้ดีที่สุด โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อความด้วยอัลกอริทึมต้นไม้ตัดสินใจ นาอ์ฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน นอกจากนี้ยังได้นำเสนอวิธีการแทนข้อความด้วยฐานความรู้ภาษาเพื่อเพิ่มประสิทธิภาพการจำแนกข้อเสนอแนะให้ดียิ่งขึ้น

2.1.4 การประเมินประสิทธิภาพการจำแนกประเภทข้อความ

การประเมินประสิทธิภาพของกระบวนการจำแนกข้อความ (หรือเอกสาร) ได้อาศัยการประเมินประสิทธิภาพจากการค้นคืนสารสนเทศ (Information Retrieval: IR) ซึ่งระบบการค้นคืนสารสนเทศที่ดีควรดึงเอกสารที่เกี่ยวข้องออกมาให้ได้มากที่สุด และขจัดเอกสารที่ไม่เกี่ยวข้องออกไปให้ได้มากที่สุด สามารถวัดประสิทธิภาพของกระบวนการด้วยค่าระลึกละและค่าความแม่นยำ ได้ดังตารางที่ 2.2 โดยที่ค่าระลึกละ คือความสามารถของระบบที่จะดึงเอกสารที่เกี่ยวข้องออกมา เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด ดังสมการที่ 13 ค่าระลึกละเท่ากับ 1 หมายถึงระบบสามารถค้นคืนเอกสารที่เกี่ยวข้องทั้งหมดได้ แต่ถ้าค่าผลลัพธ์ที่ได้เป็น 0 หมายถึงระบบค้นคืนเอกสาร ได้ผลผิดพลาดทั้งหมด

ตารางที่ 2.2 เปรียบเทียบประสิทธิภาพการค้นคืนข้อมูล

Category C_i		Expert judgment	
		Yes	No
classifier judgment	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$Recall = \frac{True\ positive}{True\ positive + False\ Negative} \quad (13)$$

โดยที่ ค่าความถูกต้องเชิงบวก (True Positive) หมายถึงเอกสารที่อยู่กลุ่ม C_i และแบบจำลองทำนายว่าอยู่กลุ่ม C_i

ค่าความผิดพลาดเชิงบวก (False Positive) หมายถึงเอกสารไม่อยู่กลุ่ม C_i แต่แบบจำลองทำนายว่าอยู่กลุ่ม C_i

ค่าความผิดพลาดเชิงลบ (False Negative) หมายถึงเอกสารที่อยู่กลุ่ม C_i แต่แบบจำลองทำนายว่าไม่อยู่กลุ่ม C_i

ค่าความถูกต้องเชิงลบ (True Negative) หมายถึงเอกสารไม่อยู่กลุ่ม C_i และแบบจำลองทำนายว่าไม่อยู่กลุ่ม C_i

ในขณะที่ค่าความแม่นยำ (Precision) คือค่าความสามารถในการจัดเอกสารที่ไม่เกี่ยวข้องออกไป เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารทั้งหมดที่ทำการค้นคืนมาได้ ดังสมการที่ 14 ค่าความแม่นยำเท่ากับ 1 หมายถึงระบบสามารถค้นคืนเอกสารได้ถูกต้องโดยที่ไม่มีเอกสารที่ไม่เกี่ยวข้องปะปนอยู่

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (14)$$

โดยปกติแล้วค่าระลอกและค่าความแม่นยำมักนำมาพิจารณาร่วมกัน ระบบที่มีประสิทธิภาพดีหมายถึงค่าระลอกสูงด้วยค่าความแม่นยำที่สูงใกล้เคียงกัน สามารถวัดประสิทธิภาพได้ด้วยค่า F-measure ดังสมการที่ 15

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (15)$$

จากสมการประเมินประสิทธิภาพข้างต้นเป็นการประเมินประสิทธิภาพสำหรับชุดข้อมูลที่มี 2 กลุ่ม (Classes) สำหรับชุดข้อมูลที่มีมากกว่า 2 กลุ่ม สามารถประเมินประสิทธิภาพได้ด้วยวิธีการคำนวณค่า Micro averaging และค่า Macro averaging

ค่า Micro averaging เป็นการวัดประสิทธิภาพด้วยค่าเฉลี่ยแบบให้น้ำหนักทุกเอกสารเท่ากัน (Document-pivoted measurement) โดยจะคำนวณค่าความถูกต้องจากจำนวนเอกสาร ซึ่งจะนำจำนวนเอกสารของแต่ละกลุ่มมารวมกันเพื่อคำนวณหาค่าความถูกต้องในระดับเอกสาร

ค่า Macro averaging เป็นการวัดประสิทธิภาพด้วยค่าเฉลี่ยแบบให้น้ำหนักทุกกลุ่ม (Classes) เท่ากัน (Category-pivoted measurement) โดยจะคำนวณหาค่าความถูกต้องของแต่ละกลุ่มก่อน จากนั้นนำมาเฉลี่ยรวมกันเพื่อให้ได้ค่าความถูกต้องของระบบในระดับกลุ่ม ดังสมการ 16-19

$$Recall_{Micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \quad (16)$$

$$Precision_{Micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i} \quad (17)$$

$$Recall_{Macro} = \frac{\sum_{i=1}^{|C|} Recall_i}{m} \quad (18)$$

$$Precision_{Macro} = \frac{\sum_{i=1}^{|C|} Precision_i}{m} \quad (19)$$

2.2 ทบทวนวรรณกรรม

ปัจจุบันการวิเคราะห์ห้วงพิจารณาเน้นที่การวิเคราะห์ข้อความแสดงความคิดเห็นทางตรงเท่านั้น โดยวิเคราะห์ข้อความนำเข้าที่มีค่าแสดงขั้วความเห็น (Polar words) ที่ชัดเจน แล้วจึงจัดกลุ่มความคิดเห็นเป็นเชิงบวก (Positive) ลบ (Negative) หรือเป็นกลาง (Neutral) ในยุคแรกของการวิเคราะห์ความคิดเห็น Peter (2002) ได้อาศัยฐานความรู้ทางภาษา (Linguistic knowledge based) โดยนำเสนอการวิเคราะห์ขั้วความคิดเห็นจากคำวิเศษณ์หรือวลีวิเศษณ์ ซึ่งถือเป็นคำบ่งชี้ขั้วความเห็นที่ดี และจัดกลุ่มความคิดเห็นเป็นเชิงบวก (Thumbs up) หรือเชิงลบ (Thumbs down) ส่วนงานวิจัยในปัจจุบันได้นำเสนอวิธีวิเคราะห์ภาษาโดยอาศัยรูปแบบโครงสร้างภาษาที่เกิดขึ้นจริง (Syntactic pattern extraction) แต่ยังคงอาศัยการวิเคราะห์จากฐานความรู้ทางภาษาร่วมด้วย (Minqing Hu et al., 2004b; Alisa Kongthon et al., 2010; วรรณญา วรรณศรี, 2553) เนื่องจากคำถือเป็นหัวใจหลักในการวิเคราะห์ข้อความ

Bing Liu (2011) ได้แบ่งระดับการวิเคราะห์ความคิดเห็นออกเป็น 3 ระดับ ได้แก่

1. ระดับเอกสาร (Document level) เป็นการวิเคราะห์โดยสรุปขั้วความคิดเห็นจากทั้งเอกสาร (Peter, 2002)

2. ระดับประโยค (Sentence level) เนื่องจากในหนึ่งเอกสารประกอบไปด้วยหลาย ๆ ประโยคซึ่งแสดงความคิดเห็นต่อคุณลักษณะที่แตกต่างกัน การวิเคราะห์ในระดับเอกสารจึงไม่มี

ความละเอียดมากพอ Soo-Min Kim and Eduard Hovy (2004) และ Minqing Hu and Bing Liu (2004a) จึงนำเสนอวิธีการวิเคราะห์ความคิดเห็นในระดับประโยค

3. ระดับคุณลักษณะ (Entity and Feature/Aspect Level) เป็นการวิเคราะห์ข้อความความคิดเห็นในระดับที่ละเอียดมากขึ้น ด้วยวิธีการระบุว่าผู้เขียนชอบหรือไม่ชอบคุณลักษณะใดต่อหัวข้อที่สนใจ (Minqing Hu et al., 2004b, Bo Pang et al., 2008, วรรณญา วรรณศรี, 2552 และ Alisa Kongthon et. al, 2010) ซึ่ง Minqing Hu et al. (2004b) ได้นำเสนอวิธีการสร้างฐานความรู้ทางภาษาของคำแสดงคุณลักษณะ (Features) แบบอัตโนมัติ ด้วยเทคนิคการหาความสัมพันธ์ของข้อมูล (Association rule mining) ซึ่งนำเสนอว่าคำนามที่เกิดขึ้นร่วมกันบ่อยมีโอกาสที่คู่ของคำนามนั้นจะเป็นคุณลักษณะของหัวข้อที่สนใจ โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 1%

สำหรับการวิเคราะห์บทวิจารณ์ประเภทข้อเสนอแนะ ซึ่งปะปนอยู่กับบทวิจารณ์ประเภทอื่น (ข้อเท็จจริงและความคิดเห็น) งานวิจัยในปัจจุบันแบ่งออกเป็น 2 หัวเรื่องคือการสกัดข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น และการสกัดวลีประเภทข้อเสนอแนะ

การสกัดข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น ในงานวิจัย Vishwanath et al. (2011) ได้นำเสนอกระบวนการสกัดบทวิจารณ์ประเภทข้อเสนอแนะออกจากความคิดเห็น โดยใช้วิธีวิศวกรรมองค์ความรู้ด้วยเทคนิคการใช้ผู้เชี่ยวชาญสร้างกฎการตัดสินใจการสกัดแยกเอกสารหรือข้อความ โดยได้กำหนดกฎการตัดสินใจแยกข้อความที่เป็นข้อเสนอแนะด้วยหน้าที่ของคำที่เป็นคำกริยาช่วย (Modal verbs) เช่น “*should have*”, “*could have been*”, “*could be*” หรือ “*must be*” เป็นต้น วิธีการจำแนกข้อความด้วยเทคนิคการใช้ผู้เชี่ยวชาญในการสร้างกฎเป็นวิธีที่มีความถูกต้องสูง แต่ต้องใช้ระยะเวลาในการสร้างกฎนาน และยังมีข้อมูลจำนวนมากระยะเวลาในการสร้างกฎจะยิ่งนานมากขึ้นตามไปด้วย หากเมื่อใดเมนข้อมูลเปลี่ยนจำเป็นต้องสร้างกฎการตัดสินใจใหม่ทุกครั้ง ดังนั้นในงานวิจัยนี้ได้นำเสนอวิธีการจำแนกข้อเสนอแนะแบบอัตโนมัติ ด้วยวิธีการเรียนรู้ของเครื่อง โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อเสนอแนะของอัลกอริทึมต้นไม้ตัดสินใจ นาอ็ฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน เพื่อหาว่าอัลกอริทึมใดที่เหมาะสมกับการจำแนกข้อเสนอแนะมากที่สุด

กระบวนการสกัดหัวข้อข้อเสนอแนะ ในงานวิจัย Amar Viswanathan et al. (2011) ได้นำเสนอกระบวนการสกัดหัวข้อข้อเสนอแนะ ซึ่งอาศัย 2 แนวทางในการวิเคราะห์ก็คือ (1) การวิเคราะห์โดยอาศัยฐานความรู้ทางภาษา (Linguistic knowledge based) เทคนิคที่นำมาใช้ได้แก่ การประมวลผลภาษาธรรมชาติ ซึ่งหมายถึงการวิเคราะห์ในระดับคำ (Morphological analysis) โครงสร้างทางภาษา (Syntactic analysis) และความหมาย (Semantic analysis) งานวิจัยของ Viswanathan นี้ได้สร้างฐานความรู้ทางภาษาเก็บคำศัพท์ที่เป็นคำแสดง คำที่มีความหมายเชิงเปรียบเทียบ จำนวน และอาศัยเทคนิคการสร้างฐานความรู้แบบออนโทโลยี (Ontology) เพื่อให้

สามารถวิเคราะห์ภาษาธรรมชาติในระดับความหมายได้ (2) การวิเคราะห์ภาษาโดยอาศัยรูปแบบโครงสร้างภาษาที่เกิดขึ้นจริง (Syntactic pattern extraction) ด้วยวิธีการกำหนดกฎการใช้ภาษาของข้อเสนอแนะ (Rules) ดังตัวอย่าง

1. Patterns with explicit keywords มีคำบ่งชี้ข้อเสนอแนะที่ชัดเจน เช่น ‘suggest’, ‘recommend’ หรือ ‘I wish’ เป็นต้น

2. Patterns containing queries รูปแบบประโยคคำถาม

3. Patterns containing modal verbs ประโยคที่มีกริยาช่วย เช่น can, could, shall, may, might, must

การวิเคราะห์ข้อเสนอแนะจะทำการวิเคราะห์โครงสร้างทางภาษาก่อน จากนั้นเปรียบเทียบประโยคข้อเสนอแนะกับกฎการใช้ภาษาของข้อเสนอแนะ (Rule lookup) แล้วจึงสรุปวลีข้อเสนอแนะในแบบฟอร์มที่กำหนดไว้ในลักษณะตาราง (Frame Manager)

สำหรับงานวิจัยนี้ได้ศึกษาแนวทางจากการวิเคราะห์ข้อเสนอแนะจาก Viswanathan et al. (2011) และได้นำเสนอประเภทของข้อเสนอแนะที่จำแนกตามเจตนาการแสดงข้อเสนอแนะและรูปแบบการใช้ภาษาที่เกิดขึ้นบ่อยในประโยคข้อเสนอแนะสำหรับภาษาไทย ได้แก่

1. ข้อเสนอแนะทางตรง (S_d)

2. ข้อเสนอแนะเชิงคำถาม (S_q)

3. ข้อเสนอแนะเชิงเงื่อนไข (S_c)

ซึ่งจากการศึกษาพบว่าการจำแนกข้อเสนอแนะออกเป็นประเภท ก่อนนำไปสกัดหาวลีจะช่วยให้กระบวนการสกัดหาวลีข้อเสนอแนะมีความถูกต้องมากยิ่งขึ้น

บทที่ 3

กรอบการดำเนินงานวิจัย

3.1 การนิยามปัญหา

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาข้อเสนอแนะ และนำเสนอกระบวนการวิเคราะห์เหมืองข้อเสนอแนะ ซึ่งจากการศึกษาข้อเสนอแนะของผู้บริโภคบนอินเทอร์เน็ตพบว่าข้อเสนอแนะเป็นส่วนหนึ่งของบทวิจารณ์ผู้บริโภค (Customer reviews) ที่ประกอบด้วย

1. ข้อเท็จจริง หมายถึงประโยคแสดงสิ่งที่เกิดขึ้นจริง สามารถพิสูจน์ได้
2. ความคิดเห็น หมายถึงความรู้สึกต่อสิ่งใดสิ่งหนึ่งที่อาจมีข้อแสดงความคิดเห็นที่ชัดเจนคือข้อบวกรหรือลบ
3. ข้อเสนอแนะ หมายถึงการแสดงความคิดเห็นต่อหัวข้อใดหัวข้อหนึ่ง ในกรณีที่ไม่เห็นด้วยกับความคิดเห็นหรือการกระทำของผู้อื่นและนำเสนอความคิดเห็นใหม่ของตนเองเพื่อเป็นแนวทางให้ปฏิบัติ หรือการชี้ให้เห็นข้อบกพร่องพร้อมทั้งเสนอแนวทางแก้ไข

รูปแบบของประโยคข้อเสนอแนะจึงประกอบด้วย 3 ส่วนประกอบ คือ (1) คำระบุหัวข้อ (2) คำบ่งชี้ข้อเสนอแนะ และ (3) คำหรือวลีข้อเสนอแนะ ดังตัวอย่าง

ตัวอย่างที่ 1 : “ชอบรายการ กินอยู่คือ มาก อยากให้ เพิ่มเวลาเป็น 1 ชม.ค่ะ”

ตัวอย่างที่ 2 : “ถ้าเป็นไปได้อยากให้ เปลี่ยนเวลา รีรัน ASEAN FOCUS เป็นตอนกลางวัน ๆ ได้ไหมค่ะ”

ตัวอย่างที่ 3 : “รายการดี ๆ แบบ หนังพาไป ทำไม ถอดออกล่ะ”

ซึ่งงานวิจัยนี้ได้นำเสนอกระบวนการวิเคราะห์ข้อเสนอแนะโดยแบ่งเป็น 3 กระบวนการหลักคือ (1) กระบวนการสกัดข้อเสนอแนะ (Suggestion extraction) (2) กระบวนการจำแนกประเภทของข้อเสนอแนะ (Suggestion type classification) และ (3) กระบวนการสกัดวลีข้อเสนอแนะ สามารถนิยามปัญหาของแต่ละกระบวนการได้ดังนี้

3.1.1 การนิยามปัญหาของกระบวนการสกัดข้อเสนอแนะ

กระบวนการสกัดข้อเสนอแนะมีวัตถุประสงค์เพื่อสกัดแยกข้อเสนอแนะที่ปะปนอยู่กับบทวิจารณ์ประเภทอื่น (ข้อเท็จจริงและความคิดเห็น) สามารถนิยามปัญหา ได้ดังนี้

$$f_1 = d \rightarrow T \quad (20)$$

เมื่อ f_1 หมายถึงฟังก์ชันการจำแนกประโยคข้อเสนอแนะ (Suggestion) กับประโยคที่ไม่เป็นข้อเสนอแนะ (Non-suggestion) โดยจะวิเคราะห์ข้อมูลนำเข้าคือเซตของ d ซึ่งประกอบประโยค s_i ดังนั้น $d = (s_1, s_2, s_3, \dots, s_n)$ เมื่อ n คือจำนวนประโยคทั้งหมดที่นำมาวิเคราะห์ ซึ่ง s_i คือเวกเตอร์ที่ประกอบด้วยคำจำนวน m คำ และ T คือผลลัพธ์ของกระบวนการ โดยที่ T เป็นเซตที่ประกอบด้วย S หมายถึงประโยคข้อเสนอแนะ และ S' หมายถึงบทวิจารณ์ประเภทอื่นที่ไม่ใช่ข้อเสนอแนะ ส่วนประโยคที่ถูกจำแนกว่าเป็นข้อเสนอแนะ จะถูกนำไปวิเคราะห์ต่อในกระบวนการจำแนกประเภทข้อเสนอแนะ

3.1.2 การนิยามปัญหาของกระบวนการจำแนกประเภทข้อเสนอแนะ

กระบวนการจำแนกประเภทข้อเสนอแนะมีวัตถุประสงค์เพื่อแยกประโยคข้อเสนอแนะออกเป็น 3 ประเภท ได้แก่ ข้อเสนอแนะทางตรง ข้อเสนอแนะเชิงคำถาม และข้อเสนอแนะเชิงเงื่อนไข

การจำแนกข้อเสนอแนะออกตามประเภทจะช่วยให้กระบวนการสกัดหัวข้อข้อเสนอแนะที่ซ่อนอยู่ในประโยคมีความถูกต้องสูงกว่าการสกัดหัวข้อโดยตรงแบบไม่จำแนกประเภทออกมา ก่อน สามารถนิยามปัญหาของกระบวนการจำแนกประเภทข้อเสนอแนะ ได้ดังนี้

$$f_2 = X \rightarrow C \quad (21)$$

เมื่อ f_2 หมายถึงฟังก์ชันการจำแนกประเภทข้อเสนอแนะ ซึ่งมีข้อมูลนำเข้าคือ X โดยที่ X คือเซตประโยคข้อเสนอแนะที่ถูกสกัดได้จากฟังก์ชัน f_1 โดยจะวิเคราะห์เพื่อให้ได้ผลลัพธ์คือประโยคที่อยู่ในเซตของ C ที่ประกอบด้วยสมาชิกของประเภทข้อเสนอแนะ, $C = \{S_d, S_q, S_c\}$ ความหมายของประเภทข้อเสนอแนะดังตารางที่ 3.1 และผลลัพธ์ของฟังก์ชัน f_2 เป็นดังนี้

Y_d คือเซตของประโยคข้อเสนอแนะ ประเภทข้อเสนอแนะทางตรง

Y_q คือเซตของประโยคข้อเสนอแนะ ประเภทข้อเสนอแนะเชิงคำถาม

Y_c คือเซตของประโยคข้อเสนอแนะ ประเภทข้อเสนอแนะเชิงเงื่อนไข

ตารางที่ 3.1 ประเภทและความหมายของข้อเสนอแนะ

สัญลักษณ์	ความหมาย	คำบ่งชี้ข้อเสนอแนะ
S_c	ข้อเสนอแนะทางตรง (Explicit suggestion) หมายถึง ข้อเสนอแนะที่มีคำบ่งชี้ข้อเสนอแนะที่ชัดเจน ในการ แสดงความคิดเห็นในกรณีที่ไม่เห็นด้วยกับความ คิดเห็นหรือการกระทำของผู้อื่น และนำเสนอความ คิดเห็นใหม่ของตนเองเพื่อเป็นแนวทางให้ปฏิบัติ รูปแบบประโยคประกอบด้วยคำบ่งชี้ข้อเสนอแนะที่ ชัดเจน และมีเจตนาให้ปฏิบัติตามข้อเสนอแนะ ดังกล่าว	อยาก, ขอเสนอแนะ, ดี , น่าจะ, รบกวน, ควร, พิจารณา
S_q	ข้อเสนอแนะเชิงคำถาม หมายถึงข้อเสนอแนะที่มี รูปแบบประโยคที่ประกอบด้วยคำบ่งชี้ข้อเสนอแนะ ในเชิงคำถามอยู่ในประโยค และมีเจตนาการตั้ง คำถามถึงสิ่งที่ต้องการให้ปฏิบัติ	ทำไม, ได้ไหม
S_c	ข้อเสนอแนะเชิงเงื่อนไข หมายถึงข้อเสนอแนะที่มี รูปแบบประโยคที่ประกอบด้วยคำบ่งชี้ข้อเสนอแนะ ในเชิงเงื่อนไขอยู่ในประโยค และมีเจตนาในการ แสดงข้อเสนอแนะแบบมีเงื่อนไขของการปฏิบัติ	ถ้า, หาก

3.1.3 การนิยามปัญหาของกระบวนการสกัดวลีข้อเสนอแนะ

กระบวนการสกัดวลีข้อเสนอแนะมีวัตถุประสงค์เพื่อสกัดหาส่วนประกอบต่าง ๆ ของ ประโยคข้อเสนอแนะออกมา สามารถนิยามปัญหาของกระบวนการสกัดวลีข้อเสนอแนะ ได้ดังนี้

$$f_3 = Y \rightarrow (OBJ, SW, Suggestion) \quad (22)$$

โดยที่อินพุตของกระบวนการสกัดวลีข้อเสนอแนะนี้คือ ผลลัพธ์ที่ได้จากกระบวนการ จำแนกประเภทข้อเสนอแนะ, Y นำไปพิจารณาเพื่อสกัดหาส่วนประกอบของวลีข้อเสนอแนะ โดย

พิจารณาจากรูปแบบของประโยคข้อเสนอแนะที่เกิดขึ้นบ่อยในคลังความรู้ ซึ่งแบ่งออกตามประเภทของข้อเสนอแนะ 3 ประเภทคือ ข้อเสนอแนะทางตรง ข้อเสนอแนะเชิงคำถาม และข้อเสนอแนะเชิงเงื่อนไข และแต่ละรูปแบบของประโยคข้อเสนอแนะประกอบด้วยส่วนประกอบ (หรือหน้าที่ของคำ) ดังนี้

1. คำบ่งชี้ข้อเสนอแนะ (Suggestion indicators หรือ suggestion words: SW)
2. คำระบุหัวข้อ (Name Entities หรือ Object: OBJ)
3. วลีเสนอแนะ (Suggestion)

รูปแบบของประโยคข้อเสนอแนะแต่ละประเภท สร้างโดยผู้เชี่ยวชาญ ดังตารางที่ 3.2

ตารางที่ 3.2 รูปแบบของประโยคข้อเสนอแนะแบ่งตามประเภท

ประเภทข้อเสนอแนะ	รูปแบบประโยค
Explicit suggestions	$S_c + OBJ + Suggestion$
	$S_c + Suggestion + OBJ$
	$S_c + Suggestion$
	$OBJ + S_c + Suggestion$
	$S_a + Suggestion + OBJ$
	$S_a + Suggestion$
	$OBJ + S_a + Suggestion$
	$OBJ + Suggestion + S_a$
	$Suggestion + OBJ + S_a$
Query suggestions	$S_q + OBJ + Suggestion$
	$S_q + Suggestion + OBJ$
	$S_q + Suggestion$
	$OBJ + S_q + Suggestion$
	$OBJ + Suggestion + S_q$
Condition suggestions	$S_c + OBJ + Suggestion$
	$OBJ + S_c + Suggestion$

3.2 กรอบการวิเคราะห์เหมืองข้อเสนอแนะ

ในงานวิจัยนี้ การวิเคราะห์เหมืองข้อเสนอแนะเป็นการวิเคราะห์ข้อเสนอแนะที่ซ่อนอยู่ในบทวิจารณ์ของผู้บริโภค โดยมีวัตถุประสงค์เพื่อจำแนกบทวิจารณ์ว่าเป็นข้อเสนอแนะหรือไม่ และจำแนกบทวิจารณ์ที่เป็นข้อเสนอแนะตามประเภทที่กำหนดไว้ 3 ประเภท

จากการศึกษาทฤษฎีและวรรณกรรมที่เกี่ยวข้องกับการวิเคราะห์เหมืองข้อเสนอแนะ ทำให้ผู้วิจัยทราบถึงแนวทางกระบวนการวิเคราะห์ เพื่อกำหนดเป็นกรอบการดำเนินงานวิจัยตามวัตถุประสงค์ของงานวิจัยนี้ ซึ่งมีกระบวนการวิเคราะห์เหมืองข้อเสนอแนะดังนี้

1. รวบรวมบทวิจารณ์ที่เกี่ยวข้องกับรายการโทรทัศน์
2. กระบวนการสร้างฐานความรู้ทางภาษา
3. กระบวนการเตรียมข้อมูล
4. กระบวนการวิเคราะห์ข้อเสนอแนะ
5. ประเมินผลการวิเคราะห์ข้อเสนอแนะ

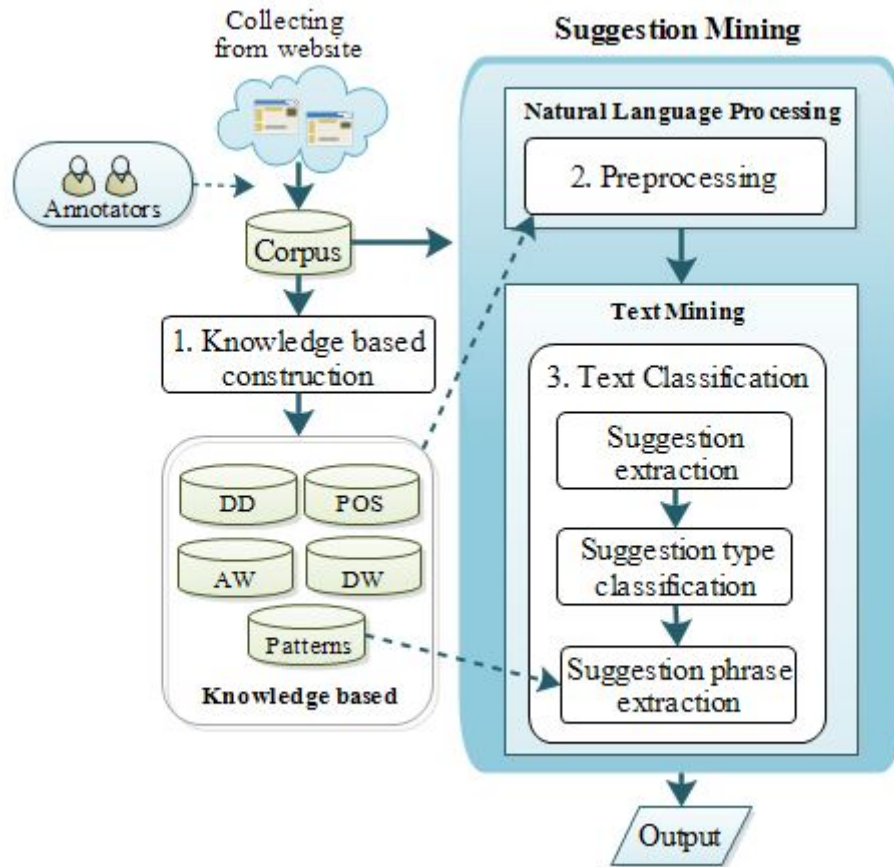
สำหรับกระบวนการหลักของการวิเคราะห์เหมืองข้อเสนอแนะ ได้แก่ กระบวนการที่

(2) – (4) คือกระบวนการสร้างฐานความรู้ทางภาษา กระบวนการเตรียมข้อมูล และกระบวนการวิเคราะห์ข้อเสนอแนะ ดังภาพที่ 3.1 ซึ่งแต่ละหัวข้อมีรายละเอียดดังต่อไปนี้

3.2.1 กระบวนการสร้างฐานความรู้ทางภาษา (Knowledge based construction)

การประมวลผลภาษาธรรมชาติจะอาศัยการวิเคราะห์จากคำ และหน้าที่ของคำเป็นหลัก เพื่อช่วยให้คอมพิวเตอร์มีความสามารถทำความเข้าใจความหมายของข้อความที่นำมาวิเคราะห์ได้ ยังมีฐานความรู้ (Knowledge based: KB) คำศัพท์และหน้าที่ของคำมาก จะยิ่งทำให้ประสิทธิภาพการจำแนกเอกสารมีความถูกต้องยิ่งขึ้น

งานวิจัยนี้ได้อาศัยชุดข้อมูลเรียนรู้ในการสร้าง 4 ฐานความรู้ทางภาษา ได้แก่ คำเฉพาะเจาะจงโดเมน (Domain Dependent: DD) หน้าที่ของคำ (Part-of-speech: POS) คำที่เกิดขึ้นร่วมกัน (Association wordlists: AW) และคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (Domain wordlists: DW) แต่ละฐานความรู้มีรายละเอียดดังนี้



ภาพที่ 3.1 กรอบงานวิจัยเหมืองข้อเสนอแนะ

3.2.1.1 คำเฉพาะเจาะจงโดเมน

ฐานความรู้นี้เป็นการสร้างคำศัพท์แบบเจาะจงหัวข้อเรื่อง (Domain specific) โดยผู้เชี่ยวชาญ โดยแบ่งเป็น 2 ประเภทคือ (1) นิพจน์ระบุค่านามตามหัวข้อที่สนใจ (Name Entity Extraction) เช่น ชื่อคนเฉพาะ, ชื่อสถานที่, ชื่อสินค้าหรือชื่อรายการ ดังตารางที่ 3.3 และ (2) คำบ่งชี้ข้อเสนอแนะ (Suggestion indicators) ซึ่งแบ่งตามประเภทได้ 3 ประเภทตามประเภทข้อเสนอแนะ ดังตารางที่ 3.4

3.2.1.2 หน้าที่ของคำ

การวิเคราะห์คำตามหน้าที่ของคำเป็นการประมวลผลภาษาธรรมชาติในระดับวากยสัมพันธ์ (หรือเชิงโครงสร้าง) ซึ่งจะช่วยให้สามารถเข้าใจความหมายของคำได้ดีขึ้น ในงานวิจัยนี้ได้เลือกใช้การกำกับคำตามหน้าที่ด้วยคลังคำไทยเล็กชิตรอน (Lexitron) ที่ได้รับความนิยมซึ่งพัฒนาโดยเนคเทค

แต่อย่างไรก็ตามข้อเสนอแนะเป็นข้อความที่เน้นการกระทำหรือการแสดงอาการของบุคคล ดังนั้นการวิเคราะห์หน้าที่ของคำกริยาอย่างละเอียดจะช่วยให้ประสิทธิภาพการจำแนกข้อเสนอแนะดียิ่งขึ้น งานวิจัยนี้ได้เลือกใช้การกำกับคำกริยาจากคลังคำไทยออร์คิด (Thai orchid corpus) เนื่องจากมีการแบ่งชนิดของคำอย่างละเอียด โดยชนิดของคำกริยาถูกแบ่งย่อยออกเป็น 3 ประเภท ได้แก่ (1) คำกริยาแสดงอาการ (Action verbs) เป็นคำที่สามารถใช้บ่งชี้การกระทำหรือแนวทางข้อเสนอแนะให้ปฏิบัติได้ชัดเจน (2) คำกริยาแสดงสถานะ (Stative verbs) และ (3) คำกริยาคุณลักษณะ (Attribute verbs) คำอธิบายหน้าที่ของคำและสัญลักษณ์ดังตารางที่ 3.5 จากตารางจะเห็นว่ากริยาแสดงสถานะและคุณลักษณะเป็นคำกริยาที่ไม่มีท่าทางแสดงอาการ จึงตั้งสมมติฐานว่าชนิดของคำกริยาทั้งสองชนิดดังกล่าวไม่สามารถใช้เป็นคำบ่งชี้ข้อเสนอแนะได้ ผู้วิจัยจึงตั้งสมมติฐานว่าการกำกับคำตามหน้าที่ของคำแบบเฉพาะเจาะจงหน้าที่ของคำกริยาจะช่วยให้สามารถจำแนกประโยคข้อเสนอแนะออกจากประโยคที่ไม่ใช่ข้อเสนอได้ ดังตัวอย่าง

ตัวอย่างการกำกับคำตามหน้าที่ของคำด้วยคลังคำไทยเล็กจิตรอน

- ไม่เป็นประโยคข้อเสนอแนะ เช่น “คุณวรรณสิงห์<OBJ3>เป็น<V>พิธีกร<N>ที่<PREP>ดี<ADJ>มาก<ADV>”
- ประโยคข้อเสนอแนะ เช่น “อยาก<S_o>ให้<AUX>เปลี่ยน<V>พิธีกร<N>”

จากการกำกับคำด้วยคลังคำไทยเล็กจิตรอน พบว่าทั้งสองประโยคประกอบด้วยหน้าที่ของคำกริยา ทำการไม่สามารถสกัดเพื่อแยกประโยคข้อเสนอแนะออกมาได้

ตัวอย่างการกำกับคำตามหน้าที่ของคำด้วยคลังคำไทยออร์คิด

- ไม่เป็นประโยคข้อเสนอแนะ เช่น “คุณวรรณสิงห์<OBJ3>เป็น<VSTA>พิธีกร<N>ที่<PREP>ดี<ADJ>มาก<ADV>”
- ประโยคข้อเสนอแนะ เช่น “อยาก<S_o>ให้<AUX>เปลี่ยน<VACT>พิธีกร<N>”

จากการกำกับคำด้วยคลังคำไทยออร์คิด พบว่าประโยคแรกมีคำกริยาแสดงสถานะซึ่งไม่มีหน้าที่แสดงอาการใด ๆ ส่วนประโยคที่สองประกอบด้วยหน้าที่ของคำกริยาที่ใช้แสดงอาการ ทำให้การกำกับหน้าที่ของคำกริยาด้วยคลังคำไทยออร์คิดสามารถวิเคราะห์เพื่อสกัดแยกประโยคข้อเสนอแนะได้ แต่ทั้งนี้การเลือกใช้คลังคำเพื่อกำกับคำตามหน้าที่ของคำขึ้นอยู่กับวัตถุประสงค์ของงานวิจัย

ตารางที่ 3.3 ตัวอย่างคำระบุนามตามหัวข้อที่สนใจ

สัญลักษณ์	ตัวอย่างคำระบุนาม
OBJ0	ชื่อสถานี ThaiPBS
OBJ1	ประเภทรายการ เช่น ข่าว, สารคดี, หนังสื เป็นต้น
OBJ2	ชื่อรายการ เช่น ตอบโจทย์, กินอยู่คือ, พื้นที่ชีวิต เป็นต้น
OBJ3	ชื่อพิธีกรและผู้ประกาศข่าวในรายการ

ตารางที่ 3.4 คำบ่งชี้ข้อเสนอแนะ

สัญลักษณ์	คำบ่งชี้ข้อเสนอแนะ
S_c	อยาก, ขอเสนอแนะ, ดี, น่าจะ, รบกวน, ควร, พิจารณา
S_q	ทำไม, ได้ไหม
S_c	ถ้า, หาก

ตารางที่ 3.5 ตัวอย่างคำกริยาแบบเฉพาะเจาะจง

สัญลักษณ์	คำอธิบายหน้าที่ของคำ
VACT	กริยาแสดงอาการ เช่น พุด กิน เดิน
VSTA	กริยาสภาวะ คือคำกริยาที่ไม่มีท่าที่แสดงอาการ เช่น เห็น รู้
VATT	กริยาคุณลักษณะ คือ คำแสดงคุณลักษณะของคำกริยา เช่น อ้วน สวย

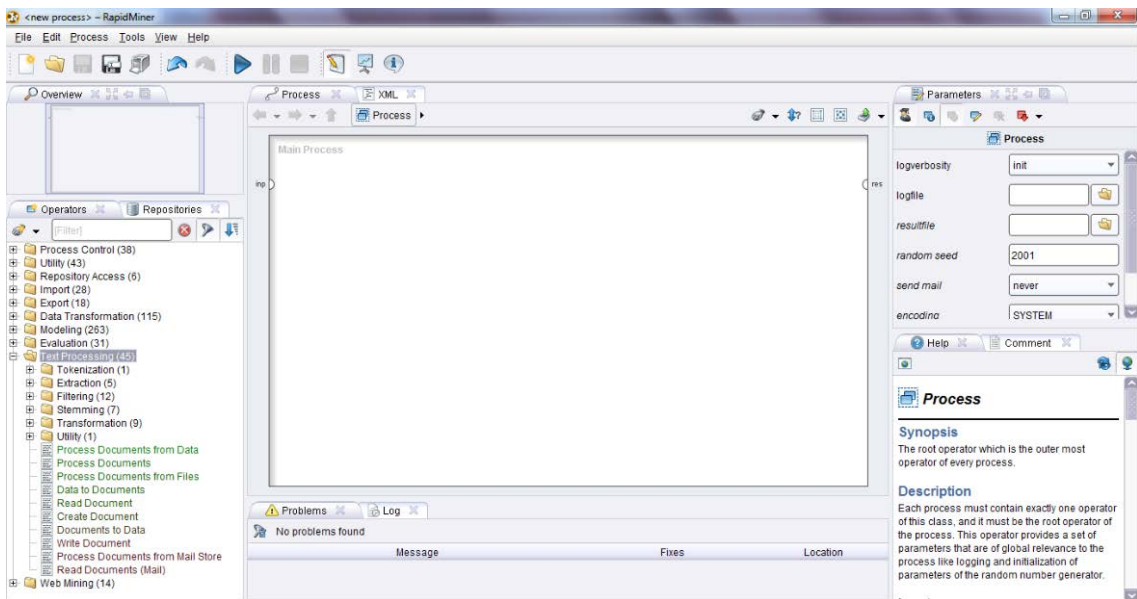
3.2.1.3 คำที่เกิดขึ้นร่วมกันบ่อย

คำหรือวลีที่ใช้แสดงอาการในข้อเสนอแนะมักถูกนำเสนอด้วยคู่ของคำนามกับคำกริยาแสดงอาการ เช่น “อยาก**ให้**เปลี่ยนพิธีกรรายการ” คำนามคือ “พิธีกร” และคำกริยาคือ “เปลี่ยน” จะเห็นได้ว่าการคู่ของคำนามและคำกริยาที่เกิดขึ้นร่วมกันสามารถใช้เป็นวลีบ่งชี้ข้อเสนอแนะได้

ในงานวิจัยนี้จึงนำเสนอวิธีการสร้างฐานความรู้ทางภาษา ด้วยวิธีการหาความสัมพันธ์ระหว่างคำที่เกิดขึ้นร่วมกันบ่อยภายในคลังข้อความเดียวกัน โดยใช้เทคนิคกฎ

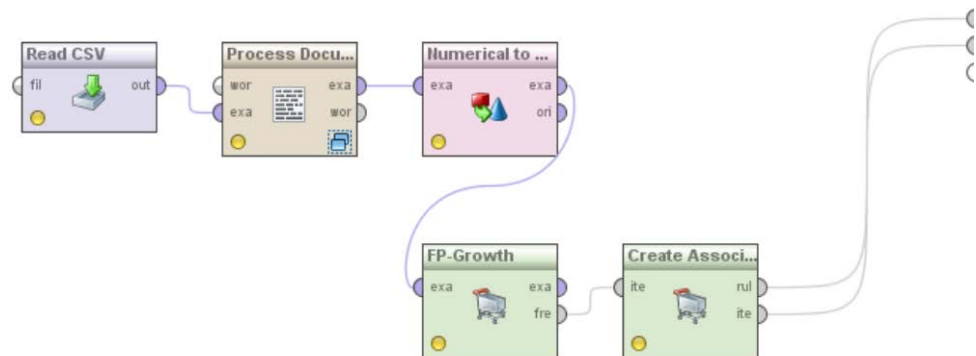
ความสัมพันธ์ที่กำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 1% (Minqing Hu et al., 2004b) เนื่องจากคุณสมบัติของข้อความที่มีลักษณะกระจายค่า (Sparse) หากกำหนดค่าสนับสนุนขั้นต่ำที่สูงเกินไป จะทำให้ได้กฎความสัมพันธ์ที่น้อยเกินไป ไม่เพียงพอต่อการนำไปวิเคราะห์ ซึ่งค่าสนับสนุนขั้นต่ำ 1% หมายความว่าคู่ของคำนามและคำกริยาที่นำมาสร้างเป็นกฎความสัมพันธ์นั้น ต้องเกิดขึ้นร่วมกันมากกว่า 1% ของจำนวนประโยคทั้งหมดในคลังข้อความ กฎความสัมพันธ์ของคำที่ได้หรือคำที่เกิดขึ้นร่วมกันบ่อย (Association Wordlists: AW) จะถูกเก็บไว้เป็นฐานความรู้ทางภาษา (Knowledge based: KB)

ผู้วิจัยเลือกใช้เครื่องมือจากโปรแกรม Rapid miner ในการค้นหาคำที่เกิดขึ้นร่วมกันบ่อย โดยรับข้อมูลนำเข้าเป็นไฟล์ CSV หน้าจอแสดงการทำงานของโปรแกรม ดังภาพที่ 3.2-3.4 และตัวอย่างกฎความสัมพันธ์ดังตารางที่ 3.6



ภาพที่ 3.2 หน้าจอโปรแกรม Rapid miner

Main Process



ภาพที่ 3.3 ตัวอย่างการนำเข้าข้อมูลเพื่อการค้นหาคำที่เกิดขึ้นร่วมกันบ่อย

process_asso – RapidMiner

File Edit Process Tools View Help

Result Overview | FrequentItemSets (FP-Growth) | Asso

Table View Annotations

No. of Sets: 28	Size	Support	Item 1	Item 2
Total Max. Size: 2	2	0.035	วัน	เวลา
	2	0.012	วัน	แพร์
Min. Size: <input type="text" value="2"/>	2	0.021	วัน	ต้นไม้
	2	0.010	วัน	องค์กร
Max. Size: <input type="text" value="2"/>	2	0.019	วัน	ช่วง
Contains Item:	2	0.010	วัน	วิทยุ
<input type="text"/>	2	0.016	วัน	ระบอบ
<input type="text"/>	2	0.011	วัน	อาคาร
<input type="text"/>	2	0.011	วัน	สัญญาณ
<input type="text"/>	2	0.012	วัน	บริษัท
<input type="text"/>	2	0.015	วัน	อาทิตย์
<input type="text"/>	2	0.011	วัน	ระยะ
<input type="text"/>	2	0.013	เวลา	ต้นไม้
<input type="text"/>	2	0.017	เวลา	ช่วง

Update View

ภาพที่ 3.4 ตัวอย่างหน้าจอแสดงผลลัพธ์ของการค้นหาคำที่เกิดขึ้นร่วมกันบ่อย
ด้วยโปรแกรม Rapid miner

ตารางที่ 3.6 ตัวอย่างคู่ของคำนามและคำกริยา (AW)

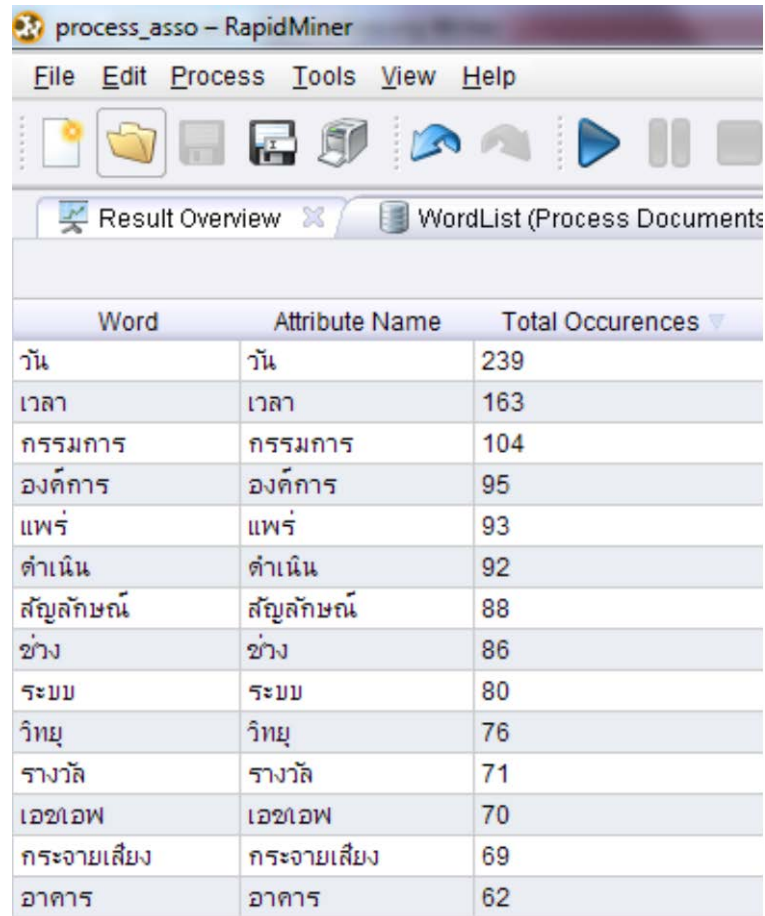
Item 1	Item 2	Support
พิธีกร (AW-N)	พูด (AW-VACT)	0.031
เนื้อหา (AW-N)	ปรับปรุง (AW-VACT)	0.025
เวลา (AW-N)	ออกอากาศ(AW-VACT)	0.019
พิธีกร (AW-N)	เปลี่ยน (AW-VACT)	0.02

3.2.1.4 คำเฉพาะเจาะจงที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน

การสร้างฐานความรู้วิธีนี้เป็นกรสร้างคำศัพท์เฉพาะเจาะจงแบบอัตโนมัติแตกต่างจากการสร้างฐานความรู้ในหัวข้อที่ 3.2.1.1 เรื่องการสร้างฐานความรู้ทางคำแบบคำเฉพาะเจาะจงโดเมน ที่เป็นวิธีการสร้างฐานความรู้โดยผู้เชี่ยวชาญทำให้คำศัพท์ที่ได้มีจำนวนจำกัดอยู่ภายใต้คลังข้อความ และขึ้นอยู่กับความรู้ของผู้เชี่ยวชาญเท่านั้น แต่การสร้างฐานความรู้ทางคำแบบคำเฉพาะเจาะจงที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน จะทำให้ได้คำศัพท์ที่ไม่เจาะจงภายในคลังข้อความ แต่สามารถสร้างฐานความรู้ทางคำได้จากแหล่งความรู้ภายนอก นำไปสู่คำศัพท์ใหม่ที่ยังไม่มีอยู่ในคลังข้อความได้

การสร้างฐานความรู้ทางภาษาในกระบวนการนี้เป็นการค้นหาคำเฉพาะเจาะจงโดเมนแบบอัตโนมัติ จากคำที่เกิดขึ้นบ่อยจากแหล่งข้อมูลภายนอก เช่น เว็บไซต์วิกิพีเดีย (www.wikipedia.com) เป็นการเก็บรวบรวมข้อมูลจากโดเมนที่ใกล้เคียงกัน เช่น โดเมนงานวิจัยคือรายการโทรทัศน์ จะทำการรวบรวมข้อมูลจากโดเมนที่ใกล้เคียงกันที่เกี่ยวข้องกับรายการโทรทัศน์ทั้งหมด เพื่อนำมาวิเคราะห์หาคำที่เกิดขึ้นบ่อย โดยเลือกเฉพาะคำที่มีความถี่สูง 100 อันดับแรกเท่านั้น เนื่องจากคำที่มีความถี่ต่ำกว่านั้นมีค่าน้อยกว่า 1% ของจำนวนคำที่พบทั้งหมด ซึ่งถือว่าคำเหล่านั้นไม่มีนัยสำคัญต่อโดเมนที่สนใจ

ผู้วิจัยเลือกใช้เครื่องมือจากโปรแกรม Rapid miner ในการค้นหาความถี่ของคำที่เกิดขึ้นภายใต้โดเมนที่ใกล้เคียงกัน ตัวอย่างหน้าจอแสดงผลการทำงานของโปรแกรม ดังภาพที่ 3.5 และตัวอย่างคำนามและคำกริยาที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน ดังตารางที่ 3.7



The screenshot shows the RapidMiner interface with a table titled 'WordList (Process Documents)'. The table has three columns: 'Word', 'Attribute Name', and 'Total Occurrences'. The data is as follows:

Word	Attribute Name	Total Occurrences
วัน	วัน	239
เวลา	เวลา	163
กรรมการ	กรรมการ	104
องค์การ	องค์การ	95
แพร์	แพร์	93
ตำแหน่ง	ตำแหน่ง	92
สัญลักษณ์	สัญลักษณ์	88
ช่วง	ช่วง	86
ระบบ	ระบบ	80
วิทยุ	วิทยุ	76
รางวัล	รางวัล	71
เอชเอฟ	เอชเอฟ	70
กระจายเสียง	กระจายเสียง	69
อาคาร	อาคาร	62

ภาพที่ 3.5 ตัวอย่างหน้าจอแสดงผลศัพท์ค่าความถี่ของคำ ด้วยโปรแกรม Rapid miner

ตารางที่ 3.7 ตัวอย่างคำเฉพาะเจาะจงที่เกิดขึ้นบ่อย (DW)

สัญลักษณ์	คำเฉพาะเจาะจงที่เกิดขึ้นบ่อย (DW)
DW-N	โทรทัศน์, สัญญาณ, ผู้ชม

3.2.1.5 รูปแบบวลีข้อเสนอแนะที่เกิดขึ้นบ่อย

การสร้างฐานความรู้ของรูปแบบวลีข้อเสนอแนะสร้างขึ้นจากผู้เชี่ยวชาญในการพิจารณารูปแบบของวลีข้อเสนอแนะที่เกิดขึ้นบ่อยในคลังบทวิจารณ์ ดังตัวอย่าง

1) ข้อเสนอแนะทางตรง

รูปแบบ $S_c + \text{Suggestion} + \text{OBJ}$

ตัวอย่างประโยค อยากให้< S_c >ผู้บริหารทบทวนบทบาทการรายงาน<Suggestion>ข่าว<OBJ>ของผู้ประกาศข่าวภาคค่ำด้วย

2) ข้อเสนอแนะเชิงคำถาม

รูปแบบ $\text{OBJ} + S_c + \text{Suggestion}$

ตัวอย่างประโยค รายการเวทีสาธารณะ<OBJ>ทำไม< S_c >ไม่มีเปิดให้แสดงความคิดเห็นสดบ้าง<Suggestion>

3) ข้อเสนอแนะเชิงเงื่อนไข

รูปแบบ $S_c + \text{OBJ} + \text{Suggestion}$

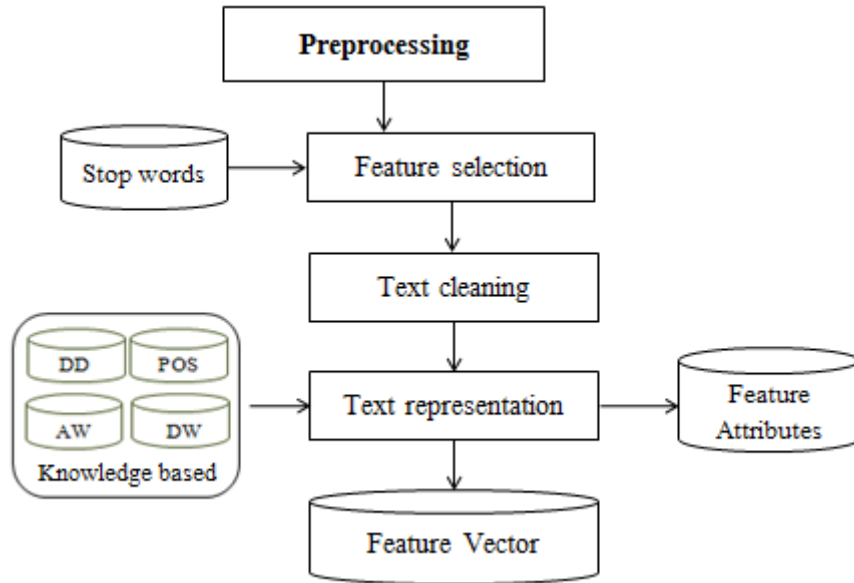
ตัวอย่างประโยค จะดีกว่านี้นะถ้า< S_c >เอารายการท่องโลกกว้าง<OBJ>มาใส่ซับไทเทิลภาษาไทยให้ด้วย<Suggestion>

3.2.2 กระบวนการเตรียมข้อมูล (Preprocessing)

กระบวนการเตรียมข้อมูล เป็นกระบวนการนำบทวิจารณ์จากแหล่งต่าง ๆ เช่น บล็อก เว็บไซต์ บอร์ด เครือข่ายสังคมออนไลน์ เป็นต้น มาผ่านกระบวนการเตรียมข้อมูล เพื่อให้ได้ตัวแทนข้อความ (Feature attributes) ที่อยู่ในรูปแบบโครงสร้างพีเจอร์เวกเตอร์ และเพื่อให้คอมพิวเตอร์สามารถนำข้อความไปประมวลผลได้ ซึ่งกระบวนการเตรียมข้อมูลประกอบด้วย 3 กระบวนการย่อย ดังภาพที่ 3.6 ซึ่งสามารถอธิบายรายละเอียดของกระบวนการย่อยแต่ละกระบวนการได้ดังต่อไปนี้

3.2.2.1 การเลือกคุณลักษณะ

เป็นกระบวนการเลือกคุณลักษณะ เบื้องต้นสำหรับภาษาไทย ด้วยวิธีการตัดคำที่ไม่มีนัยสำคัญออก สำหรับงานวิจัยนี้คำที่ไม่มีนัยสำคัญสำหรับการจำแนกข้อเสนอแนะ ได้แก่ คำหยุด, คำบุพบท, คำสันธาน, สรรพนาม, ลักษณะนาม และตัวเลข



ภาพที่ 3.6 กระบวนการเตรียมข้อมูล

ตัวอย่างประโยคนำเข้า

“ชอบ|รายงาน|พื้นที่|ชีวิต|มาก|ค่ะ| |น่าจะ|ออกอากาศ|เวลา|หัวค่ำ|กว่า|เดิม|”

เหล่านี้คือ

คำว่า “ค่ะ” คือคำหยุดที่ไม่มีนัยสำคัญ ในกระบวนการเลือกคุณลักษณะจะตัดคำ

ผลลัพธ์ คือ

“ชอบ|รายงาน|พื้นที่|ชีวิต|มาก| |น่าจะ|ออกอากาศ|เวลา|หัวค่ำ|กว่า|เดิม|”

3.2.2.2 การกลั่นกรองข้อความ

ได้แก่ การแก้ไขคำผิดให้ถูกต้อง และแก้ไขคำซ้ำซ้อนหรือคำที่มีความหมายเดียวกันให้เป็นคำเดียวกัน และการแก้ไขข้อมูลบางส่วนที่ขาดหายไปสมบูรณ์

ตัวอย่างประโยคนำเข้า

“ชอบ|รายงาน|พื้นที่|ชีวิต|มาก| |น่าจะ|ออกอากาศ|เวลา|หัวค่ำ|กว่า|เดิม|”

ถูกต้อง

คำว่า “รายงาน” สะกดผิด ในกระบวนการกลั่นกรองข้อความจะทำการแก้ไขให้

ผลลัพธ์ คือ

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หวัค้|กว่า|เดิม”

3.2.2.3 การแทนข้อความ

เป็นกระบวนการแทนข้อความด้วยฐานความรู้ทางภาษาที่ได้จากขั้นตอนการสร้างฐานความรู้ในกระบวนการก่อนหน้าในหัวข้อ 3.2.1 ทั้งสิ้น 5 วิธี ได้แก่ (1) การแทนข้อความด้วยคำ (2) การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำด้วยคลังคำเล็กชิตรอน (3) การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่และคำกริยาเฉพาะเจาะจง (4) การแทนข้อความด้วยคำ ร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (AW) และ (5) การแทนข้อความด้วยคำ ร่วมกับคู่ของคำ (AW) ที่มีระยะห่างเหมาะสมที่สุดร่วมกับคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW) ซึ่งแต่ละวิธีมีรายละเอียดดังนี้

วิธีที่ 1 การแทนข้อความด้วยคำที่พบในคลังบทวิจารณ์ (Corpus) ดังนั้นผลลัพธ์ที่ได้จากกระบวนการเตรียมข้อความด้วยวิธีที่ 1 คือ ข้อความ s_i ที่ประกอบด้วยเวกเตอร์ของคำ TF-IDF ของคำ w_{ij} ดังสมการที่ 22

$$s_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{im}) \quad (22)$$

เมื่อ m คือจำนวนคุณลักษณะที่ถูกเลือกมาเพื่อให้เป็นตัวแทนข้อความ

ตัวอย่างประโยชน์นำเข้า

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หวัค้|กว่า|เดิม”

ผลลัพธ์ คือ ข้อความ s_i ที่ประกอบด้วยคำ TF-IDF คำ ดังนี้

$s_i = (\text{ชอบ}, \text{รายการ}, \text{พื้นที่ชีวิต}, \text{มาก}, \text{น่าจะ}, \text{ออกอากาศ}, \text{เวลา}, \text{หวัค้}, \text{กว่า}, \text{เดิม})$

วิธีที่ 2 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำด้วยคลังคำเล็กชิตรอน ดังนั้นข้อมูลนำเข้า s_i จึงประกอบด้วยเวกเตอร์ของคำ TF-IDF ของคำ w_{ij} และเวกเตอร์ของคำ TF-IDF ของหน้าที่ของคำ p_{ij} ดังสมการที่ 23

$$s_1 = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{im}, p_{i1}, p_{i2}, p_{i3}, \dots, p_{in}) \quad (23)$$

เมื่อ m คือจำนวนคำที่ถูกเลือกมาเพื่อให้เป็นตัวแทน ข้อความ และ n คือจำนวนคุณลักษณะหน้าที่ของคำ

ตัวอย่างประโยคนำเข้า ประกอบด้วยคำและหน้าที่ของคำ

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หัวค่ำ|กว่า|เดิม”

<V> <N> <OBJ> <ADV> <S_i> <V> <N> <N> <ADV><ADV>

ผลลัพธ์ คือ ข้อความ s_i ที่ประกอบด้วยคำ TF-IDF คำ และหน้าที่ของคำ ดังนี้

$s_i = (\text{ชอบ, รายการ, พื้นที่ชีวิต, มาก, น่าจะ, ออก อากาศ, เวลา, หัวค่ำ, กว่า, เดิม, V, N, OBJ, ADV, S}_i)$

วิธีที่ 3 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำ ด้วยคลังคำเล็กชิตรอน ส่วนคำที่ทำหน้าที่เป็นกริยาจะทำการกำกับคำกริยาใหม่ ด้วยหน้าที่ของคำกริยาแบบเฉพาะเจาะจง ด้วยคลังคำไทยออร์คิด ข้อความนำเข้า s_i จึงประกอบด้วยเวกเตอร์คำ TF-IDF ของคำกับหน้าที่คำ เช่นเดียวกับวิธีที่ 2 แต่หน้าที่ของคำกริยา (V) จะถูกเปลี่ยนเป็นหน้าที่ของคำกริยาแบบเฉพาะเจาะจง (VACT)

ตัวอย่างประโยคนำเข้า ประกอบด้วยคำและหน้าที่ของคำ

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หัวค่ำ |กว่า|เดิม”

<VACT> <N> <OBJ> <ADV> <S_i> <VACT> <N> <N> <ADV><ADV>

ผลลัพธ์ คือ ข้อความ s_i ที่ประกอบด้วยคำ TF-IDF คำ และหน้าที่ของคำแบบเฉพาะเจาะจง ดังนี้

$s_i = (\text{ชอบ, รายการ, พื้นที่ชีวิต, มาก, น่าจะ, ออกอากาศ, เวลา, หัวค่ำ, กว่า, เดิม, VACT, N, OBJ, ADV, S}_i)$

วิธีที่ 4 การแทนข้อความด้วยคำ ร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (AW) โดยตั้งสมมติฐานว่าคำที่อยู่ห่างกันเกิน k คำ ถือว่าคู่ของคำนั้นไม่มีความสัมพันธ์กัน เมื่อ k แทนระยะห่างระหว่างคำ ดังนั้นคำใดที่ทำหน้าที่เป็นคำนามหรือคำกริยาและเป็นคู่ของคำ AW ที่มีระยะห่างระหว่างกันไม่เกิน k คำ จะถูกกำกับคำใหม่ ด้วย $AW-N$ หรือ $AW-VACT$ ดังตัวอย่าง “ออกอากาศ” กับ “เวลา” เป็นคู่ของคำ AW จึงถูกกำกับคำใหม่ ดังตัวอย่าง

ตัวอย่างประโยคนำเข้า ประกอบด้วยคำ และหน้าที่ของคำ

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หัวค่ำ|กว่า|เดิม”

<VACT> <N> <OBJ> <ADV> <S_a> <AW-VACT> <AW-N> <N> <ADV> <ADV>

ผลลัพธ์ คือ ข้อความ s_i ที่ประกอบด้วยคำ TF-IDF คำ และหน้าที่ของคำ ดังนี้

$s_i = (\text{ชอบ, รายการ, พื้นที่ชีวิต, มาก, น่าจะ, ออกอากาศ, เวลา, หัวค่ำ, กว่า, เดิม,}$

$VACT, N, OBJ, ADV, S_a, AW-VACT, AW-N)$

วิธีที่ 5 การแทนข้อความด้วยคำ ร่วมกับคู่ของคำ (AW) ที่มีระยะห่างเหมาะสมที่สุดร่วมกับคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW) ดังตัวอย่าง “หัวค่ำ” เป็นคำ DW จึงถูกกำกับคำใหม่ ดังนี้

ตัวอย่างประโยคนำเข้า ประกอบด้วยคำ และหน้าที่ของคำ

“ชอบ|รายการ|พื้นที่ชีวิต|มาก|น่าจะ|ออกอากาศ|เวลา|หัวค่ำ |กว่า | เดิม”

<VACT> <N> <OBJ> <ADV> <S_a> <AW-VACT> <AW-N> <DW-N> <ADV> <ADV>

ผลลัพธ์ คือ ข้อความ s_i ที่ประกอบด้วยคำ TF-IDF คำ และหน้าที่ของคำ ดังนี้

$s_i = (\text{ชอบ, รายการ, พื้นที่ชีวิต, มาก, น่าจะ, ออกอากาศ, เวลา, หัวค่ำ, กว่า, เดิม,}$

$VACT, N, OBJ, ADV, S_a, AW-VACT, AW-N, DW-N)$

เมื่อแทนข้อความด้วยฐานความรู้ทางภาษาของแต่ละวิธีแล้ว จากนั้นแทนข้อความให้อยู่ในรูปแบบโครงสร้างพีเจอร์เวกเตอร์ ด้วยคำ TF-IDF ที่มีสมการการคำนวณค่าน้ำหนักของคำ คือ $TF * IDF$ โดยที่ค่า TF คือค่าความถี่ของการเกิดขึ้นของคำ, w ในข้อความ, d คู่ด้วย IDF ค่าส่วนกลับความถี่ของจำนวนข้อความ, d ที่เกิดคำ, w

ตัวอย่าง ประโยค s_1, s_2 และ s_3 ดังนี้

s_1 “น่าจะ|ออกอากาศ|ใหม่|เวลา|หัวค่ำ|”

s_2 “ข่าว|หัวค่ำ|น่าจะ|เปลี่ยน|พิธีกร|ข่าว|ใหม่|”

s_3 “ควร|เปลี่ยน|พิธีกร|ข่าว|หัวค่ำ|”

จากประโยคตัวอย่างจะได้คำที่จะนำมาวิเคราะห์ คือคำ, w ในเซตของประโยคทั้งหมด, D ที่นำมาวิเคราะห์

โดยที่ $D = \{s_1, s_2, s_3\}$ และ $w = \{ \text{“ข้าว”}, \text{“ควร”}, \text{“น้ำจะ”}, \text{“พิธีกร”}, \text{“หัวคำ”}, \text{“ออกอากาศ”}, \text{“เปลี่ยน”}, \text{“เวลา”}, \text{“ใหม่”} \}$

แต่เนื่องจากขนาดของแต่ละข้อความ, s มีความยาว (คำ) ไม่เท่ากัน จึงต้องทำ Vector normalization ให้แต่ละข้อความมีความยาวเท่ากันก่อนคือ เวกเตอร์ 1 หน่วย (Unit vector) เพื่อให้ข้อความที่ถูกแปลงให้อยู่ในรูปแบบพีเจอร์เวกเตอร์สามารถนำไปวิเคราะห์ได้อย่างถูกต้องและเหมาะสม ตัวอย่างที่ได้จากการแปลงข้อความให้เป็นพีเจอร์เวกเตอร์ดังตารางที่ 3.8

ตารางที่ 3.8 ตัวอย่างการแปลงข้อความให้เป็นพีเจอร์เวกเตอร์ ด้วยค่า TF-IDF

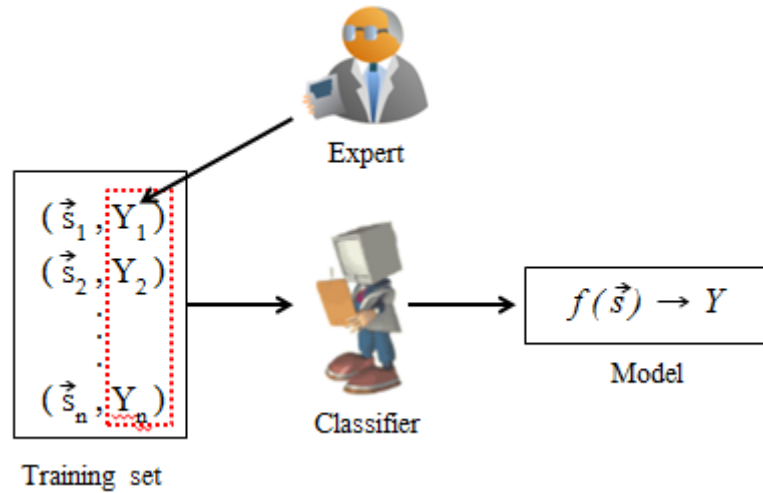
TF-IDF	ข้าว	ควร	น้ำจะ	พิธีกร	หัวคำ	ออกอากาศ	เปลี่ยน	เวลา	ใหม่
s_1	0	0	0.245	0	0	0.663	0	0.663	0.245
s_2	0.707	0	0.345	0.345	0	0	0.345	0	0.345
s_3	0.311	0.843	0	0.311	0	0	0.311	0	0

เมื่อได้ข้อความที่อยู่ในรูปแบบพีเจอร์เวกเตอร์แล้ว จะทำการเก็บคำคุณลักษณะที่ถูกเลือกไว้ในคลังคำคุณลักษณะ เพื่อนำคลังคำคุณลักษณะดังกล่าวไปใช้กับชุดข้อมูลทดสอบต่อไป และเก็บข้อความที่ถูกคำนวณให้อยู่ในรูปแบบพีเจอร์เวกเตอร์ไว้เพื่อนำเข้าสู่กระบวนการจำแนกข้อความ

3.2.3 กระบวนการจำแนกข้อความ (Text classification)

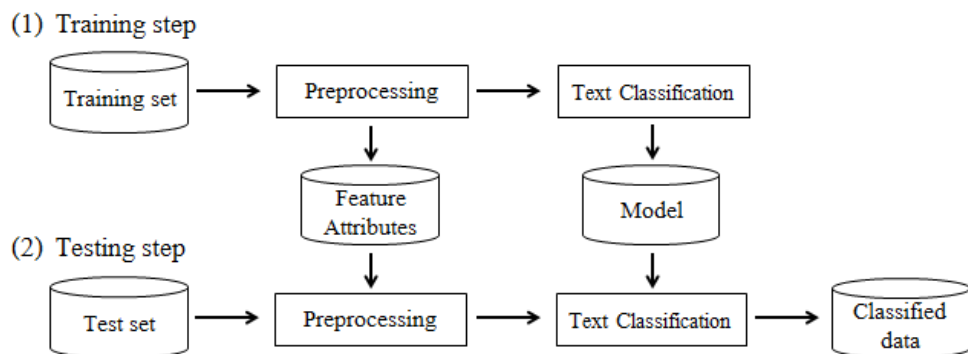
กระบวนการจำแนกข้อความเป็นกระบวนการเรียนรู้จากชุดข้อมูลเรียนรู้ที่มีการกำหนดผลเฉลยไว้แล้วโดยผู้เชี่ยวชาญ เพื่อสร้างแบบจำลอง (Model) การเลือกกลุ่มที่ดีที่สุดสำหรับการจัดกลุ่มข้อมูล จากนั้นนำชุดข้อมูลทดสอบ ที่ไม่รู้ผลเฉลย มาทดสอบแบบจำลองดังกล่าว เพื่อให้ได้กลุ่มของข้อมูลที่เหมาะสม เรียกว่าการเรียนรู้แบบมีผลเฉลย ดังภาพที่ 3.7 ซึ่งสามารถแบ่งได้เป็น 2 กระบวนการ ดังภาพที่ 3.8 ได้แก่

1. การเรียนรู้ (Training step) จากชุดข้อมูลเรียนรู้ เพื่อสร้างแบบจำลองการจำแนกข้อเสนอแนะ
2. กระบวนการทดสอบ (Testing step) แบบจำลองการจำแนกข้อเสนอแนะด้วยชุดข้อมูลทดสอบ



ภาพที่ 3.7 การเรียนรู้แบบมีผลเฉลย (Supervised learning technique)

แหล่งที่มา: Atorn Nuntiyagul, 2006.

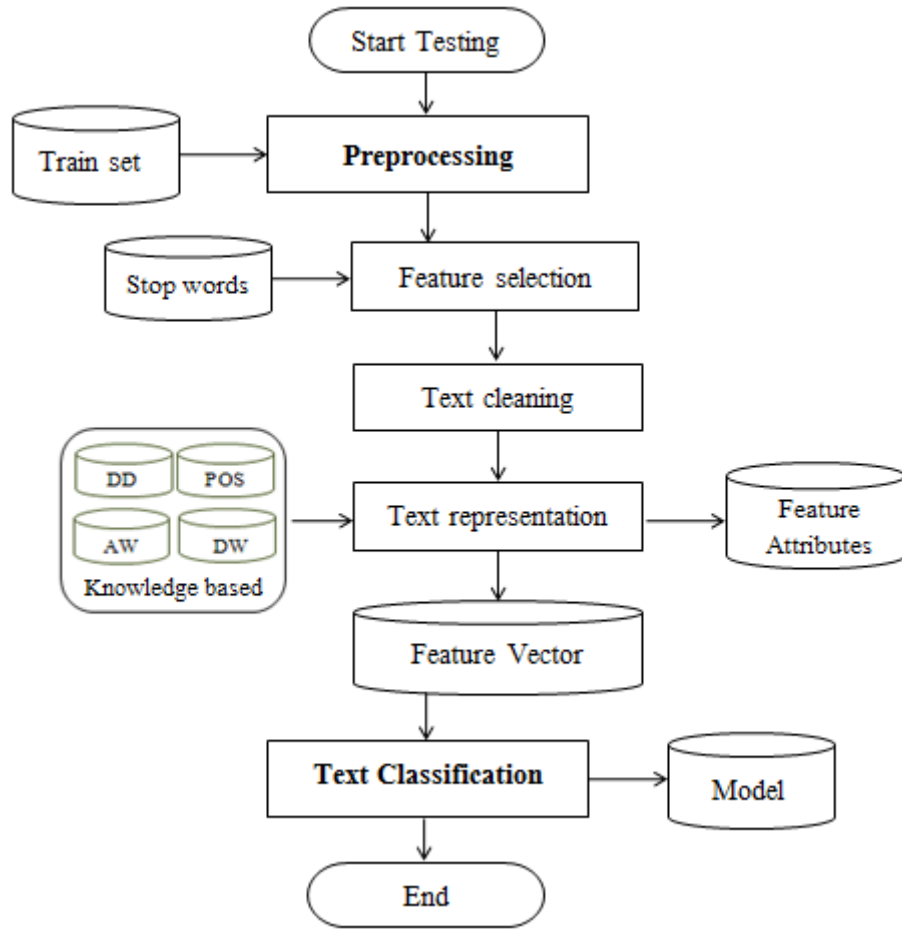


ภาพที่ 3.8 กระบวนการเรียนรู้แบบมีผลเฉลย (Supervised learning algorithm)

3.2.3.1 การเรียนรู้จากชุดข้อมูลเรียนรู้

เป็นกระบวนการสร้างแบบจำลองการจำแนกข้อเสนอแนะ สามารถอธิบายภาพรวมของกระบวนการเรียนรู้เพื่อสร้างแบบจำลองได้ ดังภาพที่ 3.9

กระบวนการเรียนรู้เพื่อสร้างแบบจำลอง เริ่มต้นที่การนำข้อมูลจากชุดข้อมูลเรียนรู้ที่เก็บอยู่ในคลังบทวิจารณ์ และถูกแบ่งส่วนให้ผู้เชี่ยวชาญกำหนดผลเฉลยไว้แล้ว ไปผ่านกระบวนการเตรียมข้อความ ตามรายละเอียดดังหัวข้อ 3.2.2



ภาพที่ 3.9 กระบวนการเรียนรู้เพื่อสร้างแบบจำลองการจำแนกข้อเสนอแนะ

กระบวนการสร้างแบบจำลองการจำแนกข้อความ ผู้วิจัยเลือกใช้เครื่องมือจากโปรแกรม Rapid miner ในการเรียนรู้จากชุดข้อมูลเรียนรู้ เมื่อได้แบบจำลองการจำแนกข้อความที่ดีที่สุดแล้ว จะเก็บแบบจำลองไว้ในคลังแบบจำลอง เพื่อนำไปใช้กับชุดข้อมูลทดสอบต่อไป

การสร้างแบบจำลองสำหรับการทำเหมืองข้อเสนอแนะในกระบวนการจำแนกข้อความนี้ได้แบ่งออกเป็น 2 กระบวนการ คือ กระบวนการสกัดข้อเสนอแนะ และกระบวนการจำแนกประเภทข้อเสนอแนะ แต่ละกระบวนการมีรายละเอียดดังนี้

1) กระบวนการสกัดข้อเสนอแนะ

กระบวนการสกัดข้อเสนอแนะเป็นกระบวนการวิเคราะห์ห้วงทวิจรรย์ ที่ถูกแทนข้อความให้อยู่ในรูปแบบพีเจอร์เวกเตอร์แล้ว จากนั้นมาเป็นข้อมูลนำเข้าของกระบวนการ มี

วัตถุประสงค์เพื่อสกัดแยกบทวิจารณ์ที่เป็นข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น (ข้อเท็จจริง และความคิดเห็น) และสร้างแบบจำลองที่เหมาะสมที่สุดสำหรับการสกัดข้อเสนอแนะ

วิธีการหาแบบจำลองที่เหมาะสมที่สุด ทำโดยเปรียบเทียบประสิทธิภาพการจำแนกข้อความของ 3 อัลกอริทึม ได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ นาอ์ฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ร่วมกับการแทนข้อความด้วยฐานความรู้ทางภาษาแบบวิธีที่ 1 (วิเคราะห์จากคำอย่างเดียว)

จากนั้นเปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะของชุดข้อมูลเรียนรู้ของแต่ละอัลกอริทึม ด้วยค่าความแม่นยำ ค่าความระลึก และค่าถ่วงดุล

จากผลการทดสอบการสกัดข้อเสนอแนะด้วยแบบจำลองที่สร้างมาจากชุดข้อมูลเรียนรู้ พบว่าซัพพอร์ตเวกเตอร์แมชชีนเป็นอัลกอริทึมที่ดีที่สุดสำหรับการสกัดข้อเสนอแนะ

เมื่อได้อัลกอริทึมที่ดีที่สุดแล้ว จะทดสอบแทนค่าข้อความด้วยวิธีการแทนความรู้ทางภาษา ดังวิธีที่ 2-5 (จากกระบวนการสร้างฐานความรู้ทางภาษา) ร่วมกับการสกัดข้อเสนอแนะด้วยอัลกอริทึมที่ดีที่สุด (ซัพพอร์ตเวกเตอร์แมชชีน) และทำการปรับค่าพารามิเตอร์ (Parameter tuning) เพื่อให้ได้ผลลัพธ์ของการสกัดข้อเสนอแนะที่ดีที่สุด โดยเปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะด้วยวิธีการแทนค่าข้อความด้วยความรู้ทางภาษาของแต่ละวิธี เพื่อให้ได้แบบจำลอง การสกัดข้อเสนอแนะ จากนั้นเก็บแบบจำลองการสกัดข้อเสนอแนะดังกล่าวไว้ในคลังแบบจำลอง เพื่อนำไปใช้กับชุดข้อมูลทดสอบต่อไป

2) กระบวนการจำแนกประเภทข้อเสนอแนะ

กระบวนการจำแนกประเภทข้อเสนอแนะเป็นกระบวนการที่นำชุดข้อมูลเรียนรู้ ที่ถูกกำหนดผลเฉลยของประเภทข้อเสนอแนะไว้แล้ว 3 ประเภทคือ ข้อเสนอแนะทางตรง ข้อเสนอแนะเชิงคำถาม และข้อเสนอแนะแบบมีเงื่อนไข นำมาเรียนรู้เพื่อสร้างเป็นแบบจำลองจำแนกประเภท ด้วยอัลกอริทึมการสกัดข้อเสนอแนะที่ดีที่สุด (ซึ่งจากผลการทดสอบ พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน เป็นอัลกอริทึมที่มีประสิทธิภาพการสกัดข้อเสนอแนะมากที่สุด) โดยมีข้อความนำเข้าเป็นข้อความที่อยู่ในรูปแบบพีเจอาร์แวกเตอร์และถูกแทนข้อความด้วยวิธีการแทนความรู้ทางภาษาที่มีประสิทธิภาพการจำแนกข้อเสนอแนะดีที่สุด (จากกระบวนการก่อนหน้า) จากนั้นเก็บแบบจำลองที่ได้จากการกระบวนการเรียนรู้ไว้ในคลังแบบจำลอง เพื่อนำไปใช้กับชุดข้อมูลทดสอบต่อไป

ผลลัพธ์ที่ได้จากกระบวนการเรียนรู้คือ คลังคำที่ถูกเลือกเพื่อเป็นตัวแทนข้อความ และแบบจำลองการวิเคราะห์เหมืองข้อเสนอแนะ ซึ่งแบ่งออกเป็น 2 แบบจำลอง คือ

แบบจำลองการสกัดหาข้อเสนอแนะ และแบบจำลองการจำแนกประเภทข้อเสนอแนะ โดยผลลัพธ์ทั้งหมดที่ได้จากกระบวนการเรียนรู้นี้จะนำไปใช้กับกระบวนการทดสอบต่อไป

3.2.3.2 การทดสอบ แบบจำลองการวิเคราะห์เหมืองข้อเสนอแนะ

กระบวนการทดสอบแบบจำลองการทำเหมืองข้อเสนอแนะ เป็นกระบวนการนำชุดข้อมูลทดสอบที่ยังไม่รู้ผลเฉลยจากคลังบทวิจารณ์มาทำการทดสอบแบบจำลองที่ได้จากขั้นตอนการเรียนรู้ โดยเริ่มต้นจากนำข้อมูลทดสอบไปผ่านกระบวนการเตรียมข้อมูล ที่ประกอบด้วยขั้นตอนการเลือกคุณลักษณะ ซึ่งคุณลักษณะดังกล่าวจะถูกดึงมาจากคลังคำที่ถูกเลือกเพื่อเป็นตัวแทนข้อความ ที่ได้จากผลลัพธ์ของกระบวนการการเรียนรู้ จากนั้นเข้าสู่ขั้นตอนการกลั่นกรองและการแทนข้อความด้วยวิธีการแทนความรู้ทางภาษา ให้อยู่ในรูปแบบพีเจอร์เวกเตอร์ด้วยค่า TF-IDF และจากนั้นทำการทดสอบแบบจำลอง ภาพรวมของกระบวนการทดสอบ แบบจำลองการวิเคราะห์เหมืองข้อเสนอแนะ (Suggestion mining) ดังภาพที่ 3.10

สำหรับงานวิจัยนี้จะทำการเปรียบเทียบประสิทธิภาพการวิเคราะห์ข้อเสนอแนะด้วยอัลกอริทึมต้นไม้ตัดสินใจ, นาอูฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ซึ่งข้อมูลที่นำมาทดสอบกับแต่ละอัลกอริทึมจะเป็นข้อมูลชุดเดียวกันทั้งหมด

การทดสอบแบบจำลองแบ่งเป็น 3 กระบวนการ คือ กระบวนการสกัดข้อเสนอแนะ กระบวนการจำแนกประเภทข้อเสนอแนะ และกระบวนการสกัดวลีข้อเสนอแนะ แต่ละกระบวนการมีรายละเอียดดังนี้

1) กระบวนการสกัดข้อเสนอแนะ

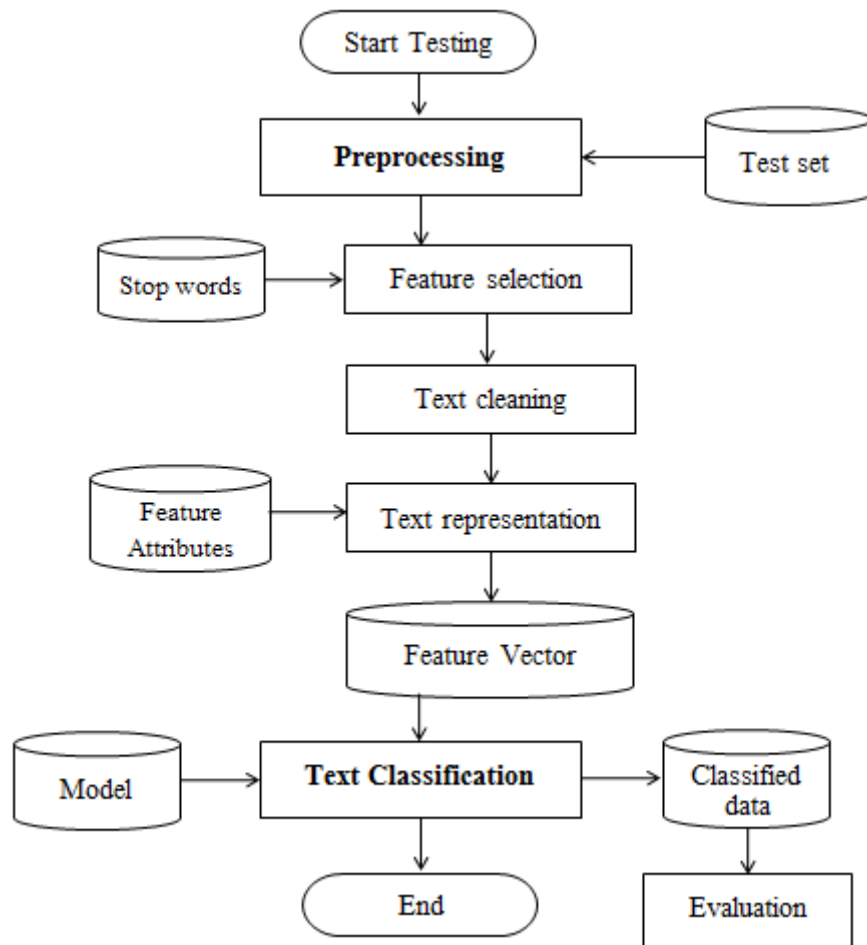
เป็นกระบวนการวิเคราะห์บทวิจารณ์ที่เป็นชุดทดสอบ คือบทวิจารณ์ที่ยังไม่รู้ผลเฉลย และถูกแทนข้อความให้อยู่ในรูปแบบพีเจอร์เวกเตอร์แล้ว จากนั้นนำมาเป็นข้อความนำเข้าของกระบวนการทดสอบการสกัดหาข้อเสนอแนะ โดยวิเคราะห์บทวิจารณ์ด้วยแบบจำลองการสกัดข้อเสนอแนะที่ได้จากผลลัพธ์ของกระบวนการเรียนรู้ และวัดประสิทธิภาพการสกัดหาข้อเสนอแนะ ด้วยค่าความแม่นยำ ค่าความระลึก และค่าถ่วงดุล

2) กระบวนการจำแนกประเภทข้อเสนอแนะ

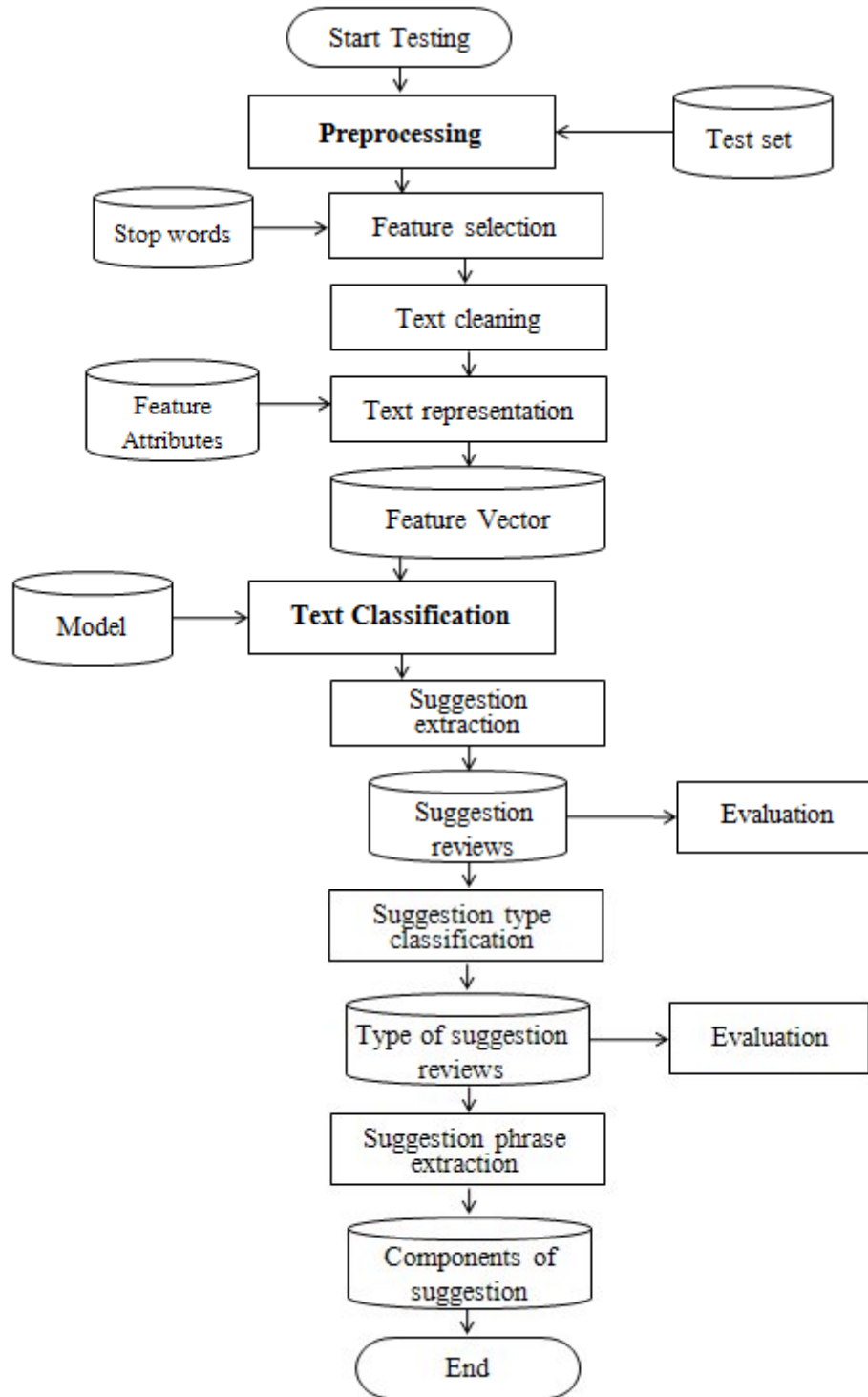
กระบวนการจำแนกประเภทข้อเสนอแนะเป็นกระบวนการที่นำผลลัพธ์ที่ถูกสกัดแยกออกมาแล้วว่าเป็นข้อเสนอแนะจากกระบวนการทดสอบก่อนหน้าในหัวข้อกระบวนการสกัดข้อเสนอแนะ มาจำแนกตามประเภทของข้อเสนอแนะ 3 ประเภท ด้วยแบบจำลองการจำแนกประเภทข้อเสนอแนะที่ได้จากกระบวนการเรียนรู้ ดังภาพที่ 3.11 จากนั้นวัดประสิทธิภาพของกระบวนการจำแนกประเภทข้อเสนอแนะด้วยค่าเฉลี่ยแบบให้น้ำหนักทุกเอกสารเท่ากัน (Micro averaging) ประกอบด้วยค่าเฉลี่ยความแม่นยำ (Micro Precision) ค่าเฉลี่ยความระลึก (Micro Recall)

3) กระบวนการสกัดวลีข้อเสนอแนะ

กระบวนการสกัดวลีข้อเสนอแนะเป็นกระบวนการสกัดหาส่วนประกอบของประโยคข้อเสนอแนะ โดยที่อินพุตของกระบวนการสกัดวลีข้อเสนอแนะนี้คือ ผลลัพธ์ที่ได้จากกระบวนการจำแนกประเภทข้อเสนอแนะ และนำผลลัพธ์ที่ได้ดังกล่าวไปพิจารณาเพื่อสกัดหาส่วนประกอบของวลีข้อเสนอแนะ



ภาพที่ 3.10 กระบวนการทดสอบแบบจำลองการวิเคราะห์เหมืองข้อเสนอแนะ



ภาพที่ 3.11 กระบวนการทดสอบแบบจำลองการทำเหมืองข้อเสนอแนะ 3 กระบวนการ

บทที่ 4

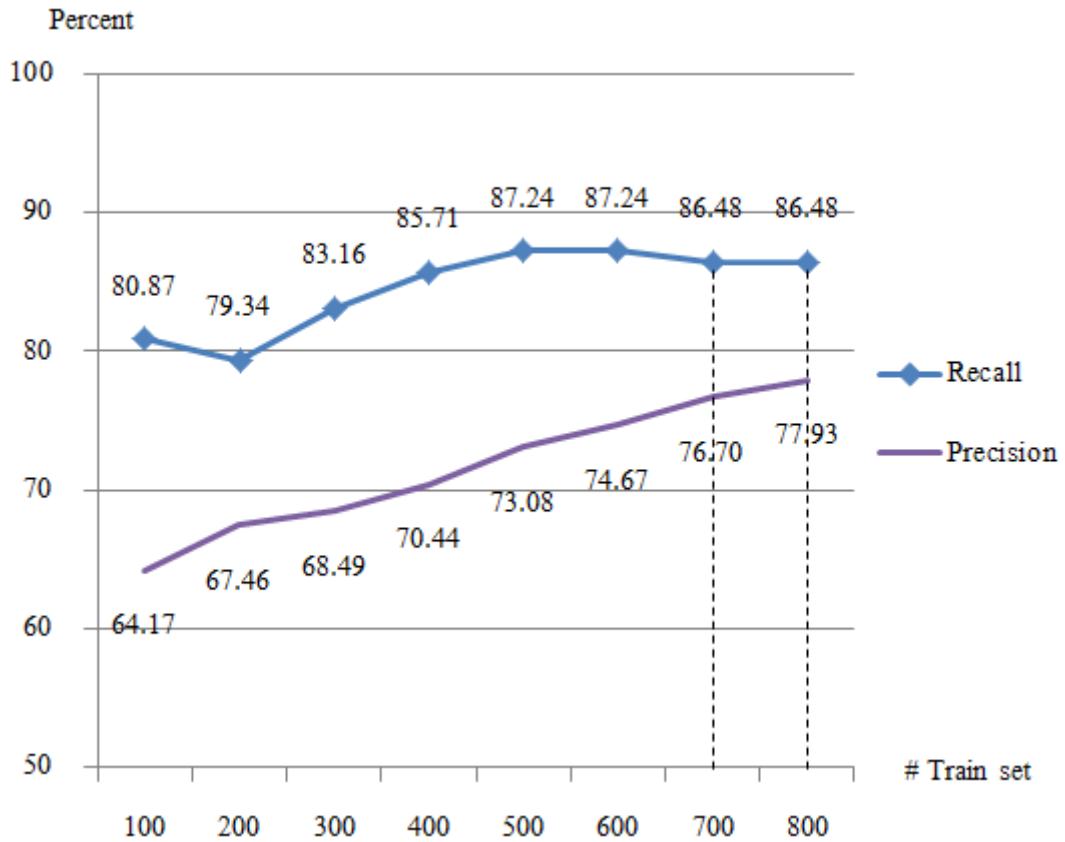
การทดลองและวัดประสิทธิภาพของกระบวนการ

การทดลองและวัดประสิทธิภาพของกระบวนการทำเหมืองข้อมูลแบบเชื่อมโยงที่กล่าวถึงในบทนี้ ประกอบด้วย 2 หัวข้อหลัก ได้แก่ (1) ข้อมูลและเครื่องมือที่ใช้ในการวิจัย และ (2) กระบวนการทดสอบการวิเคราะห์เหมืองข้อมูลแบบเชื่อมโยง ซึ่งแต่ละหัวข้อมีรายละเอียดดังต่อไปนี้

4.1 ข้อมูลและเครื่องมือที่ใช้ในการวิจัย

การทำเหมืองข้อมูลแบบเชื่อมโยงสำหรับงานวิจัยนี้ได้เก็บรวบรวมความคิดเห็นที่เกี่ยวกับรายการทีวี จากเว็บบอร์ด สมุดเยี่ยมชมเว็บไซต์ และเครือข่ายสังคมออนไลน์ จำนวนทั้งสิ้น 1,105 เอกสาร แบ่งออกเป็น 2,561 ประโยค ประกอบด้วยความคิดเห็นทั่วไป 1,774 ประโยคและข้อเสนอแนะ 787 ประโยค และแบ่งข้อมูลออกเป็น 2 ส่วนเท่า ๆ กัน คือส่วนหนึ่งสำหรับชุดข้อมูลการเรียนรู้ และอีกส่วนหนึ่งสำหรับชุดข้อมูลทดสอบ แต่ละชุดข้อมูลจะประกอบด้วยประโยคที่เป็นข้อเสนอแนะและไม่ใช่ข้อเสนอแนะ และใช้เครื่องมือในโปรแกรม Rapidminer ในการสร้างแบบจำลองจากชุดข้อมูลเรียนรู้และทดสอบแบบจำลองดังกล่าวด้วยชุดข้อมูลทดสอบ

ชุดข้อมูลการเรียนรู้เป็นชุดข้อมูลที่ให้ผู้เชี่ยวชาญอ่านและกำกับประโยคว่าเป็นข้อเสนอแนะหรือไม่ ซึ่งสามารถแบ่งเป็นประโยคข้อเสนอแนะ 395 ประโยคและไม่ใช่ข้อเสนอแนะ 419 ประโยค รวมทั้งสิ้นจำนวน 814 ประโยค การทดลองได้เลือกจำนวนชุดข้อมูลเรียนรู้ที่แตกต่างกันไป ได้แก่ ชุดข้อมูลจำนวน 50, 100, 200, 300, 400, 500, 600, 700 และ 800 นำมาทดลองสร้างแบบจำลองการสกัดข้อเสนอแนะ และนำมาทดสอบด้วยชุดข้อมูลทดสอบ ผลการทดลองดังภาพที่ 4.1 พบว่า ค่าระยะเริ่มแรกที่ที่จำนวนชุดข้อมูลทดสอบ 700-800 ซึ่งสามารถสรุปได้ว่าการเพิ่มจำนวนชุดข้อมูลเรียนรู้เพิ่มไม่ทำให้แบบจำลองการจำแนกข้อเสนอแนะดึงข้อเสนอแนะออกมาจากบทวิจารณ์ได้มากขึ้น ดังนั้นงานวิจัยนี้จึงเลือกใช้ข้อมูลชุดเรียนรู้จำนวน 814 ชุด เพื่อให้สามารถแบ่งข้อมูลเป็นสองส่วนได้อย่างเท่า ๆ กัน



ภาพที่ 4.1 เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะด้วยจำนวนชุดข้อมูลเรียนรู้ที่แตกต่างกัน

ตัวอย่างชุดข้อมูลเรียนรู้ สำหรับประโยคข้อเสนอแนะ หรือตัวอย่างที่เป็นบวก คือประโยคที่มีคำบ่งชี้ข้อเสนอแนะ (Suggestion indicator) ปรากฏอยู่ในประโยค เช่น

“ละครเรื่อง *Code Blue* ทำไม่ถึงจบเร็วครับ ดูได้ไม่กี่ตอนเองครับ”

ตัวอย่างชุดข้อมูลเรียนรู้ สำหรับประโยคที่ไม่เป็นข้อเสนอแนะ หรือตัวอย่างที่เป็นลบ แบ่งออกเป็น 2 ประเภท คือ ประโยคความคิดเห็นและประโยคข้อเท็จจริง เช่น

- ประโยคความคิดเห็น คือประโยคที่มีคำแสดงข้อความคิดเห็นปรากฏอยู่ในประโยค เช่น

“พิธีกรดี ผู้ร่วมรายการทรงภูมิและไว้ตัวเหมาะสมมากครับ เปิดผ่านแค่ 30 วินาทีแล้วต้องหยุดดูที่ช่องนี้จนจบรายการเลยครับ”

- ประโยคข้อเท็จจริง คือประโยคที่ไม่มีทั้งคำบ่งชี้ข้อเสนอแนะและคำแสดงข้อความคิดเห็นปรากฏอยู่ในประโยค เช่น

“รายการย้อนหลังของ ThaiPBS ยังขาดยุทธการณั้โลกอีกหนึ่งรายการคะ”

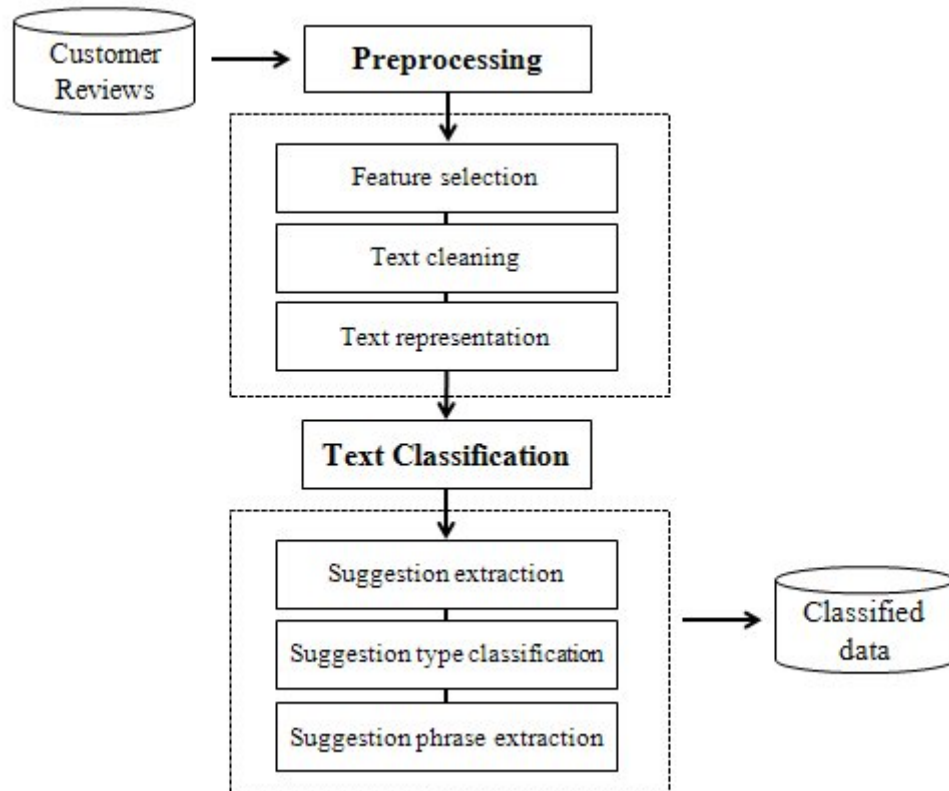
เมื่อได้จำนวนชุดข้อมูลเรียนรู้ที่เหมาะสมแล้ว นำข้อมูลชุดเรียนรู้ดังกล่าวมาสร้างแบบจำลองการจำแนกข้อเสนอแนะ โดยใช้วิธีการเรียนรู้แบบ 10-fold cross-validation เนื่องจากการเรียนรู้แบบ 10 fold cross validation สามารถให้ค่าความถูกต้องเป็นที่น่าพอใจและใช้เวลาในการทำงาน (Time complexity) ไม่มากนัก (Ron, 1995 อ้างถึงใน สิทธิโชค มุกดาสกุลภิบาล, 2551) การเรียนรู้เพื่อสร้างแบบจำลองด้วยวิธีนี้เป็นการแบ่งข้อมูลสำหรับการเรียนรู้ออกเป็น 10 ชุดย่อย (folds) ฝึกสอนด้วยชุดข้อมูล 9 ชุด ส่วนที่เหลืออีก 1 ชุดเก็บไว้สำหรับการทดสอบที่สร้างจาก 9 ชุดข้างต้น ทำการทดลองซ้ำ 10 ครั้งโดยเปลี่ยนชุดข้อมูลสำหรับฝึกสอนและทดสอบใหม่ ซึ่งทุกชุดข้อมูลเรียนรู้ถูกนำมาเป็นชุดทดสอบ เพื่อสร้างแบบจำลอง จากนั้นนำข้อมูลทดสอบจำนวน 815 ประโยค เป็นชุดทดสอบแบบจำลองที่สร้างมาจากชุดข้อมูลเรียนรู้

4.2 กระบวนการทดสอบการวิเคราะห์เหมืองข้อเสนอแนะ

กระบวนการวิเคราะห์เหมืองข้อเสนอแนะ ประกอบด้วย 5 กระบวนการ ได้แก่

1. รวบรวมบทวิจารณ์ที่เกี่ยวข้องกับรายการโทรทัศน์
2. กระบวนการสร้างฐานความรู้ทางภาษา
3. กระบวนการเตรียมข้อมูล
4. กระบวนการวิเคราะห์ข้อเสนอแนะ
5. ประเมินผลการวิเคราะห์ข้อเสนอแนะ

สำหรับการทดสอบกระบวนการวิเคราะห์เหมืองข้อเสนอแนะเกิดขึ้นหลังจากผ่านกระบวนการเรียนรู้และสร้างแบบจำลองสำหรับการวิเคราะห์เหมืองข้อเสนอแนะแล้ว จากนั้นจึงนำแบบจำลองดังกล่าวมาวิเคราะห์ข้อมูลชุดทดสอบที่ยังไม่ทราบผลเฉลย ทำการวิเคราะห์เพื่อสกัดข้อเสนอแนะ จำแนกประเภทข้อเสนอแนะ และสกัดหัวข้อข้อเสนอแนะ ซึ่งกระบวนการทดสอบแบบจำลองประกอบด้วย 2 กระบวนการหลัก (1) กระบวนการเตรียมข้อมูล และ (2) กระบวนการวิเคราะห์ข้อเสนอแนะ ดังภาพที่ 4.2



ภาพที่ 4.2 กระบวนการหลักสำหรับการวิเคราะห์ข้อเสนอแนะ

4.2.1 กระบวนการเตรียมข้อมูล

กระบวนการเตรียมข้อมูลประกอบด้วย 3 กระบวนการย่อย คือ การเลือกคุณลักษณะของข้อความ การกลั่นกรองข้อความ และการแทนข้อความ มีรายละเอียดขั้นตอนการทดสอบดังต่อไปนี้

4.2.1.1 การเลือกคุณลักษณะของข้อมูล

กระบวนการเลือกคุณลักษณะข้อมูลสำหรับการทดสอบแบบจำลองนี้ จะวิเคราะห์คำคุณลักษณะที่ได้มาจากคลังคำที่ถูกเลือกเพื่อเป็นตัวแทนข้อความ จากกระบวนการเรียนรู้ ซึ่งงานวิจัยนี้ได้ใช้วิธีเลือกคุณลักษณะสำคัญแบบเบื้องต้นด้วยวิธีการตัดคำที่ไม่มีนัยสำคัญออก ได้แก่ คำที่มีหน้าที่ของคำดังต่อไปนี้คือ คำหยุด, คำบุพบท, คำสันธาน, คำสรรพนาม, คำลักษณะนาม และตัวเลข ตัวอย่างคำที่ไม่มีนัยสำคัญ ดังภาคผนวก 1 และตัวอย่างข้อความที่ผ่านการเลือกคุณลักษณะในเบื้องต้น ดังตารางที่ 4.1

ตารางที่ 4.1 ตัวอย่างข้อความที่ผ่านการเลือกคุณลักษณะในเบื้องต้น

	ข้อความก่อนเลือกคุณลักษณะ	ข้อความหลังเลือกคุณลักษณะ
s_1	ชอบ มาก ละคร เรื่อง บ้า บ้า ย่า หย่า ขอ ให้ เพิ่ม วัน กับ เวลา ฉาย ไม่ ได้ หรือ อะ เพิ่ม เป็น จันทร์ ถึง ศุกร์ เลย ก็ ดี อะ ถ้า เป็น ไป ได้	ชอบ มาก ละคร เรื่อง บ้า บ้า ย่า หย่า ขอ ให้ เพิ่ม วัน เวลา ฉาย ไม่ ได้ หรือ เพิ่ม เป็น จันทร์ ศุกร์ เลย ก็ ดี ถ้า เป็น ไป ได้
s_2	ทำไม ละคร ต่าง ๆ ยัง ไม่ นำ ออก มา ฉาย <u>เช่น</u> Tea House <u>และ</u> Code Blue มี คน ที่ ติดตาม ชม อยู่ มาก มา อยาก ให้ นำ มา ลง เร็ว ๆ	ทำไม ละคร ไม่ นำ ออก มา ฉาย Tea House Code Blue มี คน ติดตาม ชม มาก มา อยาก ให้ นำ มา ลง เร็ว ๆ
s_3	ชอบ รายการ พื้นที่ ชีวิต มาก อะ น่าจะ ออก อากาศ เวลา หัว ค่ำ กว่า เดิม	ชอบ รายการ พื้นที่ ชีวิต มาก น่าจะ ออก อากาศ เวลา หัว ค่ำ กว่า เดิม

4.2.1.2 กระบวนการกลั่นกรองข้อความ

เนื่องจากข้อมูลที่นำมาทดสอบมีบางส่วนของข้อมูลที่ยังไม่สมบูรณ์ไม่สามารถนำมาใช้ในการจำแนกได้ เช่น คำที่เขียนผิด, คำที่มีความหมายเดียวกันแต่ใช้คำต่างกัน ผู้วิจัยได้แก้ไขคำให้ถูกต้องหรือเปลี่ยนแปลงคำให้เป็นคำเดียวกัน เพื่อความถูกต้องและเหมาะสมก่อนการนำไปวิเคราะห์

4.2.1.3 การแทนข้อความ

กระบวนการแทนข้อความประกอบด้วย 2 กระบวนการย่อยคือ การแทนข้อความด้วยฐานความรู้ทางภาษา (Knowledge based tagging) จากนั้นจึงแทนข้อความด้วยค่า TF-IDF

1) การแทนข้อความด้วยฐานความรู้ทางภาษา

ในงานวิจัยนี้ได้นำเสนอการเพิ่มฐานความรู้ทางภาษา ประกอบด้วย 5 วิธีดังตารางที่ 4.2 ซึ่งแสดงตัวอย่างการแทนข้อความด้วยฐานความรู้ทางภาษา แต่ละวิธีมีรายละเอียดดังนี้

- วิธีที่ 1 การแทนข้อความด้วยคำที่พบในคลังบทวิจารณ์

- วิธีที่ 2 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำ

ด้วยคลังคำเล็กซิตรอน

- วิธีที่ 3 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำ ด้วยคลังคำเล็กซิตรอน ส่วนคำที่ทำหน้าที่เป็นกริยาจะทำการกำกับคำกริยาใหม่ ด้วยหน้าที่ของ คำกริยาแบบเฉพาะเจาะจงจากคลังคำไทยออร์คิด

- วิธีที่ 4 การแทนข้อความด้วยคำ ร่วมกับการกำกับคู่ของคำที่เกิดขึ้น ร่วมกันบ่อย (AW) โดยกำหนดระยะห่างระหว่างคำ ไม่เกิน k คำ ซึ่งทดสอบค่า k ตั้งแต่ 3 ถึง 6 คำ

- วิธีที่ 5 การแทนข้อความด้วยคำ ร่วมกับคู่ของคำ (AW) ที่มีระยะห่าง เหมาะสมที่สุดร่วมกับคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW)

ตารางที่ 4.2 ตัวอย่างการแทนข้อความด้วยฐานความรู้ทางภาษา

<u>วิธีที่ 1</u> วิเคราะห์ด้วยคำ	
s_1	ชอบ มาก ละคร เรื่อง บ้านป่า ย่าหย่า ขอ ให้ เพิ่ม วัน เวลา ฉาย ไม่ ได้ เธอ เพิ่ม เป็น จันทร์ ศุกร์ เลย ก็ ดี ถ้า เป็น ไป ได้
s_2	ทำไม ละคร ไม่ นำ ออก มา ฉาย Tea House Code Blue มี คน ติดตาม ชม มากมาย อยาก ให้ นำ มา ลง เร็ว ๆ
s_3	ชอบ รายการ พื้นที่ ชีวิต มาก น่าจะ ออก อากาศ เวลา หัว คำ กว่า เดิม
<u>วิธีที่ 2</u> วิเคราะห์ด้วยฐานความรู้คำร่วมกับหน้าที่ของคำ (POS)	
s_1	ชอบ<V>มาก<ADV>ละคร<OBJ1>เรื่อง<N>บ้านป่า ย่าหย่า<OBJ2> ขอ<S _c >ให้ <AUX>เพิ่ม<V>วัน<N>เวลา<N>ฉาย<V>ไม่<NEG>ได้<AUX>เธอ<QUES> เพิ่ม <V>เป็น<V>จันทร์<N>ศุกร์<N>เลย<ADV>ก็<ADV>ดี<ADJ> ถ้า<S _c >เป็น ไปได้ <V>
s_2	ทำไม<S _q >ละคร<OBJ1>ไม่<NEG>นำ<V>ออก<V>มา<V>ฉาย<V> Tea House<OBJ2> Code Blue<OBJ2>มี<V>คน<N>ติดตาม<V> ชม<V>มากมาย <ADV>อยาก<S _c >ให้<AUX>นำ<V>มา<V>ลง<V>เร็ว ๆ <ADV>
s_3	ชอบ<V>รายการ<N>พื้นที่<N>ชีวิต<OBJ2>มาก<ADV>น่าจะ<S _q >ออกอากาศ<V>เวลา <N>หัว<N>คำ<N>กว่า<ADV>เดิม<ADV>

ตารางที่ 4.2 (ต่อ)

วิธีที่ 3 วิเคราะห์ด้วยฐานความรู้คำร่วมกับหน้าที่ของคำและกริยาแบบเฉพาะเจาะจง (Specific POS)

s ₁	<p>ชอบ<VACT>มาก<ADV>ละคร<OBJ1>เรื่อง<N>ป่าป่า ย่ำเหย้า<OBJ2> ขอ<S_c>ให้ <AUX>เพิ่ม<VACT>วัน<N>เวลา<N>ฉาย<VACT>ไม่<NEG>ได้<AUX>เหรอ <QUES> เพิ่ม<VACT>เป็น<VSTA>จันทร์<N>ศุกร์<N>เลย<ADV>ก็<ADV>ดี <ADJ>ถ้า<S_c>เป็นไปได้<VACT></p>
s ₂	<p>ทำไม<S_q>ละคร<OBJ1>ไม่<NEG>นำ<VACT>ออก<VACT>มา<VACT>ฉาย <VACT> Tea House<OBJ2> Code Blue<OBJ2>มี<VACT>คน<N>ติดตาม<VACT> ชม<VACT>มากมาย<ADV> อยาก<S_c>ให้<AUX>นำ<VACT>มา<VACT>ลง <VACT>เร็ว ๆ <ADV></p>
s ₃	<p>ชอบ<VACT>รายการ<N>พื้นที่ชีวิต<OBJ2>มาก<ADV>น่าจะ<S_q> ออกอากาศ <VACT> เวลา<N>หวัค้ำ<N> กว่า<ADV> เดิม<ADV></p>

วิธีที่ 4 วิเคราะห์ด้วยฐานความรู้คำร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (AW) โดย
 ทดลองหาระยะห่างตั้งแต่ 3 คำจนถึง 6 คำ

s ₁	<p>ชอบ<VACT>มาก<ADV>ละคร<OBJ1>เรื่อง<N>ป่าป่า ย่ำเหย้า<OBJ2> ขอ<S_c>ให้ <AUX>เพิ่ม<AW-VACT>วัน<AW-N>เวลา<AW-N>ฉาย<AW-VACT>ไม่<NEG> ไปได้<AUX>เหรอ<QUES> เพิ่ม<VACT>เป็น<VSTA>จันทร์<N>ศุกร์<N>เลย<ADV> ก็<ADV>ดี<ADJ> ถ้า<S_c>เป็นไปได้<VACT></p>
s ₂	<p>ทำไม<S_q>ละคร<OBJ1>ไม่<NEG>นำ<AW-VACT>ออก<VACT>มา<VACT>ฉาย <AW-VACT> Tea House<OBJ2>Code Blue<OBJ2>มี<VACT>คน<N>ติดตาม <VACT>ชม<VACT>มากมาย<ADV> อยาก<S_c>ให้<AUX>นำ<VACT>มา <VACT>ลง<VACT> เร็ว ๆ <ADV></p>
s ₃	<p>ชอบ<VACT>รายการ<N>พื้นที่ชีวิต<OBJ2>มาก<ADV>น่าจะ<S_q> ออกอากาศ <AW-VACT> เวลา<AW-N>หวัค้ำ<N> กว่า<ADV> เดิม<ADV></p>

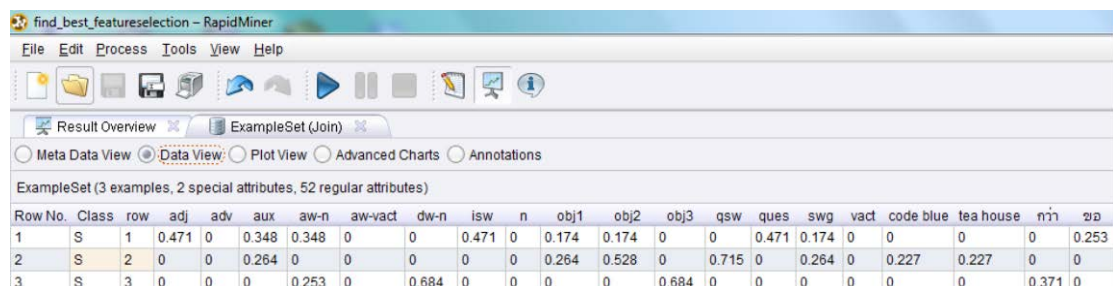
ตารางที่ 4.2 (ต่อ)

วิธีที่ 5 วิเคราะห์ด้วยฐานความรู้คำและคู่ของคำ (AW) ที่มีระยะห่างเหมาะสมที่สุดร่วมกับคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW)

s_1	ชอบ<VACT>มาก<ADV>ละคร<OBJ1>เรื่อง<N>ป่าป่า ย่ำเหย้า<OBJ2> ขอ<S _c >ให้ <AUX>เพิ่ม<AW-VACT>วัน<AW-N>เวลา<AW-N>ฉาย<AW-VACT>ไม่<NEG> ได้<AUX>เธอ<QUES> เพิ่ม<VACT>เป็น<VSTA>จันทร์<N>ศุกร์<N>เลข<ADV> ก็<ADV>ดี<ADJ> ถ้า<S _c >เป็นไปได้<VACT>
s_2	ทำไม<S _c >ละคร<OBJ1>ไม่<NEG>นำ<AW-VACT>ออก<VACT>มา<VACT>ฉาย <AW-VACT> Tea House<OBJ2>Code Blue<OBJ2>มี<VACT>คน<N>ติดตาม <VACT>ชม<VACT>มากมาย<ADV> อยาก<SWG>ให้<AUX>นำ<VACT>มา <VACT>ลง<VACT>เร็ว ๆ <ADV>
s_3	ชอบ<VACT>รายการ<N>พื้นที่ชีวิต<OBJ2>มาก<ADV>น่าจะ<S _c > ออกอากาศ <AW-VACT> เวลา<AW-N>หวั่น<DW-N>กว่า<ADV> เดิม<ADV>

2) การแทนข้อความด้วยค่า TF-IDF

การแทนข้อความด้วยค่า TF-IDF เป็นการแทนข้อความให้อยู่ในรูปเวกเตอร์สเปซโมเดล (Vector Space Model: VSM) ด้วยค่าความถี่ของคำและค่าส่วนกลับความถี่เอกสาร ที่เกิดคำ ดังตารางที่ 4.3 และตัวอย่างการแทนข้อความด้วยค่า TF-IDF ด้วยโปรแกรม Rapid miner ดังภาพที่ 4.3



Row No.	Class	row	adj	adv	aux	aw-n	aw-vact	dw-n	isw	n	obj1	obj2	obj3	qsw	ques	swg	vact	code blue	tea house	กว่า	ขอ	
1	S	1	0.471	0	0.348	0.348	0	0	0.471	0	0.174	0.174	0	0	0.471	0.174	0	0	0	0	0	0.253
2	S	2	0	0	0.264	0	0	0	0	0	0.264	0.528	0	0.715	0	0.264	0	0.227	0.227	0	0	0
3	S	3	0	0	0	0.253	0	0.684	0	0	0	0	0.684	0	0	0	0	0	0	0	0.371	0

ภาพที่ 4.3 การแทนข้อความด้วยค่า TF-IDF ด้วยโปรแกรม Rapid miner

ตารางที่ 4.3 การแทนข้อความด้วยค่า TF-IDF ของคำ (t) และหน้าที่ของคำ (p)

	Selected words				Selected POS			
	t_1	t_2	...	t_m	p_1	p_2	...	p_i
s_1	$w_{1,1}$	$w_{2,1}$...	$w_{m,1}$	$w_{p1,1}$	$w_{p2,1}$...	$w_{pi,1}$
s_2	$w_{1,2}$	$w_{2,2}$...	$w_{m,2}$	$w_{p1,2}$	$w_{p2,2}$...	$w_{pi,2}$
...
s_n	$w_{1,n}$	$w_{2,n}$...	$w_{m,n}$	$w_{p1,n}$	$w_{p2,n}$...	$w_{pi,n}$

4.2.2 กระบวนการจำแนกข้อความ

กระบวนการจำแนกข้อความหรือกระบวนการวิเคราะห์ข้อเสนอแนะแบ่งออกเป็น 3 กระบวนการย่อย ดังนี้

4.2.2.1 การสกัดข้อเสนอแนะ

ในกระบวนการสกัดข้อเสนอแนะนี้จะนำข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล และถูกแทนข้อความด้วยความรู้ทางภาษาแบบวิธีที่ 1 มาเข้าสู่กระบวนการเรียนรู้เพื่อสกัดข้อเสนอแนะและสร้างแบบจำลองการแยกข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่น โดยเรียนรู้รูปแบบของข้อมูลจากชุดข้อมูลเรียนรู้ และสร้างแบบจำลองการสกัดข้อเสนอแนะ 3 แบบจำลอง ได้แก่ (1) แบบจำลองที่สร้างจากอัลกอริทึมต้นไม้ตัดสินใจ (2) แบบจำลองที่สร้างจากอัลกอริทึมเนอิว์โรน และ (3) แบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน จากนั้นนำชุดข้อมูลทดสอบ มาจำแนกหาข้อเสนอแนะด้วยแบบจำลองที่สร้างได้ 3 แบบจำลอง นำผลลัพธ์ที่ได้จากแต่ละแบบจำลอง (อัลกอริทึม) มาเปรียบเทียบประสิทธิภาพการทำนายผล เพื่อให้ได้อัลกอริทึมการสกัดข้อเสนอแนะที่เหมาะสมที่สุด

แบบจำลองที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจคือ แต่ละกิ่งของต้นไม้จะถูกแทนด้วยคำหรือคุณลักษณะที่ถูกเลือกเพื่อเป็นตัวแทนเอกสาร (Feature) ซึ่งโหนดปลายทางจะมี 2 คลาสคือ เป็นข้อเสนอแนะกับไม่เป็นข้อเสนอแนะ

แบบจำลองที่ได้จากอัลกอริทึมเนอิว์โรนคือ ผลรวมของค่าความน่าจะเป็นมีเงื่อนไขของแต่ละคำหรือคุณลักษณะที่พบในเอกสารและคลาสข้อเสนอแนะและผลรวมของค่า

ความน่าจะเป็นของคลาสใดมีค่ามากกว่า จะถือว่าเอกสารจัดอยู่ในคลาสนั้น

แบบจำลองที่ได้จากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนคือ ค่าน้ำหนัก, w ของแต่ละคำหรือคุณลักษณะที่ถูกเลือกเพื่อเป็นตัวแทนเอกสารและค่าโน้มนำ, b ที่จะนำมาใช้คำนวณหา ค่า y ในสมการ

$$y = \begin{cases} +1, \bar{w} * \bar{x} + b > 0 \\ -1, \bar{w} * \bar{x} + b < 0 \end{cases} \quad (24)$$

โดยที่คือพีเจอร์เวกเตอร์ของข้อมูลที่ได้จากแทนข้อความด้วยค่า TF-IDF ในกระบวนการเตรียมข้อมูล หากผลลัพธ์ y ที่ได้จากการคำนวณมากกว่า 0 จัดอยู่ในคลาสข้อเสนอแนะ และหากน้อยกว่า 0 จะถูกจัดอยู่ในคลาสไม่เป็นข้อเสนอแนะ

อัลกอริทึมใดในกระบวนการการสกัดข้อเสนอแนะด้วยวิธีการแทนข้อความด้วยฐานความรู้ทางภาษาแบบวิธีที่ 1 (วิเคราะห์จากคำ) ที่มีประสิทธิภาพดีที่สุด จะนำอัลกอริทึมดังกล่าวมาใช้ในการสกัดข้อเสนอแนะด้วยวิธีการแทนข้อความด้วยฐานความรู้ทางภาษาแบบวิธีที่ 2-5 เพื่อให้ได้วิธีการแทนข้อความด้วยฐานความรู้ทางภาษาและอัลกอริทึมที่เหมาะสมกับการทำเหมืองข้อเสนอแนะที่สุด

ดังนั้นข้อมูลนำเข้าที่นำมาทดสอบเพื่อหาวิธีการแทนข้อความด้วยฐานความรู้ทางภาษา ประกอบด้วย

- วิธีที่ 1 การแทนข้อความด้วยคำที่พบในคลังข้อความ
- วิธีที่ 2 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำ ด้วยคลังคำเล็กชิตรอน
- วิธีที่ 3 การแทนข้อความด้วยคำ ร่วมกับการกำกับคำตามหน้าที่ของคำ ด้วยคลังคำเล็กชิตรอน ส่วนคำที่ทำหน้าที่เป็นกริยาจะทำการกำกับคำกริยาใหม่ ด้วยหน้าที่ของคำกริยาแบบเฉพาะเจาะจง ด้วยคลังคำไทยออร์คิด
- วิธีที่ 4 การแทนข้อความด้วยคำ ร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (AW) โดยกำหนดระยะห่างระหว่างคำ ไม่เกิน k คำ ซึ่งทดสอบค่า k ตั้งแต่ 3 ถึง 6 คำ
- วิธีที่ 5 การแทนข้อความด้วยคำ ร่วมกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (AW) ที่มีระยะห่างเหมาะสมที่สุด และคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW)

หลังจากที่แทนข้อความด้วยฐานความรู้ทางภาษาแล้ว นำชุดข้อมูลทดสอบไปทดสอบกับแบบจำลองการสกัดข้อเสนอแนะ เพื่อให้ได้วิธีการแทนข้อความหรืออินพุตเวกเตอร์ที่

เหมาะสมกับการสกัดข้อเสนอแนะ จากนั้นนำอินพุตเวกเตอร์ดังกล่าวไปวิเคราะห์เพื่อเพิ่มประสิทธิภาพโดยการปรับค่าพารามิเตอร์และเลือกใช้คอร์เนลต่าง ๆ ที่เหมาะสมกับชุดข้อมูลหรืออินพุตเวกเตอร์

4.2.2.2 การจำแนกประเภทข้อเสนอแนะ

กระบวนการจำแนกประเภทข้อเสนอแนะมีวัตถุประสงค์เพื่อจำแนกข้อเสนอออกเป็น 3 ประเภท ได้แก่ ข้อเสนอแนะทางตรง (S_c) ข้อเสนอแนะเชิงคำถาม (S_q) และข้อเสนอแนะเชิงเงื่อนไข (S_o) ซึ่งกระบวนการจำแนกประเภทข้อเสนอแนะนี้จะนำข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูลและแทนข้อความด้วยความรู้ทางภาษา ด้วยวิธีที่มีประสิทธิภาพการสกัดข้อเสนอแนะในกระบวนการก่อนหน้า (หัวข้อ 4.2.2.1) มาเข้าสู่กระบวนการจำแนกประเภทข้อเสนอแนะ เพื่อสร้างแบบจำลอง โดยเรียนรู้รูปแบบของข้อมูลจากชุดข้อมูลเรียนรู้ ด้วยอัลกอริทึมที่ได้จากกระบวนการก่อนหน้า จากนั้นนำชุดข้อมูลทดสอบ มาจำแนกหาประเภทของข้อเสนอแนะด้วยแบบจำลองที่สร้างได้

4.2.2.3 การสกัดวลีข้อเสนอแนะ

กระบวนการสกัดวลีข้อเสนอแนะเป็นกระบวนการสกัดหาส่วนประกอบของประโยคข้อเสนอแนะ โดยทำการวิเคราะห์รูปแบบประโยคในคลังบทวิจารณ์ที่ได้ทำการสกัดข้อเสนอแนะและจำแนกประเภทข้อเสนอแนะ และมีการกำกับหน้าที่ของคำไว้แล้ว เปรียบเทียบกับรูปแบบประโยคข้อเสนอแนะในคลังฐานความรู้ทางภาษาที่ได้จากผู้เชี่ยวชาญ หากรูปแบบของประโยคในคลังบทวิจารณ์เหมือนกับรูปแบบประโยคในคลังฐานความรู้ จะสามารถแสดงส่วนประกอบของประโยคข้อเสนอแนะออกมาได้

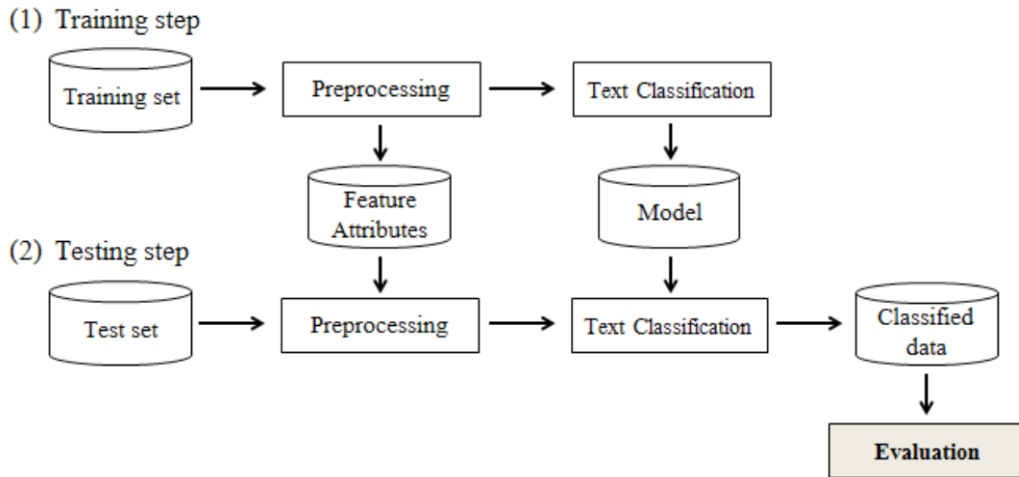
4.3 การวัดประสิทธิภาพของกระบวนการ

จากการทดสอบด้วยชุดข้อมูลทดสอบ จำนวนทั้งสิ้น 815 ข้อความ ด้วยแบบจำลองที่สร้างมาจากชุดข้อมูลเรียนรู้ ดังภาพที่ 4.4 นำมาวัดประสิทธิภาพของการจำแนกข้อเสนอแนะด้วยกระบวนการที่นำเสนอ เปรียบเทียบกับการจำแนกข้อความด้วยผู้เชี่ยวชาญ สามารถสรุปผลการวัดประสิทธิภาพของกระบวนการ ได้ดังนี้

4.3.1 การวัดประสิทธิภาพของกระบวนการสกัดข้อเสนอแนะ

กระบวนการวัดประสิทธิภาพของการสกัดข้อเสนอแนะ ประกอบด้วย 3 ขั้นตอนย่อย ได้แก่ (1) การวัดประสิทธิภาพของอัลกอริทึมค้นไม้ตัดสินใจ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์

แมชชีน เพื่อหาว่าอัลกอริทึมใดเหมาะสมกับการวิเคราะห์รูปแบบของประโยคข้อเสนอแนะ



ภาพที่ 4.4 กระบวนการวัดประสิทธิภาพงานวิจัย

มากที่สุด โดยวิเคราะห์จากการใช้คำเพียงอย่างเดียวเป็นอินพุตเวกเตอร์ของระบบ (2) การวัดประสิทธิภาพของรูปแบบอินพุตเวกเตอร์ต่างๆ ที่เหมาะสมที่สุดในการเป็นตัวแทนที่จะใช้วิเคราะห์ประโยคข้อเสนอแนะ และ (3) กระบวนการปรับค่าพารามิเตอร์ (Parameter tuning) เพื่อให้อัลกอริทึมที่นำมาวิเคราะห์มีประสิทธิภาพสูงสุด

4.3.1.1 การวัดประสิทธิภาพการสกัดข้อเสนอแนะของอัลกอริทึมต้นไม้ตัดสินใจ นาอ์ฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ผลลัพธ์ที่ได้จากกระบวนการสกัดข้อเสนอแนะของแต่ละอัลกอริทึม ดังตารางที่ 4.4

เมื่อเปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะอัลกอริทึมต้นไม้ตัดสินใจ นาอ์ฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนมี

ตารางที่ 4.4 เปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ

Classifier	Precision	Recall	F-measure
Decision Tree	80.15	48.47	60.41
Naïve Bayes	70.02	72.70	71.33
SVM	77.63	86.73	81.93

ประสิทธิภาพการสกัดข้อเสนอแนะที่ดีที่สุด ซึ่งมีค่าความแม่นยำเท่ากับ 77.98% ค่าระลอกเท่ากับ 86.73% และค่าถ่วงดุล เท่ากับ 81.98% อัลกอริทึมที่มีประสิทธิภาพรองลงมาคือ อัลกอริทึมเนอิวเบีย และต้นไม้ตัดสินใจ ตามลำดับ (อัลกอริทึมต้นไม้ตัดสินใจมักนำมาทดสอบร่วมด้วย เพื่อนำผลลัพธ์ที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจมาเป็นค่าประสิทธิภาพพื้นฐาน ในการเปรียบเทียบผลลัพธ์กับอัลกอริทึมอื่น)

ผลลัพธ์ที่ได้จากการทดลองสอดคล้องกับผลการทดลองของงานวิจัยในอดีตที่ผ่านมา (Thorsten, 1998; Yiming Yang et al., 1999; Basu et al., 2002) ที่สรุปไว้ว่าซัพพอร์ตเวกเตอร์แมชชีนเป็นอัลกอริทึมที่เหมาะสมสำหรับการวิเคราะห์เอกสารประเภทข้อความ (Text classification) เนื่องจากซัพพอร์ตเวกเตอร์แมชชีนมีกระบวนการแบ่งกลุ่มข้อมูลด้วยเส้นไฮเปอร์เพลน ซึ่งไม่ขึ้นอยู่กับมิติของข้อความนำเข้าแต่จะขึ้นอยู่กับซัพพอร์ตเวกเตอร์ที่ใช้วิเคราะห์เท่านั้น ทำให้ซัพพอร์ตเวกเตอร์แมชชีนมีกระบวนการในการเรียนรู้ที่ไม่ซับซ้อนเท่าอัลกอริทึมอื่น ๆ และสามารถลดความไม่ยืดหยุ่นเพื่อรองรับกับข้อมูลชุดใหม่ (Overfitting) ได้ นอกจากนี้ซัพพอร์ตเวกเตอร์แมชชีนจะมีเส้นไฮเปอร์เพลนแบ่งชุดข้อมูลด้วยเส้นตรงแล้ว (Linear separate line) ยังสามารถแบ่งชุดข้อมูลที่ไม่สามารถแบ่งได้ด้วยเส้นตรงได้ โดยเลือกใช้เคอร์เนลต่าง ๆ ที่เหมาะสมได้ ทำให้ซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการจำแนกข้อความมากที่สุด

4.3.1.2 การวัดประสิทธิภาพของรูปแบบอินพุตเวกเตอร์ต่าง ๆ ที่เหมาะสมที่สุดในการเป็นตัวแทนที่จะใช้วิเคราะห์ประโยคข้อเสนอแนะ

เมื่อได้อัลกอริทึมการสกัดข้อเสนอแนะที่เหมาะสมที่สุดแล้ว คือ อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน จากนั้นนำข้อเสนอแนะมาแทนข้อความด้วยฐานความรู้ทางภาษาให้มีรูปแบบข้อมูลอินพุตที่แตกต่างกัน เพื่อพิจารณาว่าข้อมูลอินพุตรูปแบบใดเป็นรูปแบบที่ให้ประสิทธิภาพการสกัดข้อเสนอแนะที่ดีที่สุด ซึ่งประกอบด้วยข้อมูลอินพุตที่ใช้วิธีการแทนข้อความแตกต่างกัน 5 วิธี ดังนี้

วิธีที่ 1 วิธีการแทนข้อความด้วยคำ

วิธีที่ 2 วิธีการแทนข้อความด้วยคำร่วมกับหน้าที่ของคำ

วิธีที่ 3 วิธีการแทนข้อความด้วยคำร่วมกับหน้าที่ของคำและกริยาแบบ

เฉพาะเจาะจง

วิธีที่ 4 วิธีการแทนข้อความด้วยคำร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อยโดยทดลองหาระยะห่างตั้งแต่ 3 คำจนถึง 6 คำ

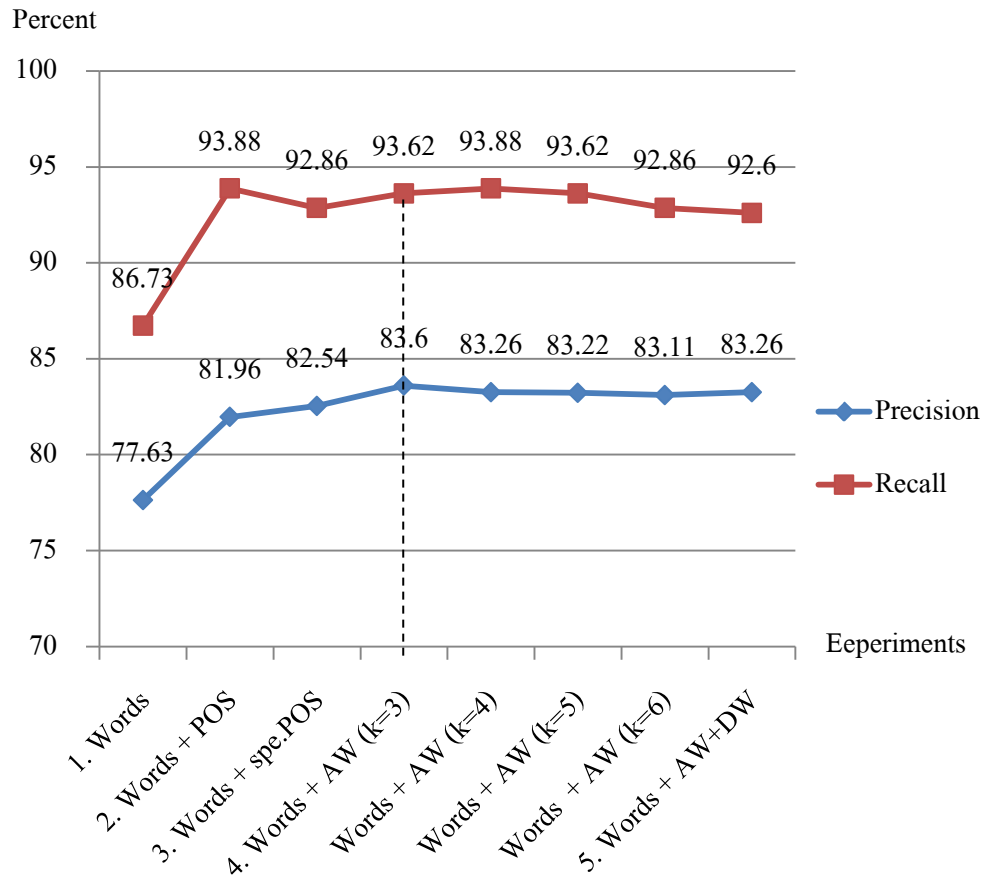
วิธีที่ 5 วิธีการแทนข้อความด้วยคำและคู่ของคำที่เกิดขึ้นร่วมกันบ่อยที่มีระยะห่างเหมาะสมที่สุดร่วมกับคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน

ผลลัพธ์ของสัปดาห์ข้อเสนอแนะด้วยวิธีการแทนข้อความทั้ง 5 วิธี ด้วยอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีนแบบเส้นตรง แสดงตารางที่ 4.5 และภาพแสดงการเปรียบเทียบประสิทธิภาพการสัปดาห์ข้อเสนอแนะ ดังภาพที่ 4.5

เมื่อเปรียบเทียบประสิทธิภาพของกระบวนการจำแนกข้อเสนอแนะ ด้วยการแทนข้อความทั้ง 5 วิธี ร่วมกับการจำแนกข้อเสนอแนะด้วยอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีน พบว่าการจำแนกข้อเสนอแนะ ด้วยวิธีการแทนข้อความด้วยคำร่วมกับการกำกับฐานความรู้ทางภาษาคู่ของคำที่เกิดขึ้นร่วมกันบ่อย ในวิธีที่ 4 ที่มีระยะห่างระหว่างคำ (k) ไม่เกิน 3 คำ ให้ประสิทธิภาพการสัปดาห์ข้อเสนอแนะที่ดีที่สุด เนื่องจากคำที่มีระยะห่างระหว่างกันเกิน 3 คำ จะมีความสัมพันธ์กันน้อยลงหรือไม่มีความสัมพันธ์กัน

ตารางที่ 4.5 เปรียบเทียบประสิทธิภาพการสัปดาห์ข้อเสนอแนะ ด้วยอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีนแบบเส้นตรง ด้วยวิธีการทดสอบการแทนข้อความ 5 วิธี

Experiment	Precision	Recall	F-measure
1. Words	77.63	86.73	81.93
Words + ...			
2. + POS	81.96	93.88	87.52
3. + spe.POS	82.54	92.86	87.40
4. + AW ($k=3$)	83.60	93.62	88.33
+ AW ($k=4$)	83.26	93.88	88.25
+ AW ($k=5$)	83.22	93.62	88.11
+ AW ($k=6$)	83.11	92.86	87.71
5. + AW+DW	83.26	92.60	87.68



ภาพที่ 4.5 ภาพแสดงการเปรียบเทียบประสิทธิภาพการสกัดข้อเสนอแนะ

4.3.2 การปรับค่าพารามิเตอร์

เมื่อได้อินพุตเวกเตอร์ที่เหมาะสมในการวิเคราะห์รูปแบบของประโยคข้อเสนอแนะ (ซึ่งหมายถึงวิธีการสร้างอินพุตเวกเตอร์แบบวิธีที่ 4) จะทำการเลือกใช้เคอร์เนลซัพพอร์ตเวกเตอร์ (SVM kernel) และปรับค่าพารามิเตอร์แต่ละเคอร์เนลให้เหมาะสม เพื่อพิจารณาว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนที่เคอร์เนลใดมีประสิทธิภาพการสกัดข้อเสนอแนะที่ดีที่สุด ในงานวิจัยนี้ได้เลือกใช้ 3 รูปแบบ ได้แก่ Linear SVM, ฟังก์ชันเคอร์เนล Radial basic และฟังก์ชันเคอร์เนล Polynomial จากนั้นจะทำการปรับค่าพารามิเตอร์ของแต่ละฟังก์ชัน ซึ่งมีรายละเอียดดังนี้

Linear SVM เป็นฟังก์ชันที่สามารถปรับค่าพารามิเตอร์ง่ายที่สุด เนื่องจากมีค่าพารามิเตอร์ C ค่าเดียวที่ใช้ในการปรับ ผลลัพธ์ดังตารางที่ 4.6

SVM ฟังก์ชันเคอร์เนล Radial basic มีการปรับค่าพารามิเตอร์ 2 ตัว ได้แก่ พารามิเตอร์ C และ gamma ผลลัพธ์ดังตารางที่ 4.7

ตารางที่ 4.6 เปรียบเทียบประสิทธิภาพการสกดข้อเสนอนะ ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์
แมชชีนแบบเส้นตรง เมื่อมีการปรับค่าพารามิเตอร์ C ที่แตกต่างกัน

Linear SVM	Precision	Recall	F-measure
C = 0	83.60%	93.62%	88.33%
C = 1	83.71%	94.39%	88.73%
C = 2	82.59%	94.39%	87.46%
C = 10	84.41%	89.80%	87.02%
C = 100	84.41%	89.80%	87.02%

ตารางที่ 4.7 เปรียบเทียบประสิทธิภาพการสกดข้อเสนอนะ ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์
แมชชีนแบบฟังก์ชันเคอร์เนล Raidial basic เมื่อมีการปรับค่าพารามิเตอร์ C และ
gamma ที่แตกต่างกัน

Linear SVM	Precision	Recall	F-measure
C = 1, gamma = 0.1	81.55%	85.71%	83.58%
C = 1, gamma = 0.5	84.86%	90.05%	87.38%
C = 1, gamma = 1	86.57%	88.78%	87.66%
C = 1, gamma = 2	84.08%	89.50%	86.71%

ฟังก์ชันเคอร์เนล Polynomial การปรับค่าพารามิเตอร์ 3 ตัวได้แก่ พารามิเตอร์ C, gamma และ degree ผลลัพธ์ดังตารางที่ 4.8 หลังจากปรับค่าพารามิเตอร์แล้ว พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน แบบเคอร์เนล Polynomial ที่ค่าพารามิเตอร์ C เท่ากับ 1, gamma เท่ากับ 1 และ degree เท่ากับ 1 มีประสิทธิภาพการสกดข้อเสนอนะที่ดีที่สุด ดังตารางที่ 4.9

ตารางที่ 4.8 เปรียบเทียบประสิทธิภาพการสัดข้อเสนอแนะ ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์
แมชชีนแบบฟังก์ชันเคอร์เนล Polynomial เมื่อมีการปรับค่าพารามิเตอร์ C, gamma
และ degree ที่แตกต่างกัน

Linear SVM	Precision	Recall	F-measure
C = 1, gamma = 1, degree = 1	85.75%	93.62%	89.51%
C = 1, gamma = 1, degree = 2	87.14%	91.58%	89.30%
C = 1, gamma = 1, degree = 3	87.19%	88.52%	87.85%
C = 1, gamma = 1, degree = 4	86.11%	86.99%	86.55%
C= 2, gamma = 2, degree = 1	85.14%	92.09%	88.48%

ตารางที่ 4.9 ตารางแสดงประสิทธิภาพการสัดข้อเสนอแนะด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์
แมชชีน แบบเคอร์เนล Polynomial ที่ C = 1, gamma = 1 และ degree = 1

SVM Polynomial kernel Parameters: C = 1, gamma = 1, degree = 1	Actual Suggestion	Actual non-Suggestion	Total predict	class precision
predict Suggestion	367	61	428	85.75%
predict non-Suggestion	25	362	387	93.54%
Total actual	393	423		
class recall	93.62%	85.58%		
F - measure	89.51%			

4.3.3 การวัดประสิทธิภาพของกระบวนการจำแนกประเภทข้อเสนอแนะ

เมื่อได้วิธีการแทนข้อความด้วยคำ ร่วมกับการกำกับคู่ของคำที่เกิดขึ้นร่วมกันบ่อย ที่มีระยะห่างระหว่างกันไม่เกิน 3 คำ (วิธีที่ 4) และอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีน เคอร์เนล Polynomial ที่ค่าพารามิเตอร์ C เท่ากับ 1, gamma เท่ากับ 1 และ degree เท่ากับ 1 ที่เหมาะสมที่สุด สำหรับการจำแนกข้อเสนอแนะแล้ว กระบวนการสุดท้ายคือการจำแนกประเภทข้อเสนอแนะ ออกเป็น 3 ประเภท โดยใช้ชุดข้อมูลเรียนรู้จำนวน 392 ประโยคและชุดข้อมูลทดสอบจำนวน 395 ประโยค ผลลัพธ์ที่ได้ ดังตารางที่ 4.10

ตารางที่ 4.10 ประสิทธิภาพการจำแนกประเภทข้อเสนอแนะ

Test set	true S_c	true S_q	true S_e	Precision
pred. S_c	304	8	7	95.30%
pred. S_q	3	47	0	94.00%
pred. S_e	2	0	24	92.31%
Recall	98.38%	85.45%	77.24%	
Micro precision = 94.94%, Micro recall = 94.94%				
Macro precision = 93.87%, Macro recall = 87.08%				

กระบวนการจำแนกประเภทข้อเสนอแนะ เป็นกระบวนการจำแนกข้อความแบบหลายกลุ่ม (จำแนกออกเป็น 3 ประเภท) ดังนั้นจึงเลือกใช้วิธีการวัดประสิทธิภาพการจำแนกประเภทด้วยค่าเฉลี่ยแบบให้น้ำหนัก 2 วิธีคือ

1. การวัดประสิทธิภาพการจำแนกประเภทด้วยค่าเฉลี่ยแบบให้น้ำหนักกับทุกเอกสารเท่ากัน (Micro-averaged หรือ Document-pivoted measure) ซึ่งพบว่ามีค่า Micro precision และ Micro recall เท่ากับ 94.94% และ 94.94% ตามลำดับ

2. การวัดประสิทธิภาพการจำแนกประเภทด้วยค่าเฉลี่ยแบบให้น้ำหนักกับทุกประเภทเท่ากัน (Macro-averaged หรือ Category-pivoted measure) ซึ่งพบว่ามีค่า Macro precision และ Macro recall เท่ากับ 93.87% และ 87.08% ตามลำดับ

ประสิทธิภาพการจำแนกประเภทข้อเสนอแนะถือว่ามีความถูกต้องสูง เนื่องจากข้อความนำเข้าเป็นข้อความที่ถูกจำแนกไว้แล้วว่าเป็นข้อเสนอแนะ สามารถช่วยลดความผิดพลาดในการ

จำแนกประเภทข้อเสนอแนะลงได้ นอกจากนั้นประโยชน์ของการจำแนกประเภทข้อเสนอแนะคือ จะช่วยให้การสกัดหาวลีข้อเสนอแนะที่ซ่อนอยู่ภายในประโยคมีความถูกต้องมากยิ่งขึ้น

4.3.4 กระบวนการสกัดวลีข้อเสนอแนะ

กระบวนการสกัดวลีข้อเสนอแนะเป็นกระบวนการนำผลลัพธ์ที่ได้จากกระบวนการจำแนกประเภทข้อเสนอแนะไปสกัดหาวลีข้อเสนอแนะ โดยพิจารณาจากรูปแบบของประโยค ข้อเสนอแนะที่เกิดขึ้นบ่อยภายในคลังบทวิจารณ์ ซึ่งรูปแบบของประโยคข้อเสนอแนะจะถูกจำแนกประเภทตามที่กำหนดไว้แล้ว 3 ประเภท แต่ละประเภทประกอบด้วยรูปแบบการใช้ภาษาที่แตกต่างกัน ดังตารางที่ 4.11 และตัวอย่างประโยคข้อเสนอแนะและวลีข้อเสนอแนะที่สกัดได้จากรูปแบบของประโยคข้อเสนอแนะที่แบ่งตามประเภท 3 ประเภท ดังตารางที่ 4.12

ตารางที่ 4.11 รูปแบบของประโยคข้อเสนอแนะแบ่งตามประเภท

ประเภทข้อเสนอแนะ	รูปแบบประโยค
Explicit suggestions	$S_c + OBJ + Suggestion$
	$S_c + Suggestion + OBJ$
	$S_c + Suggestion$
	$OBJ + S_c + Suggestion$
	$S_a + Suggestion + OBJ$
	$S_a + Suggestion$
	$OBJ + S_a + Suggestion$
	$OBJ + Suggestion + S_a$
	$Suggestion + OBJ + S_a$
Query suggestions	$S_q + OBJ + Suggestion$
	$S_q + Suggestion + OBJ$
	$S_q + Suggestion$
	$OBJ + S_q + Suggestion$
	$OBJ + Suggestion + S_q$
Condition suggestions	$S_c + OBJ + Suggestion$
	$OBJ + S_c + Suggestion$

ตารางที่ 4.12 แสดงตัวอย่างประโยคข้อเสนอแนะและวลีข้อเสนอแนะที่สกัดได้จากรูปแบบของ
ประโยคข้อเสนอแนะที่แบ่งตามประเภท

Explicit suggestions

$S_c + OBJ + Suggestion$

อยากให้<S_c>คุณลานบุญ<OBJ>ปรับปรุงวิธีการอ่านข่าว<Suggestion>ครับ ทั้ง
น้ำเสียง การเว้นวรรคการทอดเสียงไม่เป็นธรรมชาติเลยฟังแล้วอึดอัด แถมใบหน้าที่
ไม่นิ่งต้องขยับคิ้ว เอียงคอ พยักหน้าฯลฯทุกพยางค์ เวียนหัวและไม่มีสมาธิในการชม
ข่าวเลยครับดีเพื่อก่อนะครับ

หัวข้อ: คุณลานบุญ

คำบ่งชี้ข้อเสนอแนะ: อยากให้

ข้อเสนอแนะ: ปรับปรุงวิธีการอ่านข่าว

$S_c + Suggestion + OBJ$

อยากให้<S_c>ผู้บริหารทบทวนบทบาทการรายงาน<Suggestion>ข่าว<OBJ>ของผู้
ประกาศข่าวภาคค่ำด้วย

หัวข้อ: ข่าว

คำบ่งชี้ข้อเสนอแนะ: อยากให้

ข้อเสนอแนะ: ผู้บริหารทบทวนบทบาทการรายงานข่าวของผู้ประกาศข่าวภาค
ค่ำด้วย

$S_c + Suggestion$

ขอติ<S_c>เรื่องภาพที่นำมาประกอบรายการควรปรับปรุงสัดส่วนภาพให้เหมาะสม
<Suggestion>

หัวข้อ: -

คำบ่งชี้ข้อเสนอแนะ: ขอติ

ข้อเสนอแนะ: เรื่องภาพที่นำมาประกอบรายการควรปรับปรุงสัดส่วนภาพให้
เหมาะสม

ตารางที่ 4.12 (ต่อ)

OBJ + S _c + Suggestion
เมื่อไรรายการ <u>หนังพาไป</u> <OBJ> จะกลับมาฉายอีกครั้ง <u>อยากให้</u> <S _c > <u>รายการนี้</u> กลับมาฉายใหม่เพราะที่บ้านชอบกันมาก<Suggestion>
หัวข้อ: หนังพาไป
คำบ่งชี้ข้อเสนอแนะ: อยากให้
ข้อเสนอแนะ: รายการนี้กลับมาฉายใหม่เพราะที่บ้านชอบกันมาก
S _a + Suggestion + OBJ
ควร<S _a > <u>แก้ไขภาพที่นำมาประกอบรายการ</u> <Suggestion> <u>พินิจนคร</u> <OBJ> เพราะ ไม่สอดคล้องกันกับบทพูด
หัวข้อ: พินิจนคร
คำบ่งชี้ข้อเสนอแนะ: ควร
ข้อเสนอแนะ: แก้ไขภาพที่นำมาประกอบรายการพินิจนครเพราะไม่สอดคล้อง กับบทพูด
S _a + Suggestion
รบกวน<S _a > <u>ตั้งค่าเสียงเริ่มต้นในคลิปรายการThaiPBSย้อนหลัง</u> ให้ลดลงเหลือสัก ครึ่งของระดับvolume เปิดปั๊บดังมาก<Suggestion>
หัวข้อ: -
คำบ่งชี้ข้อเสนอแนะ: รบกวน
ข้อเสนอแนะ: ตั้งค่าเสียงเริ่มต้นในคลิปรายการThaiPBSย้อนหลัง ให้ลดลง เหลือสักครึ่งของระดับ volume เปิดปั๊บดังมาก
OBJ + S _a + Suggestion
รายการ <u>กินอยู่...คือ</u> <OBJ> <u>น่าจะ</u> <S _a > <u>นำเสนอว่าวัตถุดิบในประเทศอะไรบ้างที่แทน</u> <u>วัตถุดิบของต่างประเทศได้</u> <Suggestion>
หัวข้อ: กินอยู่...คือ
คำบ่งชี้ข้อเสนอแนะ: น่าจะ
ข้อเสนอแนะ: นำเสนอว่าวัตถุดิบในประเทศอะไรบ้างที่แทนวัตถุดิบของ ต่างประเทศได้

ตารางที่ 4.12 (ต่อ)

OBJ+ Suggestion + S _q	
วันที่	19-11-54 รายการstudent channel<OBJ>และรายการetvทีวีเข้มonet เปิดดูแล้ว
หัวข้อ:	student channel
คำบ่งชี้ข้อเสนอแนะ:	กรณ
ข้อเสนอแนะ:	รายการetvทีวีเข้มonet เปิดดูแล้วไม่ใช่เนื้อหาของรายการเลขกรณาแก้ไขให้ด้วย
Suggestion + OBJ+ S _q	
หัวข้อ:	ThaiPBS
คำบ่งชี้ข้อเสนอแนะ:	ควรพิจารณา
ข้อเสนอแนะ:	พิธีกรภาคสนามแต่งตัวไม่ค่อยสุภาพ ผู้บริหาร ThaiPBS ควรพิจารณาด้วย
Query suggestions	
S _q + OBJ + Suggestion	
หัวข้อ:	หนัง
คำบ่งชี้ข้อเสนอแนะ:	ทำไม
ข้อเสนอแนะ:	มีแต่ตลก ๆ เอามาฉายเร็ว ๆ หน่อยไม่ได้หรือ
S _q + Suggestion + OBJ	
หัวข้อ:	ทีวีเข้มเติมเต็มความรู้
คำบ่งชี้ข้อเสนอแนะ:	ทำไม
ข้อเสนอแนะ:	คาน์โหลดเอกสารประกอบการรับชมไม่ได้เลยสักวิชา

ตารางที่ 4.12 (ต่อ)

S _q + Suggestion	
อยากให้เอา AF 10 เป็นAllStarsดีไหม<S _q >เอาแชมป์มารวมกันใหม่ ใช้ชีวิตให้ใจ	
<Suggestion>	
หัวข้อ:	-
คำบ่งชี้ข้อเสนอแนะ:	ดีไหม
ข้อเสนอแนะ:	เอาแชมป์มารวมกันใหม่ ใช้ชีวิตให้ใจ
OBJ + S _q + Suggestion	
รายการเวทีสาธารณะ<OBJ>ทำไม<S _q >ไม่มีเปิดให้แสดงความคิดเห็นสดบ้าง	
<Suggestion>	
หัวข้อ:	เวทีสาธารณะ
คำบ่งชี้ข้อเสนอแนะ:	ทำไม
ข้อเสนอแนะ:	ไม่มีเปิดให้แสดงความคิดเห็นสดบ้าง
OBJ + Suggestion + S _q	
น่าจะท้วงอีกรอบแล้วรัฐบาลมัวทำอะไรกันอยู่นักข่าวThaiPBS<OBJ>เข้าไป	
ตรวจสอบให้หน่อย<Suggestion>ได้ไหม<S _q >ค่ะ	
หัวข้อ:	ThaiPBS
คำบ่งชี้ข้อเสนอแนะ:	ได้ไหม
ข้อเสนอแนะ:	เข้าไปตรวจสอบให้หน่อย
Condition suggestions	
S _c + OBJ + Suggestion	
จะดีกว่านี่นะถ้า<S _c >เอารายการท่องโลกกว้าง<OBJ>มาใส่ซับไทเทิลภาษาไทยให้	
ด้วย<Suggestion>	
หัวข้อ:	ท่องโลกกว้าง
คำบ่งชี้ข้อเสนอแนะ:	ถ้า
ข้อเสนอแนะ:	มาใส่ซับไทเทิลภาษาไทยให้ด้วย

ตารางที่ 4.12 (ต่อ)

OBJ + S _c + Suggestion	
<hr/>	
<p>พอดีได้เห็นช่องDMC<OBJ> ทำการถ่ายทอดสดงานฉลองครองราชย์ครบ 60 ปี โดยนำสัญญาณสดมาจาก โทรทัศน์รวมการเฉพาะกิจแห่งประเทศไทย แต่ถ่ายทอดไม่ตลอด ผมคิดว่า ถ้า<S_c>วันไหนจะมีการถ่ายทอดรายการพิเศษออกโทรทัศน์รวมการเฉพาะกิจ เป็นไปได้ไหม ที่ทางช่อง DMC จะทำการขอสัญญาณถ่ายทอดสดแบบทีวีช่องอื่น ๆ ตลอดรายการพิเศษนั้น<Suggestion></p>	
หัวข้อ:	DMC
คำบ่งชี้ข้อเสนอแนะ:	ถ้า
ข้อเสนอแนะ:	วันไหนจะมีการถ่ายทอดรายการพิเศษออกโทรทัศน์รวมการเฉพาะกิจ เป็นไปได้ไหม ที่ทางช่อง DMC จะทำการขอสัญญาณถ่ายทอดสดแบบทีวีช่องอื่น ๆ ตลอดรายการพิเศษนั้น

บทที่ 5

สรุปผล และข้อเสนอแนะ

5.1 สรุปผล

การวิเคราะห์ข้อเสนอแนะจากผู้บริโภคจะช่วยให้ธุรกิจทราบว่าผู้บริโภคให้ความสำคัญกับเรื่องใด และธุรกิจควรปรับปรุงหรือพัฒนาไปในทิศทางใด เพื่อให้สอดคล้องกับความต้องการของผู้บริโภคได้มากที่สุด และสามารถตอบสนองต่อความต้องการนั้นได้อย่างรวดเร็ว ซึ่งจะส่งผลดีต่อธุรกิจในการเพิ่มความสามารถทางการแข่งขันและโอกาสความสำเร็จของธุรกิจได้มากยิ่งขึ้น

งานวิจัยนี้จึงได้นำเสนอวิธีการวิเคราะห์ข้อเสนอแนะ ด้วยเทคนิคเหมืองข้อความร่วมกับการประมวลผลภาษาธรรมชาติ ซึ่งแบ่งกระบวนการวิเคราะห์ออกเป็น 2 กระบวนการหลัก ได้แก่ (1) กระบวนการสกัดข้อเสนอแนะที่จะช่วยลดระยะเวลาการค้นหาข้อเสนอแนะโดยสกัดข้อเสนอแนะที่ถูกปะปนอยู่กับข้อมูลอื่นที่ไม่เกี่ยวข้อง (ข้อเท็จจริงและข้อคิดเห็น) ออกมาได้ ด้วยวิธีการใช้แบบจำลองการสกัดแยกข้อเสนอแนะ ซึ่งแบบจำลองดังกล่าวนี้จะสร้างมาจากชุดข้อมูลเรียนรู้ที่ทราบผลเฉลยแล้ว และเมื่อมีข้อมูลชุดใหม่ที่ยังไม่ทราบผลเฉลยจะใช้แบบจำลองที่สร้างได้นี้ในการสกัดข้อเสนอแนะออกมาจากบทวิจารณ์ ทำให้สามารถสกัดข้อเสนอแนะออกมาจากบทวิจารณ์จำนวนมากได้ โดยไม่จำเป็นต้องใช้ผู้เชี่ยวชาญในการค้นหาข้อเสนอแนะใหม่ทุกครั้ง ซึ่งสิ้นเปลืองทั้งทรัพยากรด้านเวลาและค่าใช้จ่ายจำนวนมาก และ (2) กระบวนการจำแนกประเภทข้อเสนอแนะที่จะช่วยให้ได้ข้อมูลที่ถูกจัดกลุ่มตามเจตนาการแสดงข้อเสนอแนะที่คล้ายกันเข้าไว้ด้วยกัน ทำให้ง่ายต่อการวิเคราะห์และตีความข้อเสนอแนะที่มีโครงสร้างไม่แน่นอนได้ และสามารถนำสารสนเทศที่ได้จากการกระบวนการจำแนกประเภทข้อเสนอแนะนี้ไปสกัดหาข้อข้อเสนอแนะที่ซ่อนอยู่ภายในประโยค ทำให้ได้ข้อข้อเสนอแนะที่สามารถนำไปใช้ประโยชน์เพื่อการปรับปรุงและพัฒนาธุรกิจต่อไป

5.2 การอภิปรายผล

จากการทดลองการวิเคราะห์เหมือนข้อเสนอแนะพบว่า การวิเคราะห์ข้อเสนอแนะด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน โดยวิเคราะห์คำ ร่วมกับการแทนข้อความด้วยความรู้ทางภาษา ให้อยู่ในรูปแบบของพีเจอร์เวกเตอร์ที่ประกอบด้วยคำและหน้าที่ของคำที่ได้จากการวิเคราะห์ ความสัมพันธ์ระหว่างคู่ของคำด้วยกฎความสัมพันธ์ ได้แก่ความสัมพันธ์ของคำนามกับคำกริยา และความสัมพันธ์ระหว่างคำกริยากับคำกริยาที่เกิดขึ้นร่วมกันบ่อย ซึ่งคู่ของความสัมพันธ์นั้นต้องมีระยะห่างระหว่างกันไม่เกิน 3 คำ พิจารณาดังประโยคตัวอย่าง

“อยาก<SE>ให้<AUX>หนังพาไป<OBJ2>ที่<PREP>มี<VACT>พิธีกร<N>ชาย<N>พา<VACT>เที่ยว<VACT>ประเทศ<N>ต่าง ๆ <DET>เปลี่ยน<AW-VACT> เวลา<AW-N>มา<VACT>กลางวัน<DW-N>หน่อย<ADV>”

จากฐานความรู้ทางภาษาของคำที่เกิดขึ้นร่วมกันบ่อย (AW) พบว่าคำกริยา “เปลี่ยน” มีความสัมพันธ์กับคำนาม 2 คำ คือ “พิธีกร” และ “เวลา” ดังตารางที่ 5.1

ตารางที่ 5.1 ตัวอย่างคู่ของคำนามและคำกริยา (DW)

Item 1	Item 2	Support
พิธีกร (AW-N)	เปลี่ยน (AW-VACT)	0.02
เวลา (AW-N)	เปลี่ยน (AW-VACT)	0.01

จากประโยคตัวอย่างจะเห็นว่ากริยาแสดงอาการ “เปลี่ยน” มีความสัมพันธ์กับคำนาม “เวลา” ที่อยู่ใกล้กันมากกว่าคำนาม “พิธีกร”

แต่เมื่อแทนข้อความด้วยคำเฉพาะเจาะจง โดเมนที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน (DW) ดังวิธีการแทนข้อความด้วยฐานความรู้ทางภาษาแบบวิธีที่ 5 ซึ่งได้ตั้งสมมติฐานไว้ว่าคำที่เกิดขึ้นบ่อยภายใต้โดเมนที่ใกล้เคียงกัน คือคำที่ใช้แสดงคุณลักษณะของโดเมน เช่น เว็บไซต์, สัญญาณ, ภาษา, หน้าจอ เป็นต้น การแทนข้อความด้วยการกำกับคำจากฐานความรู้ดังกล่าวไม่ทำให้ประสิทธิภาพการจำแนกข้อเสนอแนะดีขึ้น เนื่องจากคำเหล่านั้นปรากฏอยู่ในบทวิจารณ์ทั้งประเภทข้อเท็จจริงและความคิดเห็น แต่วัตถุประสงค์ของกระบวนการนี้ คือการจำแนกข้อเสนอแนะออก

จากข้อเท็จจริงและความคิดเห็น คำนึงค่าในกลุ่ม DW จึงไม่สามารถใช้เป็นค่าบังคับเพื่อจำแนกบทวิจารณ์ประเภทข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่นได้

อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเป็นอัลกอริทึมที่มีประสิทธิภาพการวิเคราะห์ข้อเสนอแนะที่ดีที่สุดเมื่อเปรียบเทียบกับอัลกอริทึมต้นไม้ตัดสินใจและเนอรัฟเบย์ นอกจากนี้ อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนยังสามารถเพิ่มประสิทธิภาพการวิเคราะห์ข้อเสนอแนะได้ด้วยวิธีการเลือกใช้ฟังก์ชันเคอร์เนลและการปรับค่าพารามิเตอร์ที่เหมาะสม จากผลการทดลองพบว่าการวิเคราะห์ข้อเสนอแนะมีประสิทธิภาพการดีที่สุด โดยการใช้อัลกอริทึมซัพพอร์ตเวกเตอร์ ฟังก์ชันเคอร์เนล Polynomial ที่ค่าพารามิเตอร์ ดังนี้

พารามิเตอร์ C เท่ากับ 1 หมายถึง ค่า C ที่เหมาะสมสำหรับการสร้างแบบจำลองการแบ่งกลุ่มข้อมูลหรือการสกัดข้อเสนอแนะ ที่มีการกำหนดให้ยอมรับความผิดพลาดที่เกิดขึ้นของชุดข้อมูลบางชุดได้ และมีระยะห่างระหว่างกลุ่ม 2 กลุ่มข้อมูลที่เหมาะสม กล่าวคือไม่เกิดปัญหาเรื่องการเข้ากันเกินไป (Overfitting) หรือการสร้างแบบจำลองที่สามารถนำมาใช้ในการเรียนรู้ได้ดี แต่ไม่สามารถนำไปใช้ทำนายข้อมูลอื่นๆ ได้หรือทำนายได้ไม่ดี

พารามิเตอร์แกมมา เท่ากับ 1 หมายถึง การปรับค่าความกว้างเพื่อให้รองรับกับซัพพอร์ตเวกเตอร์ หากกำหนดค่าแกมมาที่ 0 หมายถึงแบบจำลองหรือเส้นไฮเปอร์เพลนที่ได้ไม่สอดคล้องกับซัพพอร์ตเวกเตอร์ค่าใดเลย และหากเพิ่มค่าแกมมาให้สูงขึ้นแบบจำลองหรือเส้นไฮเปอร์เพลนที่ได้จะชนกับซัพพอร์ตเวกเตอร์จำนวนมากขึ้น ทำให้ได้แบบจำลองที่สามารถทำนายชุดข้อมูลใหม่ได้ถูกต้องยิ่งขึ้น แต่หากกำหนดค่าแกมมาสูงจนเกินไปจะเกิดปัญหาการเข้ากันมากเกินไปได้ ซึ่งจากการทดลองวิเคราะห์การสกัดข้อเสนอแนะพบว่าค่าพารามิเตอร์แกมมาเท่ากับ 1 เป็นค่าที่เหมาะสมกับชุดข้อมูลมากที่สุด

พารามิเตอร์ดีกรี เท่ากับ 1 หมายถึง การปรับค่าความโค้งของเส้นไฮเปอร์เพลนที่เหมาะสมกับชุดข้อมูล เนื่องจากชุดข้อมูลที่นำมาเรียนรู้และทดสอบไม่สามารถแบ่งได้ด้วยเส้นตรงหรือได้แต่ประสิทธิภาพการสกัดข้อเสนอแนะไม่ดีเท่าที่ควร การปรับเส้นไฮเปอร์เพลนให้มีความโค้งที่สอดคล้องกับชุดข้อมูลจะทำให้ได้เส้นแบ่งชุดข้อมูล 2 กลุ่มออกจากกันได้เหมาะสม ซึ่งได้แก่ค่าพารามิเตอร์ดีกรี เท่ากับ 1 และหากปรับเพิ่มค่าดีกรีจะได้เส้นแบ่งที่โค้งมากขึ้น ซึ่งทำให้ได้เส้นแบ่งชุดข้อมูลที่ไม่เหมาะสม นอกจากนี้แล้วยังเพิ่มค่าดีกรีจำนวนมากขึ้นจะทำให้ต้องใช้ระยะเวลาการคำนวณที่มากขึ้นตามด้วย และหากปรับค่าดีกรีที่ต่ำที่สุดก็คือการแบ่งด้วยเส้นตรงนั่นเอง

ดังนั้นการปรับค่าพารามิเตอร์ควรพิจารณาถึงความเหมาะสมของชุดข้อมูลทดสอบและชุดข้อมูลใหม่ การกำหนดค่าพารามิเตอร์ที่สูงเกินไปจะเกิดปัญหาเรื่องการเข้ากันมากเกินไปได้ และ

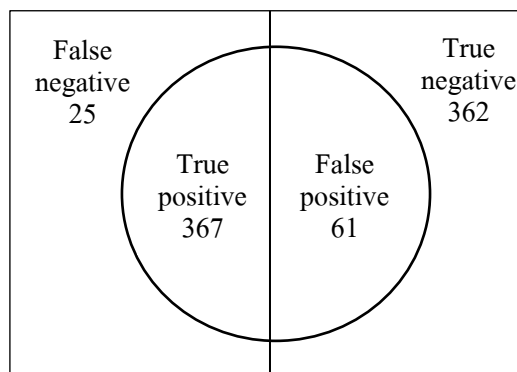
หากปรับค่าน้อยเกินไปจนไม่เหมาะสมกับข้อมูล จะทำให้ได้ผลลัพธ์ที่ไม่สามารถวิเคราะห์ข้อมูล หรือนำไปใช้ประโยชน์ได้

ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อเสนอแนะ พบว่าข้อเสนอแนะมักถูกแสดงด้วยคำกริยา แสดงอาการ เพื่อบ่งบอกถึงการกระทำของผู้อื่น หรือแนวทางให้ปฏิบัติ การวิเคราะห์ด้วยคำกริยา แสดงอาการและคำบ่งชี้วลีข้อเสนอแนะประกอบกันจะทำให้มีประสิทธิภาพการจำแนก ข้อเสนอแนะที่ดีที่สุด กล่าวคือเป็นข้อเสนอแนะที่ไม่มีความกำกวม กล่าวคือมีส่วนประกอบของ ประโยคข้อเสนอแนะที่สมบูรณ์

ประโยคข้อเสนอแนะและวลีข้อเสนอแนะมีส่วนประกอบ(หน้าที่ของคำ) ดังนี้

1. คำบ่งชี้ข้อเสนอแนะ (Suggestion indicators หรือ suggestion words: SW)
2. คำระบุหัวข้อ (Name Entities หรือ OBJ)
3. วลีเสนอแนะ หรือคู่ของคำที่เกิดขึ้นร่วมกันบ่อย (Association wordlists: AW)

การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อเสนอแนะ แบ่งเป็น 4 ส่วน ประกอบด้วย (1) ค่าความถูกต้องเชิงบวก (True positive) จำนวนทั้งสิ้น 367 ประโยค (2) ค่าความผิดพลาดเชิงบวก (False positive) จำนวนทั้งสิ้น 61 ประโยค (3) ค่าความผิดพลาดเชิงลบ (False negative) จำนวนทั้งสิ้น 25 ประโยค และ (4) ค่าความถูกต้องเชิงลบ จำนวนทั้งสิ้น 362 ประโยค ดัง ภาพที่ 5.1

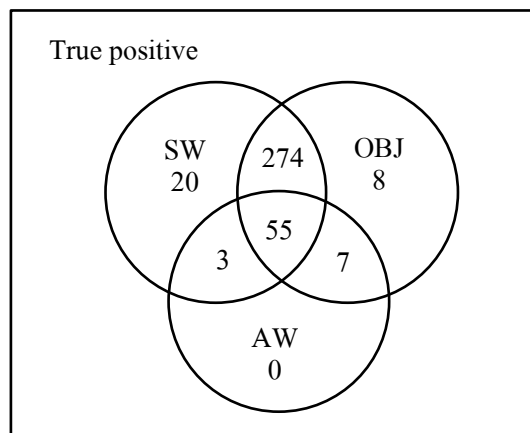


ภาพที่ 5.1 การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อเสนอแนะ

5.2.1 ความถูกต้องเชิงบวก (True positive)

ความถูกต้องเชิงบวก หมายถึง ประโยคที่ผู้เชี่ยวชาญระบุว่าเป็นข้อเสนอแนะ (S) และแบบจำลองทำนายว่าเป็นข้อเสนอแนะ (S) ซึ่งประโยคที่แบบจำลองสามารถทำนายได้ว่าเป็นข้อเสนอแนะจะต้องมีส่วนประกอบของข้อเสนอแนะครบทั้ง 3 ส่วน หรือมีส่วนประกอบอื่น ๆ ของข้อเสนอแนะ มีจำนวนทั้งสิ้น 367 ประโยค ดังภาพที่ 5.2

จากผลการทดลองพบว่าประโยคที่วิเคราะห์ถูกต้อง 75% มีคำบ่งชี้ข้อเสนอแนะ และคำระบุหัวข้อเป็นส่วนประกอบของประโยค และ 15% มีส่วนประกอบครบทั้งสามส่วน คือคำบ่งชี้ข้อเสนอแนะ คำระบุหัวข้อ และวลีข้อเสนอแนะ หากมีฐานความรู้ทางคำหรือวลีที่เป็นข้อเสนอแนะที่มากขึ้น จะทำให้การวิเคราะห์ข้อเสนอมีความถูกต้องยิ่งขึ้น ซึ่งสามารถทำได้โดยการเพิ่มจำนวนชุดข้อมูลเรียนรู้ให้มากขึ้น ตัวอย่างประโยคที่จำแนกถูกต้องว่าเป็นข้อเสนอแนะ (True positive) ดังตารางที่ 5.2



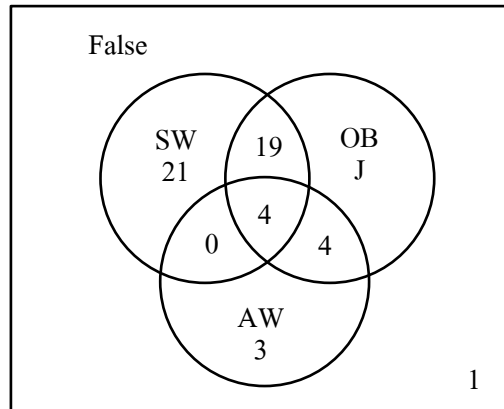
ภาพที่ 5.2 การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความถูกต้องเชิงบวก

ตารางที่ 5.2 ตัวอย่างส่วนประกอบของประโยคข้อเสนอแนะและประโยคที่มีค่าความถูกต้องเชิงบวก

ส่วนประกอบของประโยค ข้อเสนอแนะ	ตัวอย่างประโยคที่มีค่าความถูกต้องเชิงบวก
SW, OBJ, AW	อยาก<S _u >ให้<AUX>นำ<AW-VACT> พระอาทิตย์ขึ้น แรม <OBJ2>กลับมา <VACT> นาย <AW-VACT> ซ้ำ <ADV>อีก<ADV>เรื่อย<ADV> ๆ <N>
SW, OBJ, AW	พิธีกร<AW-N> ยิปโซ<OBJ3> พูด<AW-VACT> ฟัง <VACT>ไม่<NEG>รู้<VSTA>เรื่อง<N> <SPC>พูด <VACT>ซ้ำ<ADV> ๆ <N>หน่อย<ADV>ได้ไหม<S _u > หรือ<CONJ>ไม่<NEG>ขอ<S _u >ให้<AUX>เปลี่ยน<AW- VACT> พิธีกร<AW-N>
SW, OBJ	การ์ตูน<OBJ1>animation<N> <SPC>ก่อน<ADV>ข่าว <OBJ1>19.00 น.<INT>น่าจะ<S _u >นำ<VACT>ไป <VACT>เผยแพร่<VACT>ต่อ<ADV>เยอะ<ADV> ๆ <N>

5.2.2 ความผิดพลาดเชิงบวก (False positive)

ความผิดพลาดเชิงบวก หมายถึง ประโยคที่ผู้เชี่ยวชาญระบุว่าเป็นไม่ข้อเสนอแนะ (S') แต่แบบจำลองทำนายว่าเป็นข้อเสนอแนะ (S) เนื่องจากเป็นประโยคที่มีความกำกวม ผู้อ่านอาจตีความว่าเป็นหรือไม่เป็นข้อเสนอแนะ มีจำนวนทั้งสิ้น 61 ประโยค ดังภาพที่ 5.3 และตัวอย่างประโยคที่จำแนกผิดว่าเป็นข้อเสนอแนะ (False Positive) ดังตาราง 5.3



ภาพที่ 5.3 การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความผิดพลาดเชิงบวก

ตารางที่ 5.3 ตัวอย่างส่วนประกอบของประโยคข้อเสนอแนะและประโยคที่มีค่าความผิดพลาดเชิงบวก

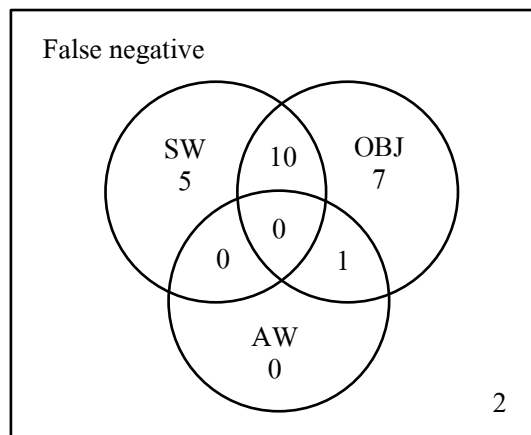
ส่วนประกอบของประโยค ข้อเสนอแนะ	ตัวอย่างประโยคที่มีค่าความผิดพลาดเชิงบวก
SW, OBJ, AW	อยาก<S₂>ทราบ<VACT>ผู้บรรยาย<N>รายการ<N>สารคดี <OBJ1>ท่อง โลก กว้าง <OBJ2>เวลา <Dom-N> <SPC>18.00 น.<INT> <SPC>และ<CONJ> ผู้เขียนบทบรรยาย<N> <SPC>เป็น<VSTA>ใคร<PRON>คะ<END> <SPC>เขียน<VACT>บทบรรยาย<N>ได้<AUX>ดี<ADJ>แล้ว<ADV>เสียง<AW-N>คนพากษ์<N>ก็<ADV>ฟัง <AW-VACT>เพลิน<ADV>ค่ะ<END>
SW, OBJ, AW	ขอ<S₂>แสดง<VACT>ความเห็น<N>เกี่ยวกับ<PREP>พิธีกร<Dom-N>หญิง<N> <SPC>รายการ<N>ย้อนสยามผ่านฟิล์มจิ๋ว<OBJ2>ที่<PREP>มี<VACT>ใน<PREP>วันอาทิตย์<N>ช่วง<N>บ่าย<N> <SPC>ว่า<AUX> <SPC>เป็น<VSTA>พิธีกร<AW-N>ที่<PREP>พูด<AW-VACT>ฟัง<VACT>ไม่<NEG>ได้<AUX>สาระ <Dom-N>เลย <ADV>

ตารางที่ 5.3 (ต่อ)

AW	เปลี่ยน<AW-VACT> พิธีกร<AW-N>เมื่อไร<ADV>จะ <AUX>กลับมา<VACT>คู่<VACT>ใหม่<ADV>
AW	สี่<Dom-N>ช่วย<AW-VACT> นำเสนอ<AW-VACT> ด้วย<ADV>นะ<END>ครับ<END> <SPC>เพื่อ<CONJ> ที่<PREP>รัฐบาล<N>จะ<AUX>ได้<AUX>แก้<VACT> ปัญหา<Dom-N>ให้<AUX>คน<N>ตกงาน<VACT>บ้าง <PRON>

5.2.3 ความผิดพลาดเชิงลบ (False negative)

ความผิดพลาดเชิงลบ หมายถึง ประโยคที่ผู้เชี่ยวชาญระบุว่าเป็นข้อเสนอแนะ (S) แต่แบบจำลองทำนายว่าไม่เป็นข้อเสนอแนะ (S') มีจำนวนทั้งสิ้น 25 ประโยค ดังภาพที่ 5.4 และตัวอย่างประโยคที่จำแนกถูกว่าไม่เป็นข้อเสนอแนะ (False Negative) ดังตารางที่ 5.4



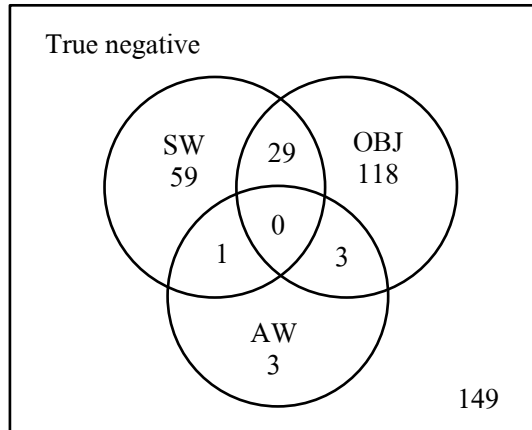
ภาพที่ 5.4 การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความผิดพลาดเชิงลบ

ตารางที่ 5.4 ตัวอย่างส่วนประกอบของประโยคข้อเสนอแนะและประโยคที่มีค่าความผิดพลาดเชิง
 ลบ

ส่วนประกอบของประโยค ข้อเสนอแนะ	ตัวอย่างประโยคที่มีค่าความผิดพลาดเชิงลบ
SW, OBJ	<p>อยาก<S_u>ให้<AUX>ทาง<N>ThaiPBS<OBJ0>นำเสนอ <VACT>ให้<AUX>ครบ<ADV>ทุก<DET>มิติ<N> <SPC>ที่<PREP>ผ่าน<VACT>มา<VACT>นั้น<PRON> ชัดเจน<ADV>ว่า<AUX> <SPC>สื่อ<Dom- N>ThaiPBS<OBJ0>สนับสนุน<VACT>แนวคิด<N> <SPC>และ<CONJ>นโยบาย<Dom-N>ของ<PREP>พรรค การเมือง<N>อนุรักษ์นิยม<N>โดย<PREP>การนำเสนอ <N>เชิง<N>บวก<ADJ>อยู่<PREP>ตลอด<ADV>เวลา <Dom-N>เรียก<VACT>ได้<AUX>ว่า<AUX>ทำ<VACT>ให้ <VACT>ภาพลักษณ์<N>ของ<PREP>พรรค<N> อนุรักษ์นิยม<N> <SPC>และ<CONJ>องค์กร<Dom-N> อิสระ<ADJ> <SPC>และ<CONJ>ทุก<DET>ฝ่าย<N></p>
SW, OBJ	<p>ThaiPBS<OBJ0> อย่า<S_u>เป็น<VSTA>เหมือน<CONJ> ช่อง<Dom-N>ฟรีทีวี<Dom-N>อื่น ๆ <ADV>เลย<ADV> ที่<PREP>ถูก<VACT>จำกัด<VACT>การนำเสนอ<N> ข้อมูล<Dom-N>อย่าง<ADV>ตรงไปตรงมา<ADV></p>
OBJ	<p><N>แบบ<N>นี้<ADJ>มัน<PRON>สื่อ<OTH>ให้<AUX> เกิด<VACT>ปัญหา<Dom-N>มากขึ้น<OTH>ไป <VACT>เหมือน<CONJ>ไม่<NEG>ให้<AUX>เกียรติ <N>ผู้อื่น<PRON>เลย<ADV>ทั้ง ๆ ที่<ADV>ยัง<PREP> ไม่<NEG>มี<VACT>หลักฐาน<N> <SPC>ระวัง<VACT> เรื่อง<N>นี้<ADJ>หน่อย<ADV>นะ<END>ครับ<END></p>

5.2.4 ความถูกต้องเชิงลบ (True Negative)

ความถูกต้องเชิงลบ หมายถึง ประโยคที่ผู้เชี่ยวชาญระบุว่าเป็นไม่ข้อเสนอแนะ (S') และแบบจำลองทำนายว่าไม่เป็นข้อเสนอแนะ (S') มีจำนวนทั้งสิ้น 362 ประโยค ดังภาพที่ 5.5 และตัวอย่างประโยคที่จำแนกถูกว่าเป็นข้อเสนอแนะ (True Negative) ดังตารางที่ 5.5



ภาพที่ 5.5 การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อเสนอแนะที่ค่าความถูกต้องเชิงลบ

ตารางที่ 5.5 ตัวอย่างส่วนประกอบของประโยคข้อเสนอแนะและประโยคที่มีค่าความถูกต้องเชิงลบ

ส่วนประกอบของประโยค ข้อเสนอแนะ	ตัวอย่างประโยคที่มีค่าความผิดพลาดเชิงลบ
OBJ	<u>ThaiPBS<OBJ0></u> <SPC>มี<VACT>การ<N>เก็บ <VACT>ข้อมูล<Dom-N>และ<CONJ>นำเสนอ<VACT> ข่าวสาร<Dom-N>ที่<PREP>มี<VACT>ความ<N> หลากหลาย<ADJ>และ<CONJ>มี<VACT>ความ<N>เป็น กลาง<VATT>มาก<ADV>เลข<ADV>ครึ่ง<END>
SW	<u>อยาก<S></u> ให้<AUX>ลูกสาว<N>และ<CONJ>เพื่อน<N> ๆ<N>ได้<AUX>ไป<VACT>ร่วม<VACT>กิจกรรม<N> ใน<PREP>การ<N>ทำ<VACT>ขนม<N>ไทย<Dom-N> ไม่<NEG>ทราบ<VACT>ว่า<AUX>จำกัด<VACT>อายุ

ตารางที่ 5.5 (ต่อ)

	<N>เท่าไร <QUES>ลูกสาว <N>อายุ <N>15 <INT> <SPC>ปี <N>แล้ว <ADV>ค่ะ <END>
SW, OBJ	อยาก <S _u >ทราบ <VACT>รายละเอียด <N>เรื่อง <N>แวนดา 3 มิติ <N>ทาง <PREP>ThaiPBS <OBJ0>จะ <AUX>แจก <VACT>ฟรี <ADV>ต้อง <AUX>ทำ <VACT>อย่างไร <QUES>บ้าง <PRON>ครับ <END>

จากศึกษารูปแบบประโยคข้อเสนอแนะ ประกอบด้วย 3 ส่วน ได้แก่ (1) คำบ่งชี้ข้อเสนอแนะ (2) คำระบุหัวข้อ และ (3) วลีเสนอแนะ แต่จากผลการทดลองพบว่ามีประโยคข้อเสนอแนะบางประโยคที่มีความกำกวม คือมีส่วนประกอบของประโยคข้อเสนอแนะไม่ครบทุกส่วนประกอบ เช่น ไม่มีคำบ่งชี้ข้อเสนอแนะที่ชัดเจน มีเพียงวลีข้อเสนอแนะเท่านั้น หรือมีคำบ่งชี้ข้อเสนอแนะอยู่ในประโยคแต่ไม่พบส่วนประกอบของวลีข้อเสนอแนะ จึงทำให้การวิเคราะห์มีความผิดพลาด ส่วนรูปแบบประโยคที่แบบจำลองสามารถวิเคราะห์ได้ถูกต้องว่าเป็นประโยคข้อเสนอแนะประกอบด้วยส่วนประกอบของประโยคอย่างน้อย 2 ส่วน ได้แก่ (1) คำบ่งชี้ข้อเสนอแนะ (2) คำระบุหัวข้อ เป็นหลักซึ่งยังขาดวลีข้อเสนอแนะ ดังนั้นการเพิ่มประสิทธิภาพของสัปดาห์วลีข้อเสนอแนะสามารถทำได้โดยการเพิ่มชุดข้อมูลเรียนรู้ เพื่อให้กระบวนการสร้างฐานความรู้ทางภาษาของวลีข้อเสนอแนะมีความรู้ทางคำหรือวลีข้อเสนอแนะที่มากขึ้น จะสามารถวิเคราะห์ประโยคข้อเสนอแนะได้ถูกต้องยิ่งขึ้น หรือใช้การวิเคราะห์ภาษาในระดับความหมายของประโยคร่วมด้วย เพื่อให้คอมพิวเตอร์สามารถทำความเข้าใจกับความหมายของประโยคข้อเสนอแนะได้ดียิ่งขึ้น

5.3 สิ่งที่ได้รับจากงานวิจัย (Contributions)

1. กระบวนการสกัดข้อเสนอแนะออกจากบทวิจารณ์ประเภทอื่นด้วยเทคนิคการเรียนรู้ของเครื่อง ที่มีเป้าหมายในการจำแนกข้อเสนอแนะออกจากความคิดเห็นทั่วไป ซึ่งในงานวิจัยของ Vishwanath (2011) ได้ใช้วิธีการจำแนกข้อเสนอแนะด้วยวิธีวิศวกรรมองค์ความรู้ ซึ่งเป็นเทคนิคการใช้ผู้เชี่ยวชาญสร้างกฎการตัดสินใจในการจำแนกข้อเสนอแนะ ซึ่งวิธีนี้มีข้อเสียคือเมื่อโดเมนข้อมูลเปลี่ยนไป จำเป็นต้องสร้างกฎการตัดสินใจใหม่ทุกครั้ง หากมีปริมาณข้อมูลมากขึ้นต้องใช้

ระยะเวลานานในการสร้างกฎ และมีข้อผิดพลาดสูง ซึ่งแตกต่างจากวิธีการเรียนรู้ด้วยเครื่อง ที่เป็นการสร้างแบบจำลองการจำแนกเอกสารหรือข้อความแบบอัตโนมัติ ด้วยวิธีการเรียนรู้จากชุดข้อมูลเรียนรู้ ทำให้มีความถูกต้องในการจำแนกข้อเสนอแนะที่สูงกว่าและไม่จำเป็นต้องใช้ทรัพยากรทั้งด้านคนและเวลามาก สามารถนำกระบวนการวิเคราะห์ที่นำเสนอไปประยุกต์ใช้กับหัวข้อเรื่องอื่น ๆ ได้

2. วิธีการสร้างฐานความรู้ทางภาษาด้วยกฎความสัมพันธ์ (Association rules mining) โดยปกติแล้วการจำแนกข้อความไม่ว่าจะเป็นการจำแนกตามหัวข้อ (Topic based) หรือตามข้อความเห็น (Polarity based) จะสามารถจำแนกได้จาก “คำเดียว” เพียงหนึ่งคำ เช่น การจำแนกตามหัวข้อสามารถจำแนกได้จากคำเดียวอย่างน้อยหนึ่งคำที่อยู่ในกลุ่มเดียวกันกับหัวข้อ หรือการจำแนกตามข้อความเห็น สามารถจำแนกได้จากคำบ่งชี้ข้อความเห็นที่ใช้ซึ่งชี้ข้อความเห็นเป็นบวก ลบ หรือเป็นกลางได้ แต่สำหรับการวิเคราะห์ข้อเสนอแนะไม่สามารถวิเคราะห์ประเภทเอกสารหรือประโยคได้จากคำเพียงคำเดียว จำเป็นต้องวิเคราะห์การเกิดขึ้นร่วมกันของคำมากกว่า 2 คำขึ้นไป ตัวอย่างประโยคที่มีคำบ่งชี้ข้อเสนอแนะ (Suggestion indicator) “อยาก” เช่น

“อยากให้มีรายการไปตลอด” คือประโยคที่ไม่เป็นข้อเสนอแนะ (Non-suggestion)

“อยากให้เปลี่ยนพิธีกร” คือประโยคข้อเสนอแนะ (Suggestion)

จะเห็นว่าคำว่า “อยาก” หนึ่งคำ สามารถปรากฏอยู่ได้ทั้งประโยคที่เป็นข้อเสนอแนะ และประโยคที่ไม่เป็นข้อเสนอแนะ ดังนั้นการวิเคราะห์ข้อเสนอแนะจึงไม่สามารถวิเคราะห์ได้จากคำบ่งชี้ข้อเสนอแนะเพียงคำเดียวได้ แต่จำเป็นต้องอาศัยการเกิดขึ้นของคำหรือวลีข้อเสนอแนะร่วมด้วย เช่นคำว่า “เปลี่ยนพิธีกร” เป็นต้น ดังนั้นผู้ศึกษาจึงนำเสนอกระบวนการสร้างฐานความรู้ทางคำศัพท์ที่มีโอกาสเป็นวลีข้อเสนอแนะขึ้น ด้วยวิธีการประยุกต์ใช้เทคนิคของกฎความสัมพันธ์ (Association rules mining) ในการวิเคราะห์เหมืองข้อเสนอแนะ

3. ประเภทของข้อเสนอแนะ

การจำแนกข้อเสนอแนะออกเป็นประเภทก่อนเข้าสู่กระบวนการสกัดหัวข้อ ข้อเสนอแนะ จะช่วยให้ประสิทธิภาพการสกัดหัวข้อข้อเสนอแนะมีความถูกต้องยิ่งขึ้น สามารถลดระยะเวลาการค้นหา การอ่านเพื่อการตีความ และสามารถนำข้อเสนอแนะไปใช้ประโยชน์ได้อย่างรวดเร็วยิ่งขึ้น

5.4 ปัญหาและอุปสรรคของงานวิจัย

กระบวนการวิเคราะห์เหมืองข้อเสนอแนะสามารถวิเคราะห์ประโยคข้อเสนอแนะได้ แต่เนื่องจากข้อความแสดงข้อเสนอแนะสำหรับภาษาไทยที่มีโครงสร้างไม่แน่นอน มีความหลากหลาย และความยืดหยุ่นของการใช้ภาษาไทย เช่น การเขียนภาษาไทยสามารถละประธานของประโยคไว้ได้ คำแสดงต่าง ๆ ที่ใช้ในการแสดงความคิดเห็นและข้อเสนอแนะ เป็นต้น จึงทำให้การวิเคราะห์เหมืองข้อเสนอแนะยังมีข้อผิดพลาดอยู่ การวิเคราะห์ภาษาในระดับความหมายจะช่วยให้สามารถวิเคราะห์ข้อเสนอแนะได้ถูกต้องยิ่งขึ้น ซึ่งในปัจจุบันมีคลังเครือข่ายคำไทย (Thai WordNet) ที่เป็นเครื่องมือในการช่วยวิเคราะห์ภาษาธรรมชาติในระดับความหมายได้ แต่ยังมีคำค่อนข้างจำกัด และอยู่ระหว่างขั้นตอนการพัฒนา

5.5 ข้อเสนอแนะและงานวิจัยในอนาคต

1. การเพิ่มจำนวนชุดข้อมูลเรียนรู้จะช่วยให้ประสิทธิภาพการวิเคราะห์เหมืองข้อเสนอแนะด้วยวิธีการเรียนรู้ของเครื่องดีขึ้น และมีข้อมูลเพียงพอต่อการวิเคราะห์หาความสัมพันธ์ของคำตั้งแต่ 2 คำขึ้นไปที่มีโอกาสจะเป็นวลีข้อเสนอแนะได้มากยิ่งขึ้น ซึ่งจะนำไปสู่การวิเคราะห์ข้อเสนอแนะที่มีความถูกต้องยิ่งขึ้น เนื่องจากมีปริมาณชุดข้อมูลเรียนรู้ที่ครอบคลุมข้อมูลที่จะเกิดขึ้นในอนาคต และปริมาณคำศัพท์ในฐานความรู้ทางภาษาที่มากขึ้น

2. การประมวลผลภาษาธรรมชาติในระดับความหมาย (Semantic) หรือในระดับที่สูงขึ้นไป จะช่วยให้สามารถวิเคราะห์ข้อเสนอแนะได้ถูกต้องสูงขึ้น เนื่องจากภาษาที่ใช้ในการแสดงข้อเสนอแนะมีรูปแบบที่หลากหลายและเฉพาะตัว เป็นไปตามธรรมชาติของการเรียนรู้ในสมองมนุษย์แต่ละคน ซึ่งแตกต่างกันไป การวิเคราะห์ในระดับความหมาย เช่นการวิเคราะห์ข้อเสนอแนะร่วมกับเครือข่ายคำไทย (Thai WordNet) จะช่วยให้สามารถเข้าใจความหมายของคำได้ดียิ่งขึ้น และสามารถระบุวลีข้อเสนอแนะได้อย่างถูกต้อง ทำให้สามารถสามารถวิเคราะห์ข้อเสนอแนะที่มีความกำกวมได้ ยกตัวอย่างประโยคข้อเสนอแนะที่มีความกำกวม เช่น “เมื่อใดรายการจะปรับปรุงเนื้อหาเสียที” ซึ่งอาจตีความเป็นข้อเสนอแนะได้ว่าต้องการให้ปรับปรุงเนื้อหาของรายการ หรือตัวอย่างประโยค “ไม่ทราบผู้บรรยายรายการสารคดีท่องโลกกว้างเวลา 18.00 น. และผู้เขียนบทบรรยายเป็นใคร” อาจตีความเป็นข้อเสนอแนะได้ว่า อยากให้เพิ่มเติมรายละเอียดในรายการให้มากขึ้น ซึ่งการวิเคราะห์ในลักษณะนี้ต้องอาศัยการวิเคราะห์ระดับความหมาย บริบทข้างเคียงร่วมด้วย จึงจะสามารถตีความตัวอย่างประโยคดังกล่าวได้ งานวิจัยต่อไป ผู้วิจัยจะทำการศึกษารูปแบบของ

ข้อเสนอแนะ และนำเสนอกระบวนการสกัดหาข้อเสนอนะให้ครอบคลุมรูปแบบของประโยค
ข้อเสนอแนะภาษาไทย รวมถึงการเพิ่มประสิทธิภาพการวิเคราะห์ข้อเสนอแนะให้มีความถูกต้อง
ยิ่งขึ้น และสามารถวิเคราะห์ข้อเสนอแนะที่มีความกำกวมได้

บรรณานุกรม

- กนกวรรณ เขียนวรรณ. 2555. การประมวลผลภาษาธรรมชาติ. ค้นวันที่ 3 ธันวาคม 2555
จาก www.mbs.mut.ac.th/paper/pdf/29.pdf
- การเรียนรู้ภาษาไทย. 2555. การแสดงความคิดเห็น. ค้นวันที่ 29 พฤศจิกายน 2555
จาก <http://www.yorwor2.ac.th/yorwor2/thaionline/comment/comment.html>
- ชูชาติ หล่ไชยะศักดิ์. 2554. **Material: Text mining.** Lecturer website: Faculty of Information
Technology. ค้นวันที่ 12 ธันวาคม 2555 จาก <http://suanpalm3.kmutnb.ac.th/teacher/choochart/>
- มหาวิทยาลัยรามคำแหง. 2555. การวิเคราะห์ข้อความ. ค้นวันที่ 12 ธันวาคม 2555
จาก <http://e-book.ram.edu/e-book/c/CT477/CT477-2.pdf>
- ยี่น ภู่วรรณ และ ชัยยงค์ วงศ์ชัยสุวัฒน์. 2535. การประมวลผลภาษาธรรมชาติ.
กรุงเทพมหานคร: สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี.
- วัลัญญา วรรณศรี. 2553. ระบบวิเคราะห์ข้อความแสดงความคิดเห็นสำหรับโรงแรม. ค้นวันที่
13 มีนาคม 2554 จาก <http://thailang.nectec.or.th/halloffame/>
- วิทยาลัยเทคโนโลยีภาคตะวันออก. 2555. การพูดแสดงความคิดเห็นและการอธิบาย. ค้นวันที่
12 ธันวาคม 2555 จาก <http://e-learning.e-tech.ac.th/learninghtml/thai2000/unit005.html>
- สมนึก สินรูปวน. 2546. การวิเคราะห์กระจายคำในประโยคภาษาไทย โดยการโปรแกรมเชิง
เจนเนติก. วิทยานิพนธ์ปริญญาโทมหาบัณฑิต สถาบันบัณฑิตพัฒนบริหารศาสตร์.
- Agrawal Rakesh and Srikant Ramakrishnan. 1994. Fast Algorithms for Mining Association
Rules in Large Databases. In **Proceedings of the 20th International Conference
on Very Large Data Bases (VLDB'94)**. Bocca, Jorge B, Matthias Jarke and Carlo
Zaniolo, eds. San Francisco: Morgan Kaufmann Publishers. Pp. 412-420.
- Basu, A., Watters, C. and Shepherd, M. 2003. Support Vector Machines for Text Categorization.
In **Proceedings of the 36th Annual Hawaii International Conference on System
Sciences (HICSS'03)**. Retrieved May10, 2012 from ACM Digital Library.

- Chirawichitchai Nivet, Sanguansatand Parinya, Meesad Phayung. 2011. Developing and Effective Automatic Thai Document Categorization. **NIDA Development Journal**. 51: 187-206.
- Feldman, Ronen and Sanger, James. 2007. **The text mining handbook**. Cambridge: Cambridge University Press.
- Hotho, Andreas, Nurnberger, Andreas and Paaß, Gerhard. 2005. A Brief Survey of Text Mining. **Journal for Computational Linguistics and Language Technology**. 20 (January):19-62. Retrieved March 20, 2012 from <http://www.kde.cs.uni-kassel.de/hotho/publication.html>
- Hu Mingqing and Liu Bing. 2004a. Mining and Summarizing Customer Reviews. In **Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. Retrieved March 13, 2011 from <http://www.informatik.uni-trier.de/~ley/db/conf/kdd/kdd2004.html>
- Hu Mingqing and Liu Bing. 2004b. Mining Opinion Features in Customer Reviews. In **Proceedings of the 19th national conference on Artificial intelligence (AAAI'04)**. Anthony G. Cohn, ed. California: AAAI Press. Pp. 755-760.
- IBM Software Business Analytics. 2012. **CRISP DM**. Retrieved February 10, 2012 from <http://www.ibm.com>
- Joachims, Thorsten. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In **Proceedings of the 10th European Conference on Machine Learning (ECML'98)**. Claire Nedellec and Céline Rouveirol, eds. London: Springer-Verlag. Pp. 137-142.
- Kim Soo-Min and Hovy, Eduard. 2004. Determining the Sentiment of Opinions. In **Proceedings of the 20th international conference on Computational Linguistic**. Retrieved December 10, 2011 from ACM Digital Library.
- Kohavi, Ron. 1995. อ้างถึงใน สิทธิโชค มุกดาสกุลภีบาล. 2551. การวัดประสิทธิภาพของ ขั้นตอนวิธีตัวจำแนก C4.5, ADTree และ Naïve Bayes ในการจำแนกข้อมูลการชุกช่อนสิ่งเสพย์ติดสำหรับไปรษณีย์ระหว่าง วิทยานิพนธ์ปริญญาโทมหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์.

- Kongthon Alisa, Angkawattanawit Niran, Sangkeetrakarn Chatchawal, Palingoon Pornpimon and Haruechaiyasak Choochart. 2010. Using an Opinion Mining Approach to Exploit Web Content in Order to Improve Customer Relationship Management. **Proceedings of Technology Management for Global Economic Growth**. Retrieved December 10, 2011 from IEEE Xplore Digital Library.
- Lai, Kenneth and Cerpa, Narciso. 2007. Support vs Confidence in Association Rule Algorithms. In **Proceedings of the 3rd International Conference on Intelligent Computing**. De-Shuang Huang, Laurent Heutte, and Marco Loog, eds. Heidelberg: Springer-Verlag Berlin. Pp. 465-474.
- Liu Bing. 2011. Tutorial Sentiment Analysis and Opinion Mining. **UIC College of Engineering**. Retrieved March 1, 2012 from <http://www.cs.uic.edu/~liub/>
- Manning, Christopher D. and Schütze, Hinrich. 2000. **Foundations of Statistical Natural Language Processing**. 2nd ed. London: The MIT Press.
- Nuntiyagul Atorn. 2006. **Text Categorization & Retrieval for Thai Item Bank using Patterned Keyword in phrase (PKIP)**. Doctoral dissertation, Mahidol University.
- Pang Bo and Lee Lillian. 2008. Opinion Mining and Sentiment Analysis. **Journal of the ACM**. 2 (January): 1-135. Retrieved December 10, 2011 from ACM Digital Library.
- Ramos, Juan. 1999. Using TF-IDF to Determine Word Relevance in Document Queries. **The College of Information Sciences and Technology**. Retrieved February 10, 2012 from CiteSeer Digital Library.
- Sukhum Khampol, Nitsuwat Supot and Haruechaiyasak Choochart. 2011. Opinion Detection in Thai Political News Columns Based on Subjectivity Analysis. **Information Technology Journal**. 14 (July): 27-31. Retrieved March 1, 2012 from <http://suanpalm3.kmutnb.ac.th/journal>
- Turney, Peter D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**. Assoc. for Computational Linguistics. Pennsylvania: Association for Computational Linguistics Stroudsburg. Pp.417-424.

- Vishwanath, J. and Aishwarya, S. 2011. User Suggestions Extraction from customer Reviews A Sentiment Analysis approach. **International Journal on Computer Science and Engineering (IJCSE)**. 3 (March): 1203-1206.
- Viswanathan Amar, Venkatesh Prasanna, Vasudevan Bintu, Balakrishnan Rajesh and Shastri Lokendra. 2011. Suggestion Mining from Customer Reviews. **AMCIS 2011 PROCEEDINGS**. Retrieved February 10, 2012 from AIS Electronic Library (AISeL).
- Witten, Ian H. and Frank, Eibe.2005. **Data Mining**. California: Morgan Kaufmann Publishers. Pp. 188-193.
- Yang Yiming and Pedersen, Jan O. 1997. A Comparative Study on Feature Selection in Text Categorization. In **Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)**. Douglas H. Fisher, ed. California: Morgan Kaufmann Publishers. Pp. 412-420.
- Zaiane, Osmar R. and Antonie, Maria-Luiza. 2002. Classifying text documents by associating terms with text categories. In **Proceedings of the 13th Australasian Database Conference (ADC'02)**. Xiaofang Zhou, ed. Melbourne: Australian Computer Society.

ภาคผนวก

ภาคผนวก ก

ตัวอย่างคำที่ไม่มีนัยสำคัญ (Stop words)

คำหยุด	คำบุพบท	คำสันธาน	คำสรรพนาม	คำลักษณนาม
ครับ	ที่	และ	ฉัน	คน
ครับผม	ซึ่ง	หรือ	ผม	ตัว
ค่ะ	ถึง	เพราะ	เธอ	ตำแหน่ง
จ้า	ตรง	เนื่องจาก	เรา	นาฬิกา
ซี	โดย	เช่น	มัน	ชั่วโมง
นะ	ใกล้	ตอน	เหล่านั้น	บท
เถอะ	ได้	ดังนั้น	แก่	ประการ
เถิด	ใน	จนกว่า	โน่น	ฝ่าย
ล่ะ	กับ	จากนั้น	พวกเขา	หน่วย
หรือ	ของ	ส่วน	พวกคุณ	หมวด
หรือ	จาก	เพื่อว่า	บางคน	อัน

ภาคผนวก ข

ตัวอย่างคำที่เกิดขึ้นร่วมกันบ่อย (Association Wordlist)

Item 1	Item 2	Support
ปรับปรุง<VACT>	เนื้อหา<N>	0.03
พิธีกร<N>	พูด<VACT>	0.025
เอา<VACT>	ฉาย<VACT>	0.022
ปรับปรุง<VACT>	การนำเสนอ<N>	0.022
พูด<VACT>	ฟัง<VACT>	0.02
เวลา<N>	ออกอากาศ<VACT>	0.018
เวลา<N>	เพิ่ม<VACT>	0.018
นำ<VACT>	ออกอากาศ<VACT>	0.018
นำ<VACT>	ภาพ<N>	0.018
นำ<VACT>	ฉาย<VACT>	0.015
นำเสนอ<VACT>	ประชาชน<N>	0.015
พิธีกร<N>	เปลี่ยน<VACT>	0.015
อ่าน<VACT>	ผิด<VACT>	0.012
ผู้ประกาศข่าว<N>	ผิด<VACT>	0.012
ใส่<VACT>	ส่วนตัว<N>	0.012
นำเสนอ<N>	เนื้อหา<VACT>	0.01
กลับมา<VACT>	ออกอากาศ<VACT>	0.01

ภาคผนวก ค

ตัวอย่างคำเฉพาะเจาะจงที่เกิดขึ้นบ่อยภายใต้โดเมน ที่ใกล้เคียงกัน (Domain Wordlist)

Words	Occurrence	Percentage
โทรทัศน์	411	100%
ช่อง	325	79%
ทีวี	232	56%
ภาพ	165	40%
เวลา	163	40%
ออกอากาศ	58	14%
สัญญาณ	56	14%
แสดง	45	11%
ประกาศ	39	9%
เสียง	37	9%
ผู้ชม	36	9%
สื่อ	34	8%
ภาษา	29	7%
ออนไลน์	21	5%
หน้าจอ	15	4%
ฝั่ง	14	3%

ประวัติผู้เขียน

ชื่อ ชื่อสกุล

นางสาวกานดา แผ้ววัฒนากุล

ประวัติการศึกษา

วิทยาศาสตรบัณฑิต

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร

ลาดกระบัง

2549

ประสบการณ์การทำงาน

พ.ศ. 2549

เจ้าหน้าที่วิเคราะห์ข้อมูล บริษัท มิราเคิล กรุ๊ป

พ.ศ. 2551 - 2553

เจ้าหน้าที่วิเคราะห์ระบบสารสนเทศงานจัดซื้อ
บริษัท เจริญโภคภัณฑ์อาหาร จำกัด (มหาชน)

ผลงานทางวิชาการ

เรื่อง Suggestion Mining and Knowledge

Construction from Thai Television Program

Reviews บทความตีพิมพ์ในงานประชุมทาง

วิชาการนานาชาติ The International

MultiConference of Engineers and Computer

Scientists 2013 จัดขึ้นที่ โรงแรม Royal Garden

Hotel ส่องกง ประเทศจีน วันที่ 13-15 มีนาคม

2556.