การจำแนกข้อความส่อเสียดในทวิตเตอร์ด้วยการใช้ความน่าจะเป็นของทวีต

นายกษิดิ์เดช ทาแป้ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศทางธุรกิจ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

SARCASM CLASSIFICATION IN TWITTER USING PROBABILITY OF TWEETS

Mr. Kasidech Tapang

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Information Technology in Business

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2016

หัวข้อวิทยานิพนธ์	การจำแนกข้อความส่อเสียดในทวิตเตอร์ด้วยการใช้ความน่าจะเป็นของทวีต
โดย	นายกษิดิ์เดช ทาแป้ง
สาขาวิชา	เทคโนโลยีสารสนเทศทางธุรกิจ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. จันทร์เจ้า มงคลนาวิน

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. พิมพ์มณี รัตนวิชา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. จันทร์เจ้า มงคลนาวิน)

.....กรรมการ

(อาจารย์ ดร. พงษ์สิน ภูแสนคำ)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร. เทพชัย ทรัพย์นิธิ)

กษิดิ์เดช ทาแป้ง : การจำแนกข้อความส่อเสียดในทวิตเตอร์ด้วยการใช้ความน่าจะเป็นของทวีต (SARCASM CLASSIFICATION IN TWITTER USING PROBABILITY OF TWEETS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. จันทร์เจ้า มงคลนาวิน, 71 หน้า.

ข้อความส่อเสียดเป็นปัญหาหนึ่งในการประมวลผลภาษาธรรมชาติเนื่องจากข้อความส่อเสียดจะกลับขั้วความคิดเห็นของข้อความทำให้การวิเคราะห์ความคิดเห็นของข้อความผิดไปจากความเป็นจริง งานวิจัยนี้ได้เสนอวิธีจำแนกข้อความส่อเสียดออกจากข้อความปกติ โดยประยุกต์ใช้ความน่าจะเป็นของข้อความ และใช้ข้อมูลความคิดเห็นของผู้บริโภคเกี่ยวกับเครือข่ายอินเตอร์เน็ตเครือข่ายหนึ่งบนเครือข่ายสังคมออนไลน์ทวิตเตอร์ในการศึกษา โดยเก็บรวบรวมข้อมูลผ่านช่องทาง Advance Search API เริ่มตั้งแต่วันที่ 25 มกราคม 2553 ถึงวันที่ 9 มิถุนายน 2559 ทั้งสิ้น 4,027 ข้อความ จากนั้นจึงประมวลผลข้อมูลเบื้องต้นโดยตัดข้อความที่มีความซ้ำซ้อน URL ที่ปรากฏอยู่ภายในข้อความ เครื่องหมายแฮชแท็กรวมถึงข้อความแฮชแทก เครื่องหมายอ้างถึง (@) และชื่อบุคคลที่ถูกอ้างถึง ตัวอักษรหรือตัวเลขที่ปรากฏติดกันมากกว่า 3 ตัวขึ้นไปรวมถึงอักขระพิเศษต่าง ๆ ในการศึกษาแบ่งการทดลองออกเป็นสองส่วน ส่วนที่หนึ่งเป็นส่วนการประมวลผลโดยเครื่อง ในส่วนนี้ข้อความแต่ละข้อความจะถูกแบ่งเป็นคำ และแปลงให้อยู่ในโมเดล bigram ซึ่งจะใช้ในการคำนวณความน่าจะเป็นของข้อความโดยใช้วิธีภาวะความควรจะเป็นสูงสุด (Maximum Likelihood Estimation) ในส่วนที่สองกำหนดให้บุคคลจำนวน 5 คนประเมินข้อความแต่ละข้อความว่าข้อความนั้นเป็นข้อความส่อเสียด ข้อความปกติ หรือไม่สามารถระบุได้ แล้วนำคะแนนประเมินมาหาคะแนนความน่าจะเป็นเฉลี่ย แล้วนำความน่าจะเป็นของข้อความที่ได้จากการคำนวณโดยเครื่องและคะแนนความน่าจะเป็นเฉลี่ยที่ได้จากการประเมินของมนุษย์มาตรวจสอบระดับความสัมพันธ์โดยใช้สหสัมพันธ์ของเพียร์สัน จากผลการทดลองพบว่าค่า P-Value มีค่าเป็น 0.015 ซึ่งสรุปได้ว่าความน่าจะเป็นของข้อความที่คำนวณโดยเครื่องมีความสัมพันธ์ไปในทิศทางเดียวกันกับการจำแนกข้อความส่อเสียดโดยมนุษย์

สาขาวิชา เทคโนโลยีสารสนเทศทางธุรกิจ ลายมือชื่อนิสิต .....

ปีการศึกษา 2559 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

5781506626 : MAJOR INFORMATION TECHNOLOGY IN BUSINESS
KEYWORDS: SARCASM CLASSIFICATION / PHRASE PROBABILITY / NATURAL LANGUAGE PROCESSING / SOCIAL NETWORKS / TWITTER

KASIDECH TAPANG: SARCASM CLASSIFICATION IN TWITTER USING PROBABILITY OF TWEETS. ADVISOR: ASST. PROF. JANJAO MONGKOLNAVIN, Ph.D., 71 pp.

Sarcasm is one of the issues in Natural Language Processing since it inverts the real sentiment of a phrase; from positive to negative. This study proposes an approach to classify sarcastic phrases, on the topic of one telecommunication service provider in Thailand gathered from Twitter using Advance Search API. The data consists of 4,027 phrases, from 25th of January 2010 to 9th of June 2016. The phrases will be preprocessed by removing duplication, URL, hashtag as well as its content, mention (@) including the users, characters or numbers repeated more than 3 times consecutively. The experiment consists of two parts, phrase probability estimation by machine and by a group of five people. For the machine part, each phrase segmented into words, which are converted into a bigram model. The phrase probability is calculated from the bigram model using Maximum Likelihood Estimation. For the human part, each person rates each phrase whether it is sarcastic, typical or uncertain, then the average score is computed for each phrase. The relationship of the results from both parts is measured by using Pearson's Correlation Test. The test shows that P-Value is 0.015 which can be concluded that they are correlated in the same direction.

Field of Study: Information Technology in Business Student's Signature .....
Advisor's Signature .....
Academic Year: 2016

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลงได้ด้วยดี เนื่องจากได้รับความกรุณาอย่างสูงจากผู้ช่วยศาสตราจารย์ ดร. จันทร์เจ้า มงคลนาวิน อาจารย์ที่ปรึกษางานวิจัยที่กรุณาให้คำแนะนำปรึกษาตลอดจนปรับปรุงแก้ไขข้อบกพร่อง ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยตระหนักถึงความตั้งใจจริงและความทุ่มเทของท่าน และกราบขอบพระคุณเป็นอย่างสูงไว้

ขอกราบขอบพระคุณบิดา มารดา ที่ให้คำปรึกษาในเรื่องต่าง ๆ รวมทั้งเป็นกำลังใจที่ดีให้ข้าพเจ้าเสมอมา

ขอขอบคุณศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ซึ่งให้ความอนุเคราะห์ในส่วนของฐานข้อมูลคำศัพท์ และโปรแกรมตัดคำ

และสุดท้ายนี้ ขอขอบคุณผู้มีพระคุณทุกท่านที่ได้กล่าวมาข้างต้น รวมถึงอีกหลาย ๆ ท่านที่อาจไม่ได้กล่าวถึงมา ณ โอกาสนี้

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การประมวลผลภาษาธรรมชาตินั้นมีความสำคัญต่อการประมวลผลด้วยคอมพิวเตอร์เป็นอย่างมาก เนื่องจากข้อมูลที่มนุษย์ใช้กันอยู่ในชีวิตประจำวันนั้นมักจะอยู่ในรูปแบบของภาษาธรรมชาติเช่น ภาษาพูด ภาษาเขียน ดังนั้นหากคอมพิวเตอร์สามารถประมวลผลข้อมูลซึ่งอยู่ในรูปแบบของภาษาธรรมชาติได้อย่างถูกต้องและตรงตามความหมายที่แท้จริง จะส่งผลให้เราสามารถเข้าถึงและใช้ประโยชน์จากข้อมูลจำนวนมหาศาลที่มีอยู่บนโลกนี้ได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ในปัจจุบันการประมวลผลภาษาธรรมชาติได้ถูกนำมาใช้ในด้านต่าง ๆ มากมาย เช่น การใช้งานในด้านแปลภาษาโดยแปลจากภาษาหนึ่งไปยังภาษาอื่น ๆ (Machine Translation) การย่อสรุปความซึ่งสามารถสรุปใจความสำคัญหรือหัวข้อที่เกี่ยวข้องของบทความออกมาในรูปแบบภาษาธรรมชาติได้ นอกจากนี้การประมวลผลภาษาธรรมชาติยังสามารถช่วยในการจำแนกบทความออกเป็นหมวดต่าง ๆ เพื่อให้ง่ายต่อการจัดเก็บและประมวลผลในลำดับถัดไปอีกด้วย อย่างไรก็ตามคุณลักษณะของภาษาจัดว่าเป็นส่วนสำคัญที่ทำให้เทคนิคการประมวลผลภาษาธรรมชาติของภาษาหนึ่งอาจไม่เหมาะสมที่จะนำมาใช้กับอีกภาษาหนึ่ง เช่น ภาษาไทยซึ่งถูกจัดอยู่ในภาษาประเภทไม่ตัดคำ (Haruechaiyasak, 2010) ซึ่งแตกต่างจากภาษาอื่น ๆ เช่น ภาษาอังกฤษ เนื่องจากในภาษาไทยไม่มีตัวอักษรใดที่แบ่งคำแต่ละคำออกจากกัน ทำให้ต้องมีกระบวนการตัดคำเบื้องต้นก่อนจึงจะสามารถนำข้อความดังกล่าวมาประมวลผลต่อได้

ในปัจจุบันที่เครือข่ายสังคมออนไลน์มีบทบาทสำคัญอย่างมากต่อชีวิตของมนุษย์ทั้งในด้านของการติดต่อสื่อสารระหว่างครอบครัว เพื่อน คนรู้จัก หรือแม้กระทั่งระหว่างลูกค้าและบริษัทซึ่งเป็นไปได้หลากหลายรูปแบบเช่น การจัดโครงการส่งเสริมการขายสินค้าและการให้บริการหลังการขายผ่านทางสื่อออนไลน์ เช่น เฟสบุ๊ค ฝ่ายบริการลูกค้าบนเครือข่ายสังคมออนไลน์พันทิป หรือแม้กระทั่งการให้ข้อเสนอแนะและการติชมผ่านทางทวิตเตอร์ ทำให้ผู้ให้บริการสามารถรับทราบความคิดเห็นของลูกค้าได้โดยทันที และสามารถนำข้อเสนอแนะดังกล่าวไปปรับปรุงคุณภาพของผลิตภัณฑ์และการให้บริการให้ดียิ่งขึ้น ซึ่งจัดว่าเป็นการสร้างข้อได้เปรียบทางการแข่งขันได้อย่างมาก

อย่างไรก็ตามเนื่องจากความคิดเห็นของผู้ใช้งานนั้นในบางครั้งอาจไม่ได้อยู่ในรูปแบบที่สามารถตีความได้ตรงตามตัวอักษรซึ่งส่งผลให้การประมวลผลภาษาธรรมชาติมีความคลาดเคลื่อนไปจากความหมายที่แท้จริง เช่น “ตั้งแต่เกิดมา ไม่เคยใช้สบู่อะไรที่ดีแบบนี้มาก่อนเลย” ผู้เขียนอาจไม่ได้หมายความว่าสบู่นี้ดี

อย่างที่พูด แต่เป็นการเขียนในเชิงส่อเสียดว่าสบู่ก้อนนี้ไม่มีคุณภาพ เป็นต้น ซึ่งหากคอมพิวเตอร์ไม่สามารถจำแนกได้ว่าความคิดเห็นนั้นเป็นความคิดเห็นจริงหรือเป็นความคิดเห็นในเชิงส่อเสียด จะส่งผลต่อการสรุปผลความคิดเห็นที่ผู้ใช้งานมีต่อประเด็นหรือเรื่องนั้น ๆ โดยเฉพาะอย่างยิ่งในกรณีความคิดเห็นเชิงส่อเสียดเป็นความคิดเห็นในเชิงลบอย่างมาก ในขณะที่คอมพิวเตอร์ประมวลผลออกมาเป็นความคิดเห็นเชิงบวกอย่างมาก ทำให้ข้อสรุปที่ได้จากการประมวลผลไม่ตรงตามความเป็นจริง งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเทคโนโลยีทางด้านการประมวลผลภาษาธรรมชาติและการเรียนรู้ของเครื่องมาใช้ในการจำแนกข้อความส่อเสียด (Sarcasm Classification) ออกจากข้อความปกติในภาษาไทยเพื่อสรุปความหมายที่แท้จริงของผู้เขียนข้อความซึ่งช่วยให้การวิเคราะห์ข้อความเป็นไปได้อย่างถูกต้องและแม่นยำมากยิ่งขึ้น

1.2 นิยามศัพท์

มีคำจำกัดความมากมายเกี่ยวกับข้อความส่อเสียดทั้งเช่น ในงานวิจัยของ Gonzalez (2011) ได้ให้คำจำกัดความของข้อความส่อเสียดว่า “การส่อเสียดหมายถึงการเปลี่ยนแปลงขั้วความคิดเห็นของข้อความจากความคิดเห็นขั้วบวกหรือลบไปยังทิศทางตรงกันข้าม” หรือในงานวิจัยของ Tungthamthiti (2014) ซึ่งให้คำจำกัดความว่า “การส่อเสียดเป็นรูปแบบหนึ่งของการสื่อสารซึ่งผู้ใช้ต้องการที่จะเยาะเย้ยบุคคลอื่นโดยใช้คำที่มีความหมายตรงกันข้ามกับสิ่งที่ผู้ใช้ต้องการจะสื่อ” หรือแม้กระทั่งในพจนานุกรมราชบัณฑิตยสถานซึ่งให้ความหมายว่า “คำประชดเหน็บแนม แกล้งทำให้เกินควรหรือพูดแดกดันเพราะความไม่พอใจ”[1] เป็นต้น จากคำจำกัดความของข้อความส่อเสียดของงานวิจัยและคำจำกัดความในพจนานุกรมที่กล่าวมาข้างต้นจะเห็นว่าคำจำกัดความดังกล่าวนั้นมีส่วนที่คล้ายคลึงกัน จึงสรุปเป็นคำนิยามสำหรับใช้ในการวิจัยนี้ว่า “ข้อความส่อเสียดคือข้อความที่มีความหมายในทิศทางตรงกันข้ามกับที่ผู้เขียนต้องการจะสื่อ” ซึ่งส่งผลให้การวิเคราะห์ความคิดเห็น (Sentiment Analysis) เกิดความผิดพลาดได้ เนื่องจากข้อความส่อเสียดให้ทิศทางความคิดเห็นตรงกันข้ามกับความหมายที่ผู้ใช้ต้องการ

เนื่องจากผลกระทบของข้อความส่อเสียดที่กล่าวมาข้างต้น จึงมีงานวิจัยบางส่วนซึ่งกล่าวถึงองค์ประกอบที่ใช้การจำแนกข้อความส่อเสียดออกจากข้อความปกติ โดยสามารถแบ่งออกได้เป็นสองกลุ่มใหญ่คือกลุ่มที่จำแนกโดยการใช้ลักษณะของภาษา และกลุ่มที่จำแนกโดยการใช้องค์ประกอบอื่น ๆ (Pt´cek, Habernal, and Hong, 2014; Pt´cek and Steinberger, 2014) โดยองค์ประกอบลักษณะของภาษาหมายถึง การปรากฏของคำในประโยครวมถึงคำที่มักจะปรากฏในข้อความส่อเสียด เช่น การที่ประโยคปรากฏคำว่า “ever” มักจะมีแนวโน้มที่จะเป็นประโยคส่อเสียดมากกว่าประโยคที่ไม่มีคำนี้ปรากฏ

[1] http://dictionary.sanook.com/search/dict-th-th-pleang/ประชด

นอกจากนี้แล้วยังมีการใช้ลักษณะอื่น ๆ ที่ไม่เกี่ยวกับคำศัพท์และการใช้ภาษา ได้แก่ จำนวนการใช้สัญลักษณ์พิเศษต่าง ๆ เช่น การใช้เครื่องหมายอัศเจรีย์ หรือแม้กระทั่งการใช้ตัวอักษรพิมพ์ใหญ่ในข้อความและจำนวนไอคอนแสดงอารมณ์ก็ถูกใช้เป็นองค์ประกอบในการจำแนกข้อความส่อเสียดเช่นกัน

อย่างไรก็ตามการประยุกต์เทคนิคข้างต้นจะใช้คำเฉพาะและสัญลักษณ์พิเศษมาประกอบการสรุปว่าข้อความเป็นข้อความปกติหรือข้อความส่อเสียดอาจประยุกต์ใช้ได้กับเฉพาะบางภาษา ซึ่งอาจขึ้นอยู่กับลักษณะและวัฒนธรรมของแต่ละภาษาและบุคคล เช่น ประโยค “This is the best thing I have EVER” จะมีแนวโน้มที่จะเป็นประโยคเชิงส่อเสียดเนื่องจากมีการใช้ตัวอักษรพิมพ์ใหญ่ในคำว่า “EVER” อย่างไรก็ตามลักษณะดังกล่าวจะไม่สามารถนำมาประยุกต์กับภาษาไทยได้ เนื่องจากภาษาไทยไม่มีลักษณะของการใช้ตัวอักษรพิมพ์เล็กและพิมพ์ใหญ่ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาถึงความเป็นไปได้ในการคำนวณความน่าจะเป็นของข้อความจากองค์ประกอบของคำที่ใช้เพื่อที่จะใช้ในการจำแนกข้อความส่อเสียดออกจากข้อความปกติรวมถึงศึกษาข้อจำกัดและปัจจัยที่เกี่ยวข้องโดยมีความเชื่อว่า ข้อความส่อเสียดน่าจะเป็นข้อความที่มีองค์ประกอบของคำที่ใช้ผิดแปลกไปจากองค์ประกอบของคำที่ใช้ในข้อความปกติธรรมดาทั่วไป

1.3 วัตถุประสงค์การวิจัย

1. เพื่อศึกษาและพัฒนาตัวแบบสำหรับจำแนกข้อความส่อเสียดออกจากข้อความปกติในภาษาไทยโดยใช้ข้อมูลความคิดเห็นของผู้บริโภคเกี่ยวกับการบริการอินเทอร์เน็ตในประเทศไทยเป็นกรณีศึกษา

2. เพื่อศึกษาถึงปัจจัยต่าง ๆ ที่มีผลต่อประสิทธิภาพของการจำแนกข้อความส่อเสียดออกจากข้อความปกติในภาษาไทย

1.4 ประโยชน์ที่คาดว่าจะได้รับ

แนวทางในการพัฒนาตัวแบบสำหรับจำแนกข้อความส่อเสียดจากข้อความปกติในภาษาไทยจากข้อมูลความคิดเห็นของผู้บริโภคบนเครือข่ายสังคมออนไลน์ทวิตเตอร์

1.5 ขอบเขตการวิจัย

สำหรับการเรียนรู้ และการทดสอบ ตัวแบบการวิเคราะห์ประโยคแสดงความคิดเห็นเกี่ยวกับการให้บริการอินเทอร์เน็ตในประเทศไทย ผู้วิจัยศึกษาจากข้อความแสดงความคิดเห็นเกี่ยวกับบริษัทผู้ให้บริการอินเตอร์เน็ตรายหนึ่งในประเทศไทยโดยเป็นข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ เนื่องจากข้อมูลการ

ใช้บริการอินเตอร์เน็ตนั้นเป็นข้อมูลที่สาธารณะที่ทุกคนสามารถเข้าถึงได้ โดยเก็บข้อมูลตัวอย่างที่เกี่ยวข้องกับการใช้บริการอินเตอร์เน็ตของผู้ให้บริการรายนั้นระหว่างวันที่ 25 มกราคม 2553 ถึง 9 มิถุนายน 2559 จำนวน 4,027 ข้อความ ซึ่งประกอบไปด้วยข้อความปกติจำนวน 3,913 ข้อความ ข้อความที่ผู้วิจัยไม่แน่ใจว่าเป็นข้อความส่อเสียดหรือไม่ 6 ข้อความ และ ข้อความส่อเสียดจำนวน 108 ข้อความ[2]

[2] ข้อความส่อเสียดจำนวน 108 ข้อความ ที่จำแนกโดยผู้วิจัยโดยใช้เกณฑ์การจำแนกโดยดูจากแฮชแทก #ประชด และวิจารณญาณของผู้วิจัย เป็นหลัก

บทที่ 2
การทบทวนวรรณกรรม

ในบทนี้จะเป็นการนำเสนอวรรณกรรมในอดีต (Literature Review) ที่เกี่ยวข้อง เพื่อชี้ให้เห็นถึงการศึกษาหรือสำรวจในประเด็นการพัฒนาระบบตรวจหาข้อความส่อเสียดบนทวิตเตอร์ (Sarcasm Detection System on Twitter) ซึ่งประกอบด้วยหัวข้อย่อยคือ (1) การประมวลผลภาษาธรรมชาติ (2) เครือข่ายสังคมออนไลน์ (3) รูปแบบของข้อมูลที่ได้จากทวิตเตอร์ (4) การวิเคราะห์ความคิดเห็น (5) การตรวจสอบข้อความส่อเสียด และ (6) ข้อจำกัดของการวิจัยที่ผ่านมา เพื่อชี้ให้เห็นถึงความสำคัญของการวิเคราะห์และจำแนกข้อความส่อเสียดซึ่งส่งผลต่อความถูกต้องในการวิเคราะห์ความคิดเห็นที่ผู้ใช้เครือข่ายสังคมออนไลน์มีต่อสินค้าและบริการ

2.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

ในปัจจุบันข้อมูลข่าวสารต่าง ๆ บนโลกอินเตอร์เน็ตล้วนมีความสำคัญอย่างมากสำหรับองค์กร เนื่ององค์กรสามารถนำจากข้อมูลต่าง ๆ เช่น ผลตอบรับของสินค้าและบริการ (Feedback) มาใช้เพื่อปรับปรุงคุณภาพของสินค้าและบริการภายในองค์กรให้ดียิ่งขึ้น (Fundin and Elg, 2010) ข้อมูลดังกล่าวเป็นข้อมูลที่ไม่มีโครงสร้างที่แน่นอน (Unstructured Data) จึงจำเป็นต้องนำเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาใช้เพื่อแปลงข้อมูลให้อยู่ในรูปแบบที่คอมพิวเตอร์เข้าใจและสามารถนำไปใช้ในการประมวลผลอื่น ๆ ต่อได้ (Bao et al., 2012) โดยมีขั้นตอนหลักคือ 1. การจัดการข้อมูลก่อนการวิเคราะห์ (Data Preprocessing) 2. การวิเคราะห์คำ (Lexical Analysis) และ 3. การวิเคราะห์วากยสัมพันธ์ (Syntax Analysis) เพื่อนำไปสร้างวากยสัมพันธ์ (Parse Tree) (Sibarani et al., 2013)

นอกจากนี้ยังมีการใช้การประมวลผลภาษาธรรมชาติในด้านต่าง ๆ เช่น ใช้เพื่อสรุปใจความสำคัญของบทความ ใช้เพื่อหาประโยคหรือคำที่เกี่ยวข้อง ใช้เพื่อจำแนกชนิดของบทความ หรือใช้เพื่อแปลภาษาจากภาษาหนึ่งเป็นอีกภาษาหนึ่งเป็นต้น เนื่องจากศาสตร์ในการประมวลผลธรรมชาตินั้นเกี่ยวเนื่องกับเรื่องของปัญญาประดิษฐ์เป็นหลัก ซึ่งการพัฒนาเกี่ยวกับการประมวลผลภาษาธรรมชาติยังคงดำเนินต่อไปอย่างไม่หยุดยั้ง และมีความเชื่อว่าในอนาคตเครื่องจักรหรืออุปกรณ์จะสามารถเข้าใจและเรียนรู้ข้อมูลและสารสนเทศในรูปแบบของภาษาธรรมชาติได้ดีมากยิ่งขึ้นตามลำดับ (Chopra, Prashar, & Sain, 2013)

2.1.1 การประมวลผลภาษาธรรมชาติในภาษาต่างประเทศ (Natural Language Processing in Foreign Language)

เนื่องจากศักยภาพและประโยชน์ที่ได้รับจากการประมวลผลภาษาธรรมชาติ ทำให้การทำเหมืองความคิดเห็น (Opinion Mining) และการวิเคราะห์ความคิดเห็น (Opinion Analysis) ได้รับความสนใจเป็นอย่างมาก (Wei et al., 2009; Baccianella et al., 2010; Osherenko, 2010) โดยส่วนมากแล้วงานวิจัยในส่วนนี้มักจะเกี่ยวกับการวิเคราะห์บทวิจารณ์ของสินค้าว่าเป็นไปในทางบวกหรือทางลบ โดยมีตั้งแต่การวิเคราะห์ในระดับเอกสาร (Moraes et al., 2013; Pang et al., 2002) ไปจนถึงการวิเคราะห์ในระดับประโยค (Kim and Hovy, 2004; Wilson et al. 2009) ตัวอย่างเช่น หากมีการเขียนบทวิจารณ์ของสินค้าใดสินค้าหนึ่ง ระบบก็จะประมวลผลข้อความภายในบทวิจารณ์ของสินค้านั้นว่าผู้เขียนมีความคิดเห็นเกี่ยวกับสินค้านั้น ๆ เป็นไปในทางบวกหรือทางลบ เป็นต้น

นอกจากการใช้การประมวลผลภาษาธรรมชาติเพื่อวิเคราะห์บทวิจารณ์ของสินค้าที่กล่าวไปข้างต้นแล้ว ยังมีการใช้การประมวลผลภาษาธรรมชาติเพื่อการวิเคราะห์ทางการแพทย์ (Rodrigues et al., 2015) ซึ่งเป็นการศึกษาพฤติกรรมและอารมณ์ของผู้ป่วยโรคมะเร็งในเครือข่ายสังคมออนไลน์ของประเทศบราซิล ซึ่งการศึกษาดังกล่าวจะสำรวจการแสดงความคิดเห็นของผู้ป่วยซึ่งเขียนอยู่ในรูปแบบภาษาโปรตุเกสและวิเคราะห์การแสดงความคิดเห็นดังกล่าวว่าอารมณ์หรือความคิดเห็นของผู้ป่วยนั้นเป็นไปในทิศทางบวก ลบ หรือเป็นกลางซึ่งใช้การวิเคราะห์แบบอิงจากคลังศัพท์ในระดับประโยค นอกจากนี้ยังมีการใช้การวิเคราะห์ภาษาธรรมชาติในภาษาอื่น ๆ เช่น ภาษาเช็ก (Pt'cek & Steinberger, 2014) หรือในภาษาอาราบิก (El-Orfali, 2014; Hsinchun & Salem, 2008) โดยภาษาที่มีการศึกษาด้านการประมวลผลภาษาธรรมชาติอย่างแพร่หลาย คือ ภาษาอังกฤษ

2.1.2 การประมวลผลภาษาธรรมชาติในภาษาไทย (Natural Language Processing in Thai Language)

มีงานวิจัยที่ทำการศึกษาการประมวลผลภาษาธรรมชาติในภาษาไทย (Chamlertwat, et al., 2011; Chamlertwat, et al., 2012; Haruechaiyasak, et al., 2013) เช่น การทำเหมืองความคิดบนไมโครบล็อก (Chamlertwat et al., 2011) ซึ่งเป็นการทำเหมืองข้อมูลเพื่อวิเคราะห์จากความรู้สึกของผู้บริโภคจากไมโครบล็อกแบบอัตโนมัติ โดยแบ่งทัศนคติออกเป็นระดับต่าง ๆ เช่น “เห็นด้วย” หรือ “เห็นด้วยอย่างยิ่ง” โดยใช้ความคิดเห็นจากผู้บริโภคเกี่ยวกับสมาร์ทโฟนเป็นกรณีศึกษาและเก็บรวบรวมข้อมูลจากทวิตเตอร์

นอกจากนี้ยังมีการทำวิจัยเกี่ยวกับการวิเคราะห์ความคิดเห็นของผู้บริโภคต่อผู้ให้บริการเครือข่ายโทรศัพท์มือถือในประเทศไทย (Haruechaiyasak et al., 2013) โดยมีการจำแนกเจตนาของหัวข้อต่าง ๆ

ออกเป็น 4 ประเภท คือ ประกาศ ขอร้อง คำถาม และความคิดเห็น จากนั้นจึงวิเคราะห์ทัศนคติของข้อความดังกล่าวว่ามีความคิดเห็นในเชิงบวกหรือเชิงลบต่อผู้ให้บริการเครือข่ายโทรศัพท์มือถือ การศึกษาดังกล่าวพบว่าการวิเคราะห์ความคิดเห็นของผู้ใช้งานเครือข่ายโดยใช้ฐานคำศัพท์ทั่วไป (General Term) และฐานคำศัพท์ชี้นำ (Clue Term) นั้นให้ความแม่นยำสูงสุดถึงร้อยละ 91.64

ลักษณะเฉพาะที่เห็นได้ชัดในการประมวลผลภาษาธรรมชาติในภาษาไทยนั้นคือภาษาไทยถูกจัดอยู่ในประเภทภาษาที่ไม่ตัดคำ (Unsegmented Language) (Haruechaiyasak, 2010) กล่าวคือภาษาไทยเป็นภาษาที่ไม่มีการใช้ตัวอักขระใด ๆ ในการบ่งบอกขอบเขตของคำอย่างชัดเจนหากเปรียบกับภาษาต่างประเทศอื่น ๆ ตัวอย่างเช่นประโยค “I go to school” ในภาษาอังกฤษ คำแต่ละคำจะถูกแบ่งแยกจากกันอย่างชัดเจนเนื่องจากมี “สเปซบาร์” เป็นตัวอักษรที่ใช้ในการแบ่งคำแต่ละคำออกจากกัน แต่ในภาษาไทย เช่น ประโยค “ฉันไปโรงเรียน” นั้นไม่มีการแบ่งคำแต่ละคำออกจากกันอย่างชัดเจน โดยจะต้องอาศัยเทคนิคการตัดคำในการบอกขอบเขตของคำซึ่งยังไม่มีเทคนิคที่ให้ความถูกต้องได้โดยสมบูรณ์แบบ จึงทำให้เกิดปัญหาในเรื่องของคำที่ไม่รู้จัก และคำกำกวม ตัวอย่างเช่น

ข้อความ “รบกวนออกมาอธิบายไวไวเถอะข่ะ นี่ขี้เกียจด่าแล้ว เน็ท 4G แย่มากๆ ต้องใช้ไวไฟคนข้างบ้านอยู่ กลัวเค้าปิดโมเด็ม”

สามารถตัดคำโดยใช้ LexTo[3] ซึ่งเป็นเครื่องมือที่ใช้ในการตัดคำโดยใช้เทคนิคตัดคำที่ยาวที่สุด (Longest Matching) โดยสามารถตัดคำได้เป็น

“|รบกวน|ออกมา|อธิบาย|ไวไว|เถอะ|ข่ะ| |นี่|ขี้เกียจ|ด่า|แล้ว| |เน็ท| |4|G| |แย่|มากๆ| |ต้อง|ใช้|ไวไฟ| |คน|ข้าง|บ้าน|อยู่| |กลัว|เค้า|ปิด|โมเด็ม|”

โดยคำที่ใช้ *ตัวเอียง* หมายถึงคำที่รู้จัก คำที่ ขีดเส้นใต้ หมายถึงคำกำกวม และคำที่ ขีดเส้นใต้สองเส้นหมายถึงคำไม่รู้จัก ซึ่งแสดงให้เห็นถึงข้อจำกัดในเรื่องคำไม่รู้จักและคำกำกวมสำหรับการตัดคำในภาษาไทย

[3] http://www.sansarn.com/lexto/

2.1.3 การตัดคำในภาษาไทย (Thai Word Segmentation)

การตัดคำ คือ การแบ่งข้อความออกเป็นหน่วยเล็ก ๆ ที่ต่อเนื่องกันซึ่งเรียกว่าหน่วยคำ (Morpheme) โดยใช้ลักษณะของการรู้จำ (Recognition) เพื่อหาขอบเขตของแต่ละหน่วยคำ ความยากง่ายหรือวิธีการที่ใช้ในการตัดคำนั้นขึ้นอยู่กับลักษณะเฉพาะของภาษานั้น ๆ กล่าวคือหากภาษาต่างกันและมีลักษณะเฉพาะไม่เหมือนกันย่อมใช้วิธีตัดคำที่แตกต่างกัน (Charoenpornsawat, 1999)

การตัดคำในภาษาไทยได้รับการพัฒนาจากหน่วยงานวิจัยต่าง ๆ ทั้งของภาครัฐและเอกชน ซึ่งแต่ละวิธีย่อมมีข้อดีและข้อเสียแตกต่างกัน เช่น ความเร็วที่ใช้ในการตัดคำ ความถูกต้อง หรือเนื้อที่ที่ใช้ในการเก็บข้อมูลคลังคำศัพท์ โดยเทคนิคที่ประยุกต์ใช้ในการตัดคำในภาษาไทยสามารถแบ่งออกเป็น 3 วิธีหลัก ดังต่อไปนี้ (Haruechaiyasak, 2010)

1. หลักการตัดคำโดยใช้กฎไวยากรณ์ทางภาษา (Rule-Based Technique) วิธีนี้เป็นการตัดคำโดยใช้กฎไวยากรณ์ของภาษา โดยระบุกฎเกณฑ์ของภาษาลงไป เช่น การใช้สัญลักษณ์เว้นวรรคในภาษาไทยระหว่างประโยคนั้นอาจแสดงถึงการจบประโยคของข้อความนั้น ๆ หรือแม้กระทั่งการขึ้นย่อหน้าใหม่ซึ่งแสดงให้เห็นถึงการสิ้นสุดของข้อความ เป็นต้น

วิธีตัดคำโดยใช้กฎไวยากรณ์ทางภาษานั้นมีข้อจำกัดมาก คือ ผลที่ได้จากการตัดคำนั้นอาจเป็นกลุ่มคำที่สามารถแบ่งแยกย่อยออกไปได้อีกซึ่งทำให้ความแม่นยำของการตัดคำโดยใช้เทคนิคนี้ค่อนข้างต่ำ แต่วิธีนี้มีข้อดีคือสามารถทำงานได้อย่างรวดเร็วและไม่ต้องสำรองข้อมูลฐานคำศัพท์ไว้ในหน่วยความจำ

2. หลักการตัดคำโดยใช้พจนานุกรม (Dictionary–Based Technique) วิธีนี้เป็นการตัดคำโดยการเก็บคำศัพท์ภาษาไทยไว้ในพจนานุกรม จากนั้นนำข้อความไปค้นหาและเปรียบเทียบกับฐานคำศัพท์ในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรมีการตัดคำในตำแหน่งใด การตัดคำโดยใช้พจนานุกรมในภาษาไทยนั้นมีขั้นตอนการทำงานหลักอยู่ 2 ขั้นตอน คือ ขั้นตอนแรกจะทำการตัดคำโดยเทียบจากพจนานุกรม และขั้นตอนที่สองจะเทียบคำที่ได้จากขั้นตอนแรกว่าสามารถตัดคำได้อีกหรือไม่

วิธีนี้นอกจากมีความแม่นยำที่ค่อนข้างสูงแล้ว ยังสามารถปรับปรุงฐานคำศัพท์ใหม่ ๆ ได้โดยการปรับปรุงหรือเพิ่มเติมคำศัพท์ใหม่ ๆ เข้าไปยังฐานข้อมูล แต่วิธีนี้ค่อนข้างสิ้นเปลืองหน่วยความจำในการเก็บข้อมูลคำศัพท์ค่อนข้างมาก

หลักการตัดคำโดยใช้พจนานุกรมนี้ยังถูกพัฒนาไปยังหลากหลายรูปแบบเช่น การใช้วิธีการเทียบคำที่ยาวที่สุด (Longest Matching) โดยเป็นวิธีการตัดคำโดยเทียบจากพจนานุกรมโดยเปรียบเทียบคำที่ยาวที่สุดก่อน ยกตัวอย่างเช่น “ฉันไปโรงหนัง” ข้อความนี้เมื่อไม่สามารถเทียบจากพจนานุกรมได้ก็จะลดลงเหลือ “ฉันไปโรงหนั” -> “ฉันไปโรงห” -> “ฉันไปโรง” จนได้สายอักขระ “ฉัน” ซึ่งสามารถเทียบคำในพจนานุกรมได้จึงตัดเป็นคำแรกของข้อความ จากนั้นทำการย้อนกลับเพื่อหาคำอื่น ๆ ดังตารางที่ 2.1

ตาราง 2.1 ผลการตัดคำทั้งหมดด้วยวิธีเทียบคำที่ยาวที่สุด

ส่วนของคำที่ยาวที่สุดที่ตัดได้	ส่วนหลังที่เหลือที่จุดย้อนกลับ
ฉัน	ไปโรงหนัง
ไป	โรงหนัง
โรงหนัง	-

วิธีนี้มีข้อจำกัดเนื่องจากลักษณะของการพยายามที่จะตัดคำที่ยาวที่สุด ซึ่งทำให้บางครั้งมีการเลือกคำที่ยาวเกินไปและส่งผลให้การตัดคำมีความผิดเพี้ยนเช่น “ไปหามเหสี” (Go to see the queen.) หากใช้วิธีการตัดคำโดยเทียบคำที่ยาวที่สุดจำได้ “ไป | หาม | เห | สี” ซึ่งจะพบคำว่า “หาม” ก่อนคำว่า “หา” เสมอ จึงทำให้การตัดคำดังกล่าวไม่สื่อความหมายใด ๆ

อีกหนึ่งวิธีที่นิยมใช้สำหรับการตัดคำโดยใช้พจนานุกรมคือการตัดคำโดยเลือกแบบเหมือนมากที่สุด (Maximal Matching) ซึ่งเป็นวิธีการตัดคำให้ได้จำนวนคำที่เป็นไปได้น้อยที่สุด ซึ่งพัฒนาโดย วิรัช ศรเลิศล้ำวาณิช (2536) โดยวิธีนี้จะแก้ปัญหาที่ปรากฏในเทคนิคการเทียบคำที่ยาวที่สุดที่กล่าวมาข้างต้น โดยขั้นตอนแรกจะใช้วิธีของการเทียบคำที่ยาวที่สุดก่อนจากนั้นทำการย้อนกลับ (Backtracking) ทีละคำเพื่อหาต้นทุน (Cost) หรือจำนวนคำที่สามารถเป็นไปได้ และเลือกตัดคำตามทางเลือกที่มีจำนวนคำที่เป็นไปได้น้อยที่สุด ยกตัวอย่างเช่น “ไปหามเหสี” นั้นหากใช้เทคนิคการตัดคำแบบเทียบคำที่ยาวที่สุดจะได้ “ไป | หาม | เห | สี” จากนั้นจะทำการย้อนกลับโดยเริ่มตั้งแต่คำที่สองคือ “หาม” ซึ่งสามารถแบ่งแยกเป็นคำอื่นได้อีกนั่นคือ “หา” จากนั้นจะสร้างทางเลือกที่เป็นไปได้ทั้งหมดดังในตารางที่ 2.2 หลังจากนั้นจึงหาค่าต้นทุนในแต่ละทางเลือกที่เป็นไปได้ จากนั้นทำการเลือกความเป็นไปได้ที่มีค่าต้นทุนต่ำที่สุดซึ่งในที่นี้คือ “ไป | หา | มเหสี” เป็นต้น

ตาราง 2.2 ความเป็นไปได้ทั้งหมดของการตัดคำทั้งหมดด้วยวิธีเลือกแบบเหมือนมากที่สุด

ความเป็นไปได้	จำนวนคำที่เป็นไปได้
ไป | หาม | เห | สี	4
ไป | หา | มเห | สี	4
ไป | หา | มเหสี	3

3. หลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Technique) วิธีนี้เป็นการใช้คลังข้อมูลในการตัดคำซึ่งใช้หลักวิธีการทางสถิติเข้ามาใช้ในการประมวลผล โดยใช้คลังข้อมูลทางภาษา (Corpus) เป็นฐานความรู้ในการตัดคำ วิธีนี้ไม่จำเป็นต้องเก็บฐานข้อมูลคำศัพท์ไว้ในหน่วยความจำและสามารถแก้ปัญหาคำศัพท์ที่ไม่รู้บางส่วนได้โดยการเรียนรู้จากคลังข้อมูลทางภาษาที่ถูกป้อนเข้าไป ข้อเสียของวิธีนี้คือต้องใช้คลังข้อมูลทางภาษาเป็นจำนวนมากและใช้เวลาในการเรียนรู้ที่ค่อนข้างนานซึ่งขึ้นอยู่กับอัลกอริทึมที่ใช้

2.2 เทคนิคที่ใช้ในการวิเคราะห์ความคิดเห็น

การวิเคราะห์ความคิดเห็นคือการใช้ความรู้เกี่ยวกับการประมวลผลภาษาธรรมชาติและการวิเคราะห์ข้อความเพื่อที่จะระบุถึงข้อมูลหรือความคิดเห็นที่ผู้เขียนข้อความต้องการจะสื่อถึง ซึ่งโดยส่วนมากแล้วการวิเคราะห์ความคิดเห็นนั้นจะถูกนำไปใช้กับเครือข่ายสังคมออนไลน์ในด้านการตลาดและการบริการลูกค้า[4] ตัวอย่างเช่น การวิเคราะห์ความคิดเห็นของผู้ใช้งานต่อตัวสินค้าว่าผู้ใช้งานมีความพึงพอใจกับตัวสินค้าในด้านต่าง ๆ เช่น ราคา คุณภาพ หรือความยากง่ายในการใช้งานว่าเป็นไปในทิศทางใดโดยมีจุดประสงค์เพื่อนำข้อมูลที่ได้มาปรับปรุงสินค้าและบริการให้ดียิ่งขึ้น

การวิเคราะห์ความคิดเห็นนั้นมีเทคนิคการวิเคราะห์ที่หลากหลายโดยขึ้นอยู่กับรูปแบบของข้อความที่ทำการวิเคราะห์หรือผลลัพธ์ที่ผู้ทำการวิเคราะห์ต้องการโดยสามารถแบ่งเทคนิคการวิเคราะห์ความคิดเห็นได้เป็นหลัก ๆ อยู่ทั้งสิ้น 4 รูปแบบ (Feldman, 2013) ได้แก่ การวิเคราะห์ความคิดเห็นในระดับเอกสาร การวิเคราะห์ความคิดเห็นในระดับประโยค การวิเคราะห์ความคิดเห็นโดยอิงจากลักษณะ และการวิเคราะห์ความคิดเห็นเชิงเปรียบเทียบ

2.2.1 Document-Level Sentiment Analysis

การวิเคราะห์ความคิดเห็นในระดับเอกสารนั้นเป็นวิธีการวิเคราะห์ความคิดเห็นของผู้เขียนบทความโดยยอมรับว่าในบทความหรือข้อความนั้นมีความคิดเห็นหลักของผู้เขียนบทความเพียงคนเดียว โดยมีการแบ่งออกเป็นสองวิธีคือการเรียนรู้แบบมีผู้สอน (Supervised learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

เทคนิคการเรียนรู้แบบมีผู้สอนนั้นจะเป็นการสร้างโมเดลการวิเคราะห์จากข้อมูลสอน (Training data set) โดยจะต้องมีการระบุผลที่ต้องการหรือประเภทไว้ก่อน เช่นหากมีค่าของคลาสที่ต้องการสองคลาสคือ Positive และ Negative เทคนิคการเรียนรู้แบบมีผู้สอนนั้นจะต้องใช้ข้อมูลสอนให้เพียงพอกับทั้งสองคลาสดังกล่าว โดยมีอัลกอริทึมที่ใช้ในการแบ่งกลุ่มทั่ว ๆ ไปเช่น Support Vector Machine (SVM) , Naïve Bayes, Logistic Regression หรือ K-Nearest Neighbor (KNN) เป็นต้น

เทคนิคการเรียนรู้แบบไม่มีผู้สอนนั้นจะเป็นการจำแนกกลุ่มของข้อมูล โดยเทคนิคการเรียนรู้แบบไม่มีผู้สอนนั้นจะแตกต่างจากการเรียนรู้แบบมีผู้สอนคือจะไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน แต่จะจำแนกกลุ่มของข้อมูลตามความเหมือน และความแตกต่างกันของตัวแปลที่ใช้ศึกษา

---

[4] https://en.wikipedia.org/wiki/Sentiment_analysis สืบค้นข้อมูล ณ วันที่ 10 เดือน ธันวาคม พ.ศ. 2558

2.2.2 Sentence-Level Sentiment Analysis

การวิเคราะห์ความคิดเห็นระดับประโยคนั้นแตกต่างจากการวิเคราะห์ความคิดเห็นในระดับเอกสาร ตรงที่ในข้อความหรือบทความนั้นอาจมีความคิดเห็นของผู้เขียนมากกว่าหนึ่งจุดในหัวข้อเดียวกันเช่น ความเร็วในการใช้งานอินเตอร์เน็ต โดยถึงแม้จะเป็นอินเตอร์เน็ตจากผู้ให้บริการเดียวกัน แต่บางช่วงเวลาอาจมีความเร็วในการใช้งานลดลงเป็นต้น ส่งผลให้การแสดงความคิดเห็นของผู้ใช้เปลี่ยนแปลงไปตามช่วงเวลาที่ต่างกัน การวิเคราะห์ในระดับนี้จะยอมรับว่าในแต่ละประโยคนั้นสามารถมีความคิดเห็นอยู่ในตัวประโยคจึงทำให้ต้องแยกการวิเคราะห์ออกเป็นระดับประโยค

2.2.3 Aspect-Based Sentiment Analysis

ในสองวิธีแรกนั้นสามารถใช้งานได้ในกรณีที่ความคิดเห็นนั้นอ้างอิงถึงเรื่องเดียวกัน แต่ในกรณีที่มีการเสนอความคิดเห็นในหลากหลายมุมมอง เช่น การวิจารณ์สินค้าหรือบริการต่าง ๆ นั้นจะใช้การวิเคราะห์ความคิดเห็นโดยอิงจากลักษณะ ซึ่งการวิเคราะห์วิธีนี้จะใช้กับบทความที่มีการนำเสนอความคิดเห็นของผู้เขียนมากกว่าหนึ่งจุดในหลาย ๆ มุมมองเช่น การวิจารณ์เครื่องสำอางประเภทน้ำหอม โดยผู้วิจารณ์สามารถวิจารณ์ได้มากกว่าหนึ่งมุมมอง คือ ด้านความหอม ด้านราคา ด้านกลิ่นที่ติดทนนาน โดยในแต่ละมุมมองนั้นอาจให้ความเห็นได้แตกต่างกันเป็นต้น วิธีนี้จะมีข้อดีแตกต่างกับการวิเคราะห์ความคิดเห็นในระดับเอกสารและระดับประโยคซึ่งจะทำให้ทราบถึงคุณสมบัติของสินค้าและบริการในด้านต่าง ๆ ไม่ใช่เพียงแค่สินค้าหรือบริการดังกล่าวนั้นดีหรือไม่ดีเท่านั้น

ในหัวข้อของการวิเคราะห์ความคิดเห็นโดยอิงจากลักษณะนั้นมีการศึกษาเพื่อใช้ในการวิเคราะห์หาจุดอ่อนของผลิตภัณฑ์ในเว็บไซท์ (Zhang, Xu, & Wan, 2012) ซึ่งเป็นการวิเคราะห์ความคิดเห็นโดยใช้คุณลักษณะของภาษาเป็นหลัก โดยเริ่มจากการหาคำศัพท์ที่ใช้บ่อยจากนั้นนำมาจัดเป็นหัวข้อต่าง ๆ เช่น ความหอม ความชุ่มชื่น และความสะอาดนั้นจะถูกจัดอยู่ในหมวดหมู่ของคุณสมบัติของผลิตภัณฑ์ ส่วนราคาและการโฆษณานั้นจะถูกจัดอยู่ในหมวดหมู่ของการตลาดเป็นต้น จากนั้นในแต่ละหัวข้อย่อย เช่น ราคา จะถูกนำมาจำแนกออกเป็นระดับต่าง ๆ ซึ่งขึ้นอยู่กับคลังคำศัพท์ คำที่ให้ความหมายเชิงปฏิเสธ คำที่ให้ความหมายตรงกันข้าม และคำวิเศษณ์ที่แสดงถึงระดับต่าง ๆ และนำมาเปรียบเทียบกับคู่แข่งเพื่อหาข้อดีและข้อเสียในตัวผลิตภัณฑ์เมื่อเปรียบเทียบกับคู่แข่ง

2.2.4 Comparative Sentiment Analysis

ในบางครั้งผู้เขียนบทวิจารณ์อาจไม่ให้ความคิดเห็นเกี่ยวกับสินค้าโดยตรงแต่จะเป็นการให้ความคิดเห็นในรูปแบบของการเปรียบเทียบระหว่างสินค้าสองชนิดเช่น “I drove the Honda Civic, it does not handle better than the TSX, not even close” ซึ่งเทคนิคการวิเคราะห์ความคิดเห็นเชิงเปรียบเทียบนั้นจะวิเคราะห์ข้อความดังกล่าวเพื่อหาว่าสินค้าชนิดใดดีกว่ากัน

มีงานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ความคิดเห็นแบบเปรียบเทียบ (Jindal & Liu, 2006) ซึ่งพบว่าคำที่ใช้ในการเปรียบเทียบนั้นไม่กี่คำก็เพียงพอที่จะครอบคลุมการแสดงความคิดเห็นเชิงเปรียบเทียบถึงร้อยละ 98 เช่น more, less หรือคำที่ลงท้ายด้วย –er –est ซึ่งเป็นคำในกลุ่มที่ใช้ในการเปรียบเทียบหรือคำคุณศัพท์หรือกริยาวิเศษณ์แสดงการเปรียบเทียบ (Superlative)

เนื่องจากคำดังกล่าวนั้นมี ค่าความแม่นยำ (Precision) ที่สูงมากแต่ ค่าความระลึก (Recall) ค่อนข้างต่ำจึงใช้การแบ่งกลุ่มด้วยการใช้ Naïve Bayes เพื่อตัดประโยคที่ไม่ได้มีการเปรียบเทียบออกจากการวิเคราะห์ความคิดเห็น (Ding, et al., 2009)

อย่างไรก็ตามเนื่องจากโดยธรรมชาติแล้ว ข้อความส่อเสียดส่วนมากมักจะอยู่ในรูปแบบของข้อความที่มีความคิดเห็นไปในเชิงบวกแบบสุดโต่ง ทำให้การวิเคราะห์ทิศทางความคิดเห็นของผู้ใช้อาจไม่จำเป็นในกรณีของการจำแนกข้อความส่อเสียดออกจากข้อความปกติ เนื่องจากเราทราบถึงขั้วความคิดเห็นที่แท้จริงของข้อความส่อเสียดอยู่แล้วว่าข้อความส่อเสียดมักใช้ข้อความเชิงบวกที่สะท้อนความคิดเห็นในเชิงลบ

2.3 โครงสร้างของคำในลักษณะ N-gram

การคำนวณหาความน่าจะเป็นของประโยค คือ การคำนวณหาความน่าจะเป็นที่คำต่าง ๆ จะปรากฏในประโยคในลำดับที่ระบุ โดยข้อความจะได้รับการแปลงให้อยู่ในรูปแบบ N-gram เพื่อความสะดวกในการประมวลผล

ลำดับ N-gram[5] คือลำดับที่ต่อเนื่องกันของ N รายการซึ่งโดยส่วนมากแล้วจะเป็นประโยค ข้อความหรือคำพูด หน่วยย่อยของรายการที่กล่าวถึงจำนวน N อาจเป็นจำนวนเป็นตัวอักษร หรือจำนวนคำ ในกรณีที่หน่วยของรายการเป็นคำ N-gram มักจะถูกเรียกว่า Shingles แทน ซึ่งในการศึกษาครั้งนี้จะอ้างอิงถึงกรณีที่หน่วยย่อยของรายการเป็นคำทั้งหมด ตัวอย่างของแปลงข้อความให้อยู่ในลักษณะ N-gram สามารถทำได้ดังนี้ โดยที่หากค่า N มีค่าเป็น 1 แล้วข้อความ “I went to school” จะถูกสามารถสร้างเป็น [I, went, to,

[5] https://en.wikipedia.org/wiki/N-gram

school] หรือสามารถเรียกแทนได้ว่า Unigram และในกรณีที่ N มีค่าเป็น 2 ข้อความข้างต้นก็จะสามารถจำแนกได้เป็น [I went, went to, to school] เป็นต้น ดังตารางที่ 2.3 จากนั้นนำไปสร้างองค์ประกอบทวิภาค (Binary Feature) เพื่อใช้ในการจำแนกข้อความส่อเสียดต่อไป (Pt´cek, et al., 2014)

ตารางที่ 2.3 ตัวอย่างการแปลงประโยคให้อยู่ในรูปแบบข้อมูลในลักษณะ N-gram

Sentence	**Unigram Sequence (1-Gram)**	**Bigram Sequence (2-Gram)**	**Trigram Sequence (3-Gram)**	**4-Gram Sequence**
I went to school	I, went, to, school	I went, went to, to school	I went to, went to school	I went to school
There is red pencil on the table	There, is, red, pencil, on, the, table	There is, is red, red pencil, pencil on, on the, the table	There is red, is red pencil, red pencil on, pencil on the, on the table	There is red pencil, is red pencil on, red pencil on the, pencil on the table
This shirt will be on sale on black Friday	This, shirt, will, be, on, sale, on, black, Friday	This shirt, shirt will, will be, be on, on sale, sale on, on black, black Friday	This shirt will, shirt will be, will be on, be on sale, on sale on, sale on black, on black Friday	This shirt will be, shirt will be on, will be on sale, be on sale on, on sale on black, sale on black Friday

2.4 การคำนวณความน่าจะเป็นของประโยค

ข้อมูลที่ถูกจัดให้อยู่ในลักษณะของ N-gram สามารถนำไปใช้ในการหาความน่าจะเป็นของประโยค (Jurafsky & Martin, 2014) โดยสามารถคำนวณหาความน่าจะเป็นของคำ w โดยกำหนดให้ลำดับของคำที่ปรากฏในประโยคก่อนหน้า h สามารถเขียนได้เป็น P (w | h) เช่น ถ้ากำหนดประโยคก่อนหน้า h คือ “its water is so transparent that” และต้องการทราบความน่าจะเป็นที่คำต่อไปจะเป็น the จะสามารถเขียนได้ดังนี้

$$P\,(the \mid its\ water\ is\ so\ transparent\ that)$$

ซึ่งสามารถคำนวณความน่าจะเป็นของตัวอย่างข้างต้นได้โดยใช้การนับความถี่สัมพัทธ์ (Relative Frequency) โดยการนับจำนวนประโยค “its water is so transparent that the” หารด้วยจำนวนประโยค “its water is so transparent that” ซึ่งสามารถเขียนได้ดังนี้

$$P\,(the \mid its\ water\ is\ so\ transparent\ that) = \frac{Count(its\ water\ is\ so\ transparent\ that\ the)}{Count(its\ water\ is\ so\ transparent\ that)}$$

แต่เนื่องจากวิธีการคำนวณความน่าจะเป็นจากการนับความถี่ของประโยคที่เกิดขึ้นนั้น ส่วนมากแล้วไม่สามารถใช้ได้ในทางปฏิบัติ เนื่องจากภาษามักจะมีคำหรือประโยคใหม่ ๆ เกิดขึ้นอยู่ตลอดเวลา ดังนั้นจึงมีการใช้ความน่าจะเป็นร่วม (Joint Probability) โดยจะคิดความน่าจะเป็นของคำที่อยู่ติดกันแทน เช่น กำหนดให้ประโยค W ประกอบด้วย w1, w2, w3, … , wn จะสามารถคำนวณหาความน่าจะเป็นของประโยค P(w1,w2,w3,…,wn) โดยนำกฎลูกโซ่แห่งความน่าจะเป็น (Chain rule of probability) ซึ่งแสดงในสมการที่ 2.1 มาใช้

$$P(X_1 \dots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1X_2)P(X_4|X_1X_2X_3) \dots P(X_n|X_1X_2 \dots X_{n-1})$$
$$= P(X_1) \prod_{k=2}^{n} P(X_K|X_1X_2 \dots X_{k-1}) \quad \text{สมการ 2.1}$$

โดยเมื่อนำกฎลูกโซ่แห่งความน่าจะเป็นมาใช้กับการหาความน่าจะเป็นของประโยค P(w1,w2,w3,...,wn) จะได้

$$P(w_1 \dots w_n) = P(w_1)P(w_2|\, w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3) \dots P(w_n|w_1w_2 \dots w_{n-1})$$
$$= P(w_1) \prod_{k=2}^{n} P(w_k|w_1w_2 \dots w_{k-1}) \quad \text{สมการ 2.2}$$

ซึ่งกฎลูกโซ่แสดงถึงการคำนวณหาความน่าจะเป็นร่วมของลำดับ (Joint probability of a sequence) และการคำนวณหาความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) ของคำ w โดยกำหนดคำก่อนหน้า w-1 คำ ซึ่งจากสมการจะเห็นว่า จะสามารถคำนวณความน่าจะเป็นของประโยคได้โดยนำความน่าจะเป็นแบบมีเงื่อนไขมาคูณกัน

สืบเนื่องจากสมมุติฐานของ Markov ซึ่งระบุว่าเราสามารถทำนายความน่าจะเป็นของหน่วยในอนาคตบางหน่วยได้โดยไม่จำเป็นต้องมองย้อนกลับไปยังอดีตมากเกินไป (Jurafsky & Martin, 2014) ดังนั้นจึงสามารถประมาณความน่าจะเป็นแบบมีเงื่อนไขสำหรับแบบจำลอง bigram ได้เป็น

$$P(w_n|w_1w_2 \dots w_{n-1}) \approx P(w_n|w_{n-1}) \quad \text{สมการ 2.3}$$

ดังนั้นแบบจำลอง n-gram สามารถนำมาใช้คำนวณความน่าจะเป็นของโยคได้ โดยไม่จำเป็นต้องใช้การนับจำนวนประโยคทั้งหมดมาคำนวณเหมือนวิธีแรก เช่นการใช้ bigram จะสามารถหาความน่าจะเป็นของคำ โดยกำหนดคำก่อนหน้าทั้งหมด $P(w_n | w_{n-1} w_{n-2} \dots w_1)$ ได้โดยใช้แค่ความน่าจะเป็นแบบมีเงื่อนไขของคำก่อนหน้าเท่านั้น $P(w_n | w_{n-1})$ เช่น แทนที่จะคำนวณหาความน่าจะเป็นของประโยค “Walden Pond’s water is so transparent that” ตามด้วย “the” โดยใช้

P (the | Walden Pond's water is so transparent that)

จะสามารถประมาณความน่าจะเป็นได้เป็น

P (the | that)

ซึ่งเมื่อแทนค่าสมการที่ 2.3 ลงในสมการที่ 2.2 จะสามารถหาความน่าจะเป็นที่จะเกิดคำใด ๆ ได้ดังสมการที่ 2.4

$$P(w_1 w_2 \dots w_{n-1}) \approx P(w_2|w_1) \prod_{k=2}^{n} P(w_k | w_{k-1}) \qquad \text{สมการ 2.4}$$

ตัวอย่างเช่น ชุดข้อมูลประกอบด้วยประโยค s1, s2 และ s3[6] จะสามารถหาความน่าจะเป็นของคำที่เกิดขึ้นติดกันได้ตามตารางที่ 2.4

S1 = <s> I went to school last Saturday </s>

S2 = <s> school is closed on Sunday </s>

S3 = <s> I went to cinema last night </s>

ตารางที่ 2.4 ตัวอย่างความน่าจะเป็นของคำที่เกิดติดกัน

คำ	การคำนวณ	ความน่าจะเป็น
P (I | <s>)	2/3	0.67
P (went | I)	2/2	1.0
P (to | went)	2/2	1.0
P (school | to)	1/2	0.5
P (last | school)	1/1	1.0
P (Saturday | last)	1/2	0.5

[6] สัญลักษณ์ <s> และ </s> แสดงถึงการเริ่มต้นและจบประโยคตามลำดับ

คำ	การคำนวณ	ความน่าจะเป็น
P (</s> | Saturday)	1/1	1.0
P (school | <s>)	1/3	0.33
P (is | school)	1/2	0.5
P (closed | is)	1/1	1
P (on | closed)	1/1	1
P (Sunday | on)	1/1	1
P (</s> | Sunday)	1/1	1
P (cinema | to)	1/2	0.5
P (last | cinema)	1/1	1
P (night | last)	1/2	0.5
P (</s> | night)	1/1	1

หลังจากที่สร้างตารางความน่าจะเป็นของคำที่เกิดติดกันเสร็จเรียบร้อย จะสามารถหาความน่าจะเป็นที่เกิดขึ้นของประโยคได้จากการนำความน่าจะเป็นของคำที่เกิดติดกันทั้งหมดของประโยคมาคูณกันดังตัวอย่างด้านล่าง

S1 = <s> I went to school last Saturday </s>

P (S1) = P (I | <s>) * P (went | I) * P (to | went) * P (school | to) * P (last | school) * P (Saturday | last) * P (</s> | Saturday) = 0.67 * 1.0 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 = **0.1675**

S2 = <s> school is closed on Sunday </s>

P (S2) = P (school | <s>) * P (is | school) * P (closed | is) * P (on | closed) * P (Sunday | on) * P (</s> | Sunday) = 0.33 * 0.5 * 1.0 * 1.0 * 1.0 * 1.0 = **0.165**

S3 = <s> I went to cinema last night </s>

P (S3) = P (I | <s>) * P (went | I) * P (to | went) * P (cinema | to) * P (last | cinema) * P (night | last) * P (</s> | night) = 0.67 * 1.0 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 = **0.1675**

2.5 เครือข่ายสังคมออนไลน์ (Social Network)

ในปัจจุบันเครือข่ายสังคมออนไลน์ได้เข้ามามีบทบาทเป็นอย่างมากในการสื่อสารระหว่างบริษัทและลูกค้า (Haruechaiyasak et al., 2013) ตัวอย่างเช่น การจัดโครงการส่งเสริมการขายสินค้าและการให้บริการในเครือข่ายสังคมออนไลน์ หรือแม้กระทั่งการแจ้งข่าวประกาศให้กับลูกค้า นอกจากนี้ฝ่ายลูกค้ายังมีการนำเสนอความคิดเห็นของผลิตภัณฑ์และบริการกลับมายังบริษัทอีกด้วย ด้วยเหตุนี้จึงส่งผลให้บริษัทมีความต้องการที่จะเข้าใจลูกค้าในแง่มุมต่าง ๆ เช่น ลูกค้ามีความคิดเห็นอย่างไรกับแบรนด์ ผลิตภัณฑ์ หรือบริการต่าง ๆ เพื่อที่จะนำข้อมูลที่ได้ไปปรับปรุงผลิตภัณฑ์และการให้บริการให้ดียิ่งขึ้น อีกทั้งในประเทศไทยยังมีการใช้บริการเครือข่ายสังคมออนไลน์กันอย่างแพร่หลายในการติดต่อกับลูกค้าทั้งในรูปแบบของการซื้อขายสินค้าและบริการเช่น การถามตอบปัญหาการใช้งานบริการอินเตอร์เน็ตของบริษัททรูออนไลน์บนเฟสบุ๊ค[7] หรือการนำเสนอนโยบายส่งเสริมการขายของดีแทคบนทวิตเตอร์[8]

2.5.1 เครือข่ายสังคมออนไลน์เฟสบุ๊ค (Facebook)

เฟสบุ๊คเป็นหนึ่งในเครือข่ายสังคมออนไลน์ที่ได้รับความนิยมอย่างแพร่หลายในประเทศไทยโดยมีผู้ใช้งานมากถึง 17.3 ล้านคน[9] เครือข่ายสังคมออนไลน์อย่างเฟสบุ๊ค นั้นจะให้บริการในรูปแบบของ Microblogging โดยจุดเด่นของเฟสบุ๊คนั้นคือการอนุญาตให้ผู้ใช้งานสามารถเพิ่มสถานะของตัวเองลงไปในหน้าหลักของตัวผู้ใช้งานเอง อีกทั้งยังสามารถเพิ่มสถานะหรือข้อความลงไปยังพื้นที่ส่วนตัวของเพื่อนที่ผู้ใช้งานต้องการได้อีกด้วย นอกจากนี้ผู้ใช้งานยังสามารถที่จะเพิ่มหรือแก้ไขข้อมูลส่วนตัว เช่น ชื่อ นามสกุล สถานที่ทำงาน ที่อยู่ หรือสิ่งที่สนใจ จากนั้นระบบจะทำการแนะนำบุคคลที่น่าจะเป็นเพื่อนกับผู้ใช้งานเนื่องจากมีที่ทำงานเดียวกัน จบจากสถาบันการศึกษาเดียวกัน หรือแม้กระทั่งมีความสนใจร่วมกันอีกด้วย อีกทั้งเฟสบุ๊คเองยังมีระบบที่เรียกว่า Page ซึ่งเปรียบเสมือนช่องทางของธุรกิจซึ่งสามารถใช้ในการติดต่อสื่อสารหรือเผยแพร่ข้อมูลสินค้าแก่ลูกค้าได้และลูกค้าก็สามารถที่จะนำเสนอผลตอบรับของสินค้าและบริการกลับมายังธุรกิจ อย่างไรก็ตามเฟสบุ๊คยังมีข้อจำกัดบางอย่างซึ่งทำให้ไม่สามารถรวบรวมข้อมูลการแสดงความเห็นของผู้ใช้งานมาทำการวิเคราะห์ได้เนื่องจากเฟสบุ๊คนั้นจำกัดการเข้าถึงความคิดเห็นของผู้ใช้งานหากผู้ใช้งานนั้นไม่ได้เกี่ยวข้องเป็นเพื่อนกัน (Friend) กับผู้ร้องขอข้อมูล ซึ่งจะทำให้ไม่สามารถค้นหาข้อมูลที่ต้องการได้ นอกจากนี้การเสนอความคิดเห็นนั้นจะเป็นการเสนอความคิดเห็นในรูปแบบไม่จำกัดตัวอักษร ซึ่งทำให้ใน

---

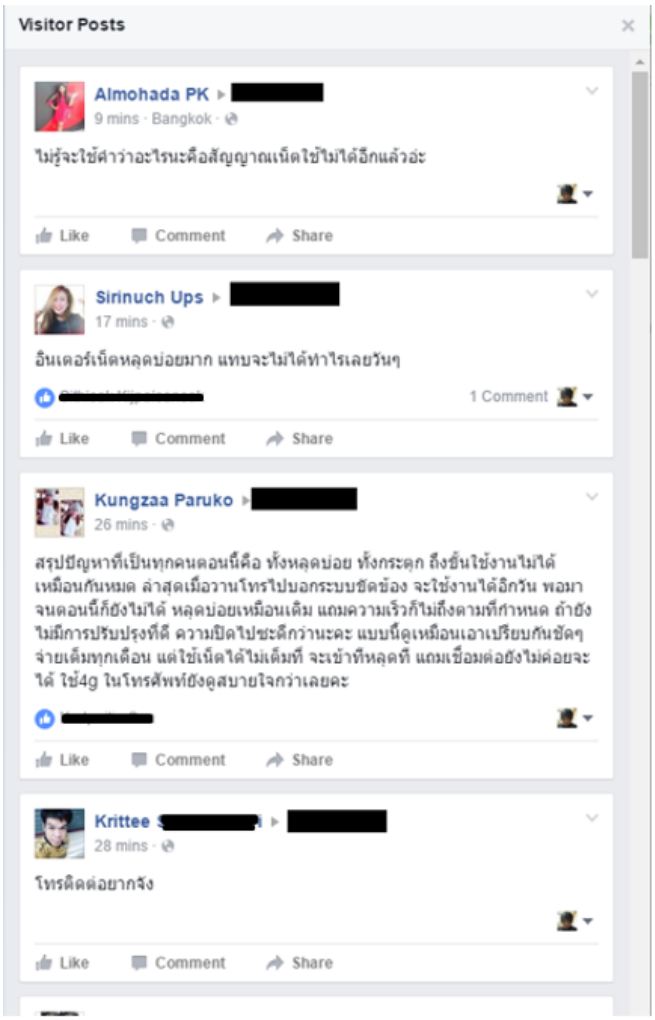
[7] https://www.facebook.com/ByTrueOnline/?fref=ts

[8] https://twitter.com/dtac

[9] รวบรวมข้อมูลสถิติผู้ใช้งานเฟสบุ๊ก หลังสิ้นสุดปี 2558 จาก http://www.statista.com/statistics/490467/number-of-thailand-facebook-users/

กรณีที่ความคิดเห็นมีความยาวอาจจะเป็นการกล่าวถึงหลายหัวข้อ หลายประเด็นปะปนกันไป ซึ่งส่งผลให้การวิเคราะห์ข้อมูลนั้นมีโอกาสเกิดความคลาดเคลื่อนจากหัวข้อที่ทำการวิเคราะห์สูง

รูปที่ 2.1 ตัวอย่างการแสดงความเห็นของผู้ใช้งานกับผู้ให้บริการบนเฟสบุ๊ค



มีการศึกษามากมายเกี่ยวกับการประมวลผลภาษาธรรมชาติซึ่งทำการวิเคราะห์บนเฟสบุ๊คเช่น การวิเคราะห์ความคิดเห็นของผู้ใช้งานและการนำไปปรับใช้กับสื่อการเรียนออนไลน์ (Ortigosa et al., 2014) ซึ่งเป็นการนำข้อความของผู้ใช้งานในเฟสบุ๊คมาวิเคราะห์เพื่อหาว่าความคิดเห็นของผู้ใช้งานเป็นไปในทิศทางบวกหรือทางลบโดยใช้เทคนิคการวิเคราะห์โดยใช้ฐานคำศัพท์ (Lexical Based) และการเรียนรู้ของเครื่อง (Machine Learning) โดยมีความแม่นยำถึง 83.27% จากนั้นจึงนำการวิเคราะห์ดังกล่าว

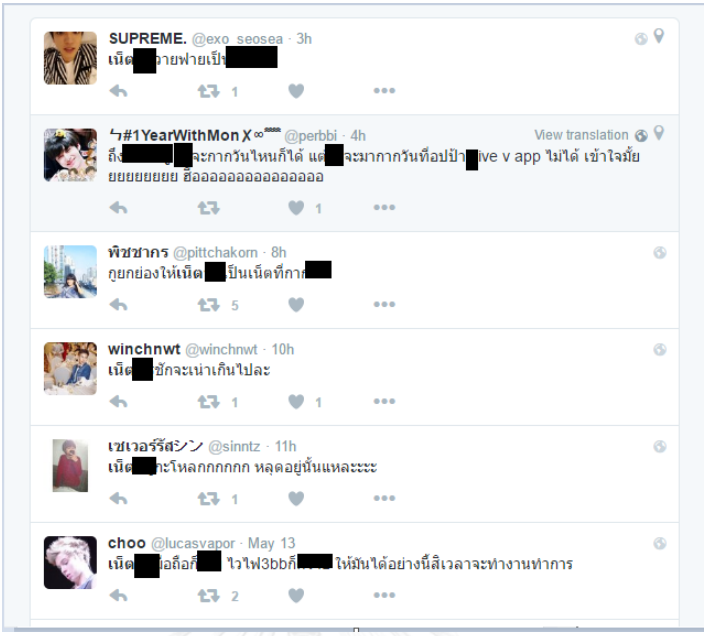
ไปใช้ในระบบสื่อการเรียนออนไลน์เพื่อวิเคราะห์ความพึงพอใจของผู้เรียนต่อบทเรียนนั้น ๆ และต่ออาจารย์ผู้สอนเพื่อที่จะนำเสนอบทเรียนที่เหมาะสมกับผู้เรียนและปรับปรุงคุณภาพการเรียนการสอนให้ดียิ่งขึ้น

นอกจากนี้ยังมีการใช้การประมวลผลภาษาธรรมชาติในการสืบสวนคดีต่าง ๆ เช่นการวิเคราะห์ข้อความก่อนตายในคดีการฆ่าตัวตาย (Pestian et al., 2012) โดยวิเคราะห์จากจดหมายอิเล็กทรอนิกส์และข้อความของผู้ใช้งานก่อนเกิดคดีฆ่าตัวตายเพื่อจำแนกอารมณ์ของผู้ตายและวิเคราะห์แรงจูงใจต่าง ๆ เพื่อหาสาเหตุและแรงจูงใจในการฆ่าตัวตายเป็นต้น

2.5.2 เครือข่ายสังคมออนไลน์ทวิตเตอร์ (Twitter)

นอกจากเฟสบุ๊คแล้วยังมีเครือข่ายสังคมออนไลน์อีกแห่งหนึ่งที่ได้รับความนิยมไม่แพ้กันคือ ทวิตเตอร์ โดยที่ทวิตเตอร์นั้นมีการดำเนินกิจการในรูปแบบ Microblogging เช่นเดียวกับเฟสบุ๊ค โดยผู้ใช้สามารถตั้งค่าข้อมูลส่วนตัว เช่น ชื่อผู้ใช้งาน อีเมล หรือรหัสผ่าน เป็นต้น การเผยแพร่ข้อความในทวิตเตอร์นั้นเรียกว่า “ทวีต” ซึ่งเป็นการนำข้อความของผู้ใช้งานไปแสดงในพื้นที่ส่วนตัวของผู้ใช้งานนั้น ๆ โดยที่ผู้ใช้งานสามารถระบุแฮชแท็ก (Hashtag) ซึ่งเปรียบเสมือนคำระบุหมวดหมู่ว่าข้อความที่ผู้ใช้เผยแพร่นั้นอยู่ในหมวดหมู่แบบไหน เช่น หากข้อความนั้นมีแฮชแท็กคำว่า “#อินเตอร์เน็ต” หมายความว่าข้อความที่ผู้ใช้งานเผยแพร่นั้นจะเป็นข้อความเกี่ยวกับอินเตอร์เน็ต เป็นต้น นอกจากนี้ทวิตเตอร์ยังถูกใช้เป็นช่องทางเผยแพร่ข้อมูลข่าวสารระหว่างบริษัทกับลูกค้าอีกด้วย โดยจะแสดงอยู่ในรูปแบบของการติดตาม (Following) ซึ่งหากผู้ใช้สนใจที่จะติดตาม (Follow) ข่าวสารของธุรกิจใด ผู้ใช้สามารถใช้การติดตามเพื่อที่จะได้รับข่าวสารเมื่อมีการนำเสนอข่าวสารในทวิตเตอร์ของธุรกิจนั้น ๆ นอกจากนี้แล้วผู้ใช้ยังสามารถตอบกลับ (Reply) ไปยังธุรกิจนั้น ๆ ผ่านข้อความส่วนตัวเพื่อติดต่อสอบถามหรือแจ้งปัญหาไปยังธุรกิจได้โดยตรง โดยผู้ใช้สามารถใช้แฮชแท็กเพื่อระบุหัวข้อที่เกี่ยวข้องของข้อความนั้น ๆ ซึ่งจุดที่แตกต่างกับเฟสบุ๊คนั้นคือเราสามารถค้นหาความคิดเห็นที่บุคคลใด ๆ ทวีตลงในทวิตเตอร์ได้โดยการใช้แฮชแท็กอีกทั้งยังมีระบบการค้นหาขั้นสูงซึ่งสามารถระบุส่วนใดส่วนหนึ่งของข้อความ ภาษาที่ใช้ หรือตัดทวีตที่มีข้อความที่ไม่ต้องการออกได้ ทำให้การรวบรวมข้อมูลเป็นไปได้ง่ายและตรงตามความต้องการมากที่สุด นอกจากนี้แล้วรูปแบบข้อความความคิดเห็นในทวิตเตอร์นั้นจะเป็นข้อความในรูปแบบจำกัดตัวอักษรที่ 140 ตัวอักษร ซึ่งส่งผลให้ความคิดเห็นที่ระบุผ่านเครือข่ายสังคมออนไลน์ทวิตเตอร์มีแนวโน้มที่จะจำกัดอยู่เพียงเรื่องเดียว อีกทั้งข้อความในทวิตเตอร์นั้นมีการใช้แฮชแท็กซึ่งสามารถใช้เป็นตัวกรองเบื้องต้นซึ่งแสดงให้เห็นว่าข้อความนั้นกล่าวถึงเรื่องใด จึงทำให้การวิเคราะห์ข้อความความคิดเห็นของผู้ใช้งานเป็นไปได้อย่างสะดวกมากยิ่งขึ้น

รูปที่ 2.2 ตัวอย่างการแสดงความเห็นของผู้ใช้งานกับผู้ให้บริการบนทวิตเตอร์



มีการศึกษามากมายเกี่ยวกับการวิเคราะห์ภาษาธรรมชาติในเครือข่ายสังคมออนไลน์ทวิตเตอร์ ในงานสัมมนา Association for the Advancement to Artificial Intelligence ครั้งที่ 4 (Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media) มีงานวิจัยเกี่ยวกับการวิเคราะห์ความคิดเห็นของผู้ใช้งานเกี่ยวกับการเมืองและการเลือกตั้งที่เกิดขึ้นในปี 2012 ของประเทศสหรัฐอเมริกา (Thelwall et al., 2011; Tumasjan, et al., 2010; Wang et al., 2012) เพื่อหาภาพรวมของผลกระทบที่เกิดขึ้นในเครือข่ายสังคมออนไลน์เมื่อมีกิจกรรมต่าง ๆ เกิดขึ้น เช่นการตอบคำถามของผู้สมัครประธานาธิบดีจะส่งผลต่อความคิดเห็นของประชาชนในเครือข่ายสังคมออนไลน์อย่างไร

นอกจากนี้ยังมีการศึกษาเกี่ยวกับการพยากรณ์ทิศทางการเปลี่ยนแปลงของหุ้นโดยดูจากความคิดเห็นต่าง ๆ ของผู้ใช้งานบนเครือข่ายสังคมออนไลน์ทวิตเตอร์ (Bollen et al., 2011) ซึ่งความแม่นยำสูงถึงร้อยละ 86.4 อย่างไรก็ตาม ความคิดเห็นหรืออารมณ์ของผู้ใช้งานนั้นจะมีผลกระทบต่อการเปลี่ยนแปลงทิศทางการเคลื่อนไหวของหุ้นในอีกสามถึงสี่วันถัดมา

จากความสามารถในการเก็บรวบรวมข้อมูลและความแม่นยำในการวิเคราะห์ด้วยการประมวลผลภาษาธรรมชาติของเครือข่ายสังคมออนไลน์ดังที่กล่าวไปข้างต้น จะเห็นได้ว่าทวิตเตอร์นั้นสามารถที่จะรวบรวมข้อมูลได้ง่ายในหัวข้อที่ต้องการผ่านทางการค้นหาข้อมูลผ่านแฮชแท็ก โดยสามารถค้นหาข้อมูลที่ต้องการโดยที่มีข้อจำกัดน้อยกว่าเฟสบุ๊ค นอกจากนี้ด้วยจำนวนตัวอักษรที่จำกัดซึ่งส่งผลให้ขอบเขตเนื้อหาของข้อความนั้นมีความเฉพาะเจาะจงกว่าเฟสบุ๊ค ผู้วิจัยจึงเลือกเครือข่ายสังคมออนไลน์ทวิตเตอร์เป็น

เครือข่ายสังคมออนไลน์ที่จะใช้ในการรวบรวมข้อมูลที่จะใช้ในการศึกษาครั้งนี้

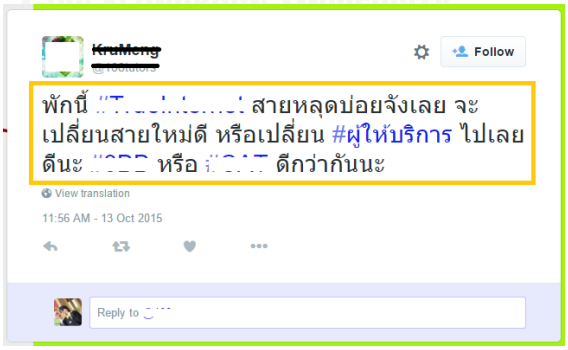
รูปแบบข้อมูลจากเครือข่ายสังคมทวิตเตอร์

ข้อมูลในทวิตเตอร์ประกอบไปด้วย 5 ส่วนหลักคือ ข้อความทวีต (Tweet) จำนวนรีทวีต (Retweet) จำนวนการกดถูกใจ (Like) จำนวนผู้ติดตาม (Follower) และจำนวนการติดตาม (Following) ซึ่งข้อความทวีตนั้นสามารถจำแนกออกเป็นข้อความและไอคอนแสดงอารมณ์ (Emoticon) ดังรายละเอียดต่อไปนี้

ข้อความทวีต (Tweet)

ข้อความทวีตคือข้อความที่ผู้ใช้งานเขียนลงไปบนทวิตเตอร์เพื่อให้ผู้ใช้รายอื่น ๆ หรือผู้ติดตามรับรู้ โดยมีจำนวนตัวอักษรสูงสุดคือ 140 ตัวอักษร ข้อความทวีตนั้นจะประกอบไปด้วยส่วนย่อยสามส่วนคือ ข้อความ ไอคอนแสดงอารมณ์ และแฮชแท็ก โดยข้อความนั้นจะอยู่ในรูปแบบตัวอักษร ไอคอนแสดงอารมณ์นั้นจะอยู่ในรูปแบบรูปภาพหรือสัญลักษณ์พิเศษและแฮชแท็กนั้นจะอยู่ในรูปแบบข้อความที่นำหน้าด้วยเครื่องหมายแฮชเช่น #hello, #holiday เป็นต้น โดยรูปที่ 2.3 จะแสดงตัวอย่างของข้อความทวีตบนทวิตเตอร์

รูปที่ 2.3 ตัวอย่างข้อความทวีตบนทวิตเตอร์



จำนวนรีทวีต (Retweet)

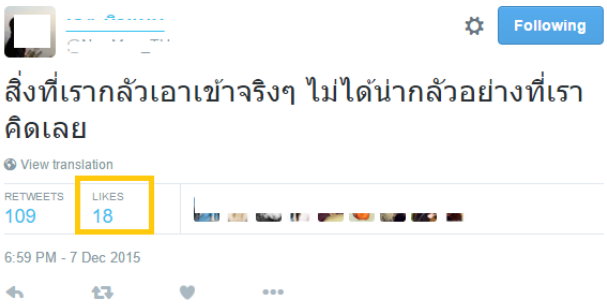
จำนวนรีทวีตคือจำนวนครั้งที่ผู้ใช้งานรายอื่น ๆ เผยแพร่ข้อความซ้ำโดยการกดปุ่ม Retweet ซึ่งการทวีตซ้ำนั้นจะช่วยให้ผู้ติดตามของผู้ที่รีทวีตนั้นเห็นข้อความดังกล่าว โดยจำนวนรีทวีตนั้นอาจจะขึ้นอยู่กับเนื้อหา แฮชแท็ก URL รูปภาพ หรือจำนวนผู้ติดตาม เป็นต้น (Suh, Lichan, Pirolli, & Chi, 2010) จำนวนการรีทวีตนั้นจะแสดงที่ด้านล่างของข้อความทวีต และอยู่ด้านหน้าจำนวนการกดถูกใจดังตัวอย่างในรูปที่ 2.4

รูปที่ 2.4 ตัวอย่างจำนวนรีทวีตบนทวิตเตอร์ซึ่งแสดงจำนวนรีทวีต 14 ครั้ง

จำนวนการกดถูกใจ (Like)

จำนวนการกดถูกใจคือจำนวนครั้งที่ผู้ใช้งานกดปุ่มถูกใจข้อความ ซึ่งจำนวนการกดถูกใจนั้นจะนับเฉพาะการกดถูกใจของผู้ใช้งานที่ไม่ซ้ำกันโดยผู้ใช้แต่ละคนที่กดปุ่มถูกใจจะสามารถกดได้เพียงครั้งเดียว หากกดปุ่มถูกใจอีกครั้งจะเป็นการยกเลิกการกดถูกใจ โดยจำนวนการกดถูกใจนั้นอาจขึ้นอยู่กับปัจจัยหลายอย่าง เช่น ข้อความทวีต แฮชแท็ก รูปภาพ หรือจำนวนผู้ติดตาม เป็นต้น จำนวนการกดถูกใจนั้นจะแสดงอยู่ด้านล่างของข้อความทวีตและถัดจากจำนวนการรีทวีตดังตัวอย่างในรูปที่ 2.5

รูปที่ 2.5 ตัวอย่างจำนวนการกดถูกใจบนทวิตเตอร์ซึ่งแสดงจำนวนการกดถูกใจจำนวน 18 ครั้ง



จำนวนการติดตาม (Following)

จำนวนการติดตามคือจำนวนที่บุคคลอื่นที่ผู้ใช้งานติดตาม โดยการติดตามนั้นจะทำให้ผู้ใช้งานเห็นข้อความทวีตของผู้ที่ถูกติดตามในหน้าข่าวสารของผู้ใช้งานเอง จำนวนการติดตามจะแสดงผลอยู่ในหน้าโพรไฟล์ของผู้ใช้งานซึ่งอยู่ระหว่างจำนวนทวีตและจำนวนผู้ติดตามดังตัวอย่างในรูปที่ 2.6

รูปที่ 2.6 ตัวอย่างจำนวนการติดตามบนทวิตเตอร์ซึ่งแสดงว่า ผู้ใช้งานติดตามผู้ใช้งานคนอื่นจำนวน 78 คน

จำนวนผู้ติดตาม (Follower)

จำนวนผู้ติดตามคือจำนวนผู้ใช้งานรายอื่น ๆ ที่กดปุ่มติดตามผู้ใช้งาน ซึ่งจำนวนผู้ติดตามนั้นจะนับเฉพาะผู้ติดตามที่ไม่ซ้ำกัน โดยผู้ติดตามนั้นจะเห็นข้อความทวีตของผู้ถูกติดตามที่ทำการทวีตซึ่งจะแสดงผลขึ้นมาบนหน้าข่าวสาร (Feed) ของผู้ติดตาม จำนวนผู้ติดตามจะแสดงผลอยู่ในหน้าโปรไฟล์ของผู้ใช้งานนั้น ๆ ซึ่งแสดงอยู่หลังจำนวนผู้ใช้คนอื่นที่ผู้ใช้รายนั้น ๆ ติดตามอยู่ ดังตัวอย่างในรูปที่ 2.7

รูปที่ 2.7 ตัวอย่างจำนวนผู้ติดตามบนทวิตเตอร์ซึ่งแสดงจำนวนผู้ติดตามผู้ใช้งานรายนี้เป็นจำนวนจำนวน 67 คน

2.6 การจำแนกข้อความส่อเสียด

ข้อความส่อเสียดหรือข้อความประชดเป็นข้อความที่ให้ความหมายไปในทางตรงกันข้ามกับสิ่งที่ผู้เขียนต้องการจะสื่อโดยมีจุดประสงค์เพื่อตั้งใจล้อเลียนเกี่ยวกับสิ่งนั้น ยกตัวอย่างเช่น ในขณะที่สภาพอากาศเลวร้าย มีลมพายุพัดอย่างรุนแรงแล้วมีใครสักคนพูดว่า “วันนี้อากาศดีจังเลย” ผู้พูดอาจไม่ได้หมายความว่าอากาศนั้นดีแต่หมายความว่าวันนี้อากาศแย่ซึ่งให้ความหมายตรงกันข้ามกับสิ่งที่ผู้เขียนต้องการจะสื่อ ดังนั้นข้อความที่ผู้พูดพูดนั้นจึงถือว่าเป็นข้อความส่อเสียด หรืออีกในกรณีหนึ่งเกี่ยวกับการเขียนวิจารณ์สินค้าและบริการตัวอย่างเช่น อินเตอร์เน็ตที่ใช้อยู่ปัจจุบันช้ามากแต่มีคนแสดงความคิดเห็นต่อผู้ให้บริการอินเตอร์เน็ตรายนี้ว่า “ไม่เคยใช้เน็ทอะไรเร็วเท่านี้มาก่อนเลย” นั่นก็ถือว่าเป็นการประชดส่อเสียด เนื่องจากผู้วิจารณ์ไม่ได้ต้องการจะสื่อความหมายตรง ๆ ตามที่ได้เขียนไว้แต่ต้องการจะสื่อว่าอินเตอร์เน็ตนั้นช้ามากเป็นต้น

2.6.1 ผลกระทบและความจำเป็นของการจำแนกข้อความส่อเสียด

เนื่องจากข้อความส่อเสียดนั้นให้ความหมายในทิศทางตรงกันข้ามกับสิ่งที่ผู้เขียนต้องการจะสื่อจึงทำให้การวิเคราะห์ความคิดเห็นนั้นเกิดความผิดพลาดและส่งผลกระทบอย่างมากต่อการวิเคราะห์ความคิดเห็น เพราะว่าข้อความส่อเสียดนั้นจะเปลี่ยนขั้ว (Polarity) ของความคิดเห็นที่วิเคราะห์ได้โดยเปลี่ยนจากความคิดเห็นที่เป็นลบให้กลายเป็นบวก ซึ่งหมายความว่าความคิดเห็นที่วิเคราะห์ได้นั้นจะกลายเป็นทิศทางตรงกันข้าม และโดยส่วนมากแล้ว ข้อความส่อเสียดนั้นมักจะเป็นไปในรูปแบบของขั้วความคิดเห็นแบบสุดโต่ง เช่น การเปลี่ยนขั้วของข้อความจากดีมาก ๆ เป็นแย่มาก ๆ เป็นต้น ดังนั้นความสามารถในการจำแนกข้อความส่อเสียดจึงเป็นส่วนสำคัญซึ่งช่วยให้การสรุปความคิดเห็นมีความถูกต้องและแม่นยำมากยิ่งขึ้น

2.6.2 งานวิจัยที่เกี่ยวข้องกับการจำแนกข้อความส่อเสียด

ในการจำแนกข้อความส่อเสียดออกจากข้อความปกตินั้น คุณลักษณะของภาษาถือว่าเป็นส่วนสำคัญที่สามารถนำมาใช้ได้ (Pt´cek et al., 2014) นอกจากนี้ยังมีส่วนประกอบอื่น ๆ ที่อาจมีส่วนช่วยในการจำแนกข้อความส่อเสียดเช่น จำนวนไอคอนแสดงอารมณ์รูปแบบต่าง ๆ ที่ปรากฏอยู่ในข้อความ การใช้ตัวอักษรพิมพ์ใหญ่ หรือแม้กระทั่งเครื่องหมายพิเศษเช่น เครื่องหมายอัศเจรีย์ (!) เป็นต้น

การจำแนกข้อความส่อเสียดโดยใช้คุณลักษณะเฉพาะของภาษา

ข้อมูลในรูปแบบ N-gram เป็นลักษณะข้อมูลที่นิยมใช้ในการจำแนกข้อความส่อเสียดอย่างแพร่หลาย (Bamman & Smith, 2015; Gonzlez-Ibez et al., 2011; Pt´cek et al., 2014) โดยจะแบ่งประโยคออกเป็นคำเช่น “I went to school” หากค่า N มีค่าเท่ากับ 1 จะสามารถสร้างแบบจำลอง N-gram ได้ดังนี้

N-gram (N=1) = I, went, to, school

ซึ่งจากการศึกษาของ Pt´cek et al (2014) นั้นจะใช้ลักษณะการปรากฏของคำในข้อความเป็นองค์ประกอบหลัก นอกจากนี้ยังมีการใช้องค์ประกอบของประโยคอื่น ๆ เช่น จำนวนเครื่องหมายพิเศษต่าง ๆ หรือไอคอนแสดงอารมณ์ โดยอันดับแรกจะสร้างแบบจำลอง N-gram โดยนับคำของข้อมูลทั้งหมดที่เกิดขึ้น และตัดเอาเฉพาะคำที่มีจำนวนการเกิดขึ้นมากกว่า 3 ครั้ง (Rajadesingan et al., 2015) ดังตารางที่ 2.5 และ 2.6 ซึ่งจะเห็นได้ว่าคอลัมน์ breakfast ในแบบจำลอง N-gram จะถูกตัดออกเนื่องจากมีจำนวนครั้งที่ปรากฏน้อยกว่า 3 ครั้ง

ตาราง 2.5 ตัวอย่างการแปลงประโยคให้อยู่ในรูปแบบ N-gram และนับจำนวนคำที่เกิดขึ้นในแต่ละหน่วยของ N-gram

I	breakfast	school	eat	...	tomorrow
20	2	13	7	...	10

ตาราง 2.6 ตัวอย่างการแปลงประโยคให้อยู่ในรูปแบบ N-gram และนับจำนวนคำที่เกิดขึ้นในแต่ละหน่วยของ N-gram หลังตัดคำที่มีจำนวนการปรากฏน้อยกว่า 3 ครั้งออก

I	school	eat	...	tomorrow
20	13	7	...	10

หลังจากนั้นตรวจสอบว่าในแต่ละประโยคมีคำใดในแบบจำลอง N-gram ปรากฏอยู่บ้างดังตัวอย่างในตารางที่ 2.7 โดยที่เลข 1 หมายความว่าคำดังกล่าวปรากฏอยู่ในประโยค และเลข 0 หมายถึงคำดังกล่าวไม่ปรากฏอยู่ในประโยค เช่น ประโยค “I go to work” ในตารางที่ 2.7 จะมีคำว่า I และ to ปรากฏ

อยู่ ดังนั้นคอลัมน์ I และ to จึงใส่เลข 1 ส่วนช่อง school นั้นเนื่องจากประโยคข้างต้นไม่ปรากฏคำว่า school จึงใส่เลข 0

ตาราง 2.7 ตัวอย่างองค์ประกอบทวิภาคโดยใช้การปรากฏของคำในประโยค ซึ่งใช้ในการจำแนกข้อความส่อเสียด

ประโยค	**I**	**to**	**...**	**school**
I went to school	1	1	...	1
I go to work	1	1	...	0
School is nice	0	0	...	1

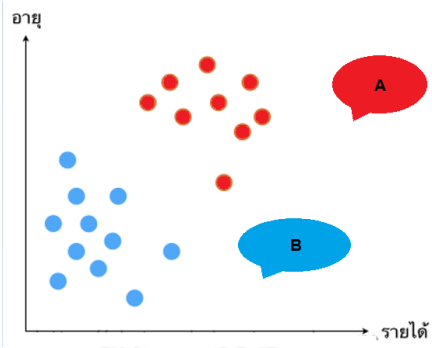
หลังจากได้ตารางองค์ประกอบทวิภาคที่แสดงถึงการปรากฏของคำในแต่ละประโยคแล้ว จะกำหนดให้คอลัมน์คำศัพท์เช่น I, to, school ในตาราง N-gram ซึ่งการปรากฏของคำศัพท์แต่ละคำในประโยคเป็นองค์ประกอบ (Feature) ซึ่งใช้ในการจำแนกข้อความส่อเสียดออกจากข้อความปกติ จากนั้นจึงใช้อัลกอริทึม Maximum Entropy (MaxEnt) และ Support Vector Machine (SVM) โดยสิ่งที่แตกต่างกันระหว่าง 2 อัลกอริทึมนั้นคือ MaxEnt จะเป็นอัลกอริทึมการจำแนกข้อมูลโดยใช้ความน่าจะเป็น โดยจะเปลี่ยนองค์ประกอบในแต่ละคอลัมน์ให้อยู่ในรูปแบบเวคเตอร์ (Vector) และคำนวณหาน้ำหนักของแต่ละองค์ประกอบ (Weight) เพื่อที่จะนำใช้จำแนกประเภทของข้อมูล (Label) สำหรับชุดขององค์ประกอบนั้น ๆ (Feature Set) โดยความน่าจะเป็นในแต่ละประเภทข้อมูล (Label) สามารถเขียนให้อยู่ในรูปแบบของสมการได้ดังสมการที่ 2.5 ซึ่งก็คือ ความน่าจะเป็นของประเภทข้อมูล “label” กำหนดชุดองค์ประกอบ “fs” สามารถหาได้จาก ผลคูณเวกเตอร์ของประเภทข้อมูล “label” หารด้วย ผลคูณเวกเตอร์ของแต่ละประเภทข้อมูลทั้งหมดรวมกัน

$$P(fs|label) = \frac{weights \bullet encode(fs,label)}{\sum_{for\ each\ l\ in\ label}(weights \bullet encode(fs,l))} \qquad \text{สมการ 2.5}$$

แต่ใน SVM นั้นจะเป็นอัลกอริทึมการจำแนกข้อมูลแบบเชิงเส้นโดยมีหลักการเพื่อที่จะหาเส้นแบ่งแยกที่ดีที่สุดที่สามารถแบ่งแยกข้อมูลในแต่ละประเภทออกจากกัน ตัวอย่างเช่น หากมีข้อมูลซึ่งแสดง

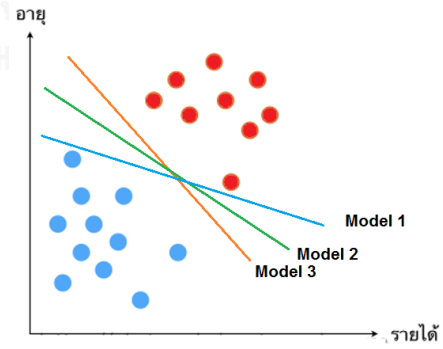
ความสัมพันธ์ระหว่าง อายุ และ รายได้ และมีประเภทของข้อมูลคือ A และ B ซึ่งสามารถเขียนแสดงให้เห็นในรูปแบบกราฟได้ดังรูปที่ 2.8 โดยกำหนดให้แกน X แทนรายได้ และแกน Y แทนอายุ

รูป 2.8 ตัวอย่างการแบ่งประเภทข้อมูล A และ B โดยอิงจากอายุและรายได้



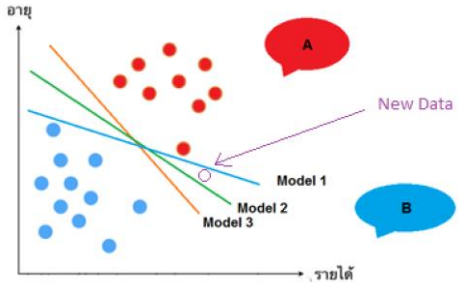
ซึ่งจากรูปที่ 2.8 สามารถแบ่งประเภทข้อมูลออกเป็นสองกลุ่มได้อย่างชัดเจนคือ A และ B ซึ่งโดยปกติแล้วจะสร้างสมการเส้นตรง (Linear Model) เพื่อแบ่งข้อมูลออกเป็นสองส่วน แต่เนื่องจากสมการเส้นตรงที่ใช้แบ่งประเภทข้อมูลออกเป็นสองกลุ่มนั้นสามารถแบ่งได้หลากหลายรูปแบบดังรูปที่ 2.9

รูป 2.9 ตัวอย่างการแบ่งประเภทข้อมูลออกเป็นสองกลุ่ม



ซึ่งจากภาพจะเห็นว่า model 1 และ model 3 ค่อนข้างอิงกับข้อมูลชุดใดชุดหนึ่งมากเกินไป (Overfitting) ซึ่งส่งผลให้ในกรณีที่มีข้อมูลชุดใหม่เกิดขึ้นห่างออกไปจากประเภทข้อมูลเดิมเล็กน้อยจะได้รับการจำแนกผิดพลาด ดังรูปที่ 2.10

รูป 2.10 ตัวอย่างปัญหา Overfitting



ซึ่งหากใช้ model 1 ข้อมูลตัวใหม่จะถูกจำแนกเป็นข้อมูลประเภท B แทนที่จะเป็นประเภท A ซึ่งอัลกอริทึม SVM นั้นจะเลือก Model ที่มีระยะห่างระหว่างประเภทข้อมูลมากที่สุดซึ่งก็คือ Model 2 ในรูปที่ 2.10

โดยจากการศึกษาพบว่าการใช้ n-gram นั้นมีความแม่นยำในการจำแนกข้อความส่อเสียดโดยใช้อัลกอริทึม MaxEnt ถึงร้อยละ 93.28 ในขณะที่ความแม่นยำในการจำแนกข้อความส่อเสียดโดยอัลกอริทึม SVM นั้นกลับให้ความแม่นยำน้อยกว่าการใช้ MaxEnt โดยมีความแม่นยำเพียงร้อยละ 92.86

เนื่องจากเทคนิคที่กล่าวมาข้างต้นนั้นเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) จึงส่งผลให้การจำแนกข้อความส่อเสียดนั้นมีประสิทธิภาพในการจำแนกข้อความส่อเสียดกับข้อมูลชุดนั้นเพียงอย่างเดียว ซึ่งหากจำเป็นต้องจำแนกข้อความส่อเสียดออกจากข้อความปกติในข้อมูลชุดใหม่ เทคนิคนี้จำเป็นที่จะต้องมีการสอน (Train) และสร้างข้อมูลเรียนรู้ (Training Data) ของข้อมูลชุดใหม่ก่อนทุกครั้ง

การวิเคราะห์ข้อความส่อเสียดโดยการใช้องค์ประกอบอื่นนอกเหนือจากภาษา

- **การนับจำนวนของเครื่องหมายพิเศษต่าง ๆ ในข้อความ**

เครื่องหมายพิเศษต่าง ๆ เช่น เครื่องหมายอัศเจรีย์ (!) เครื่องหมายปรัศนี (?) หรือสัญลักษณ์พิเศษต่าง ๆ ก็ถูกใช้เป็นส่วนหนึ่งขององค์ประกอบในการจำแนกข้อความส่อเสียดออกจากข้อความปกติ (Gonzlez-Ibez et al., 2011; Pt´cek et al., 2014) โดยใช้การนับจำนวนของสัญลักษณ์พิเศษต่าง ๆ ที่ปรากฏในข้อความเป็นองค์ประกอบเช่น ข้อความ “อินเตอร์เน็ตวันนี้ดีจัง!!!!” จะมีสัญลักษณ์พิเศษซึ่งก็คือเครื่องหมายอัศเจรีย์ปรากฏอยู่ 4 ตัว จากนั้นนำจำนวนเครื่องหมายอัศเจรีย์ที่ได้มาเปลี่ยนให้อยู่ในรูปแบบที่เหมาะสม (Normalize) โดยนำจำนวนเครื่องหมายอัศเจรีย์ในประโยคนั้นหารด้วยจำนวนเครื่องหมายอัศเจรีย์ที่มากที่สุดที่ปรากฏอยู่ในกลุ่มของข้อความที่ใช้ในการวิเคราะห์ และนำมาคูณกับค่าเฉลี่ยของ

องค์ประกอบอื่น ๆ เช่น จำนวนไอคอนแสดงอารมณ์ หรือจำนวนคำที่ใช้ตัวอักษรพิมพ์ใหญ่ ที่มากที่สุด เช่น หากองค์ประกอบไอคอนแสดงอารมณ์ในแง่บวกและแง่ลบมีจำนวนสูงสุดอยู่ที่ 3 และ 5 ไอคอนตามลำดับ และจำนวนคำที่ใช้ตัวอักษรพิมพ์ใหญ่ที่มากที่สุดในชุดข้อมูลอยู่ที่ 5 คำ ดังนั้น หากจำนวนเครื่องหมายอัศเจรีย์ที่มากที่สุดที่ปรากฏอยู่ในข้อความอื่น ๆ มีอยู่ 6 ตัว ค่าเฉลี่ยของจำนวนที่มากที่สุดขององค์ประกอบอื่น ๆ คือ จำนวนไอคอนแสดงอารมณ์ในแง่บวก รวมกับ จำนวนไอคอนแสดงอารมณ์ในแง่ลบ รวมกับ จำนวนคำที่ใช้ตัวอักษรพิมพ์ใหญ่ หารด้วย 3 ซึ่งก็คือ (3+5+5)/3 = 4.33 จากนั้นจึงสามารถหาค่าขององค์ประกอบเครื่องหมายพิเศษในข้อความได้โดยนำจำนวนสัญลักษณ์พิเศษที่ปรากฏอยู่ในข้อความ หารด้วย จำนวนสัญลักษณ์พิเศษที่มากที่สุดที่ปรากฏอยู่ในข้อความอื่น ๆ คูณด้วยค่าเฉลี่ยของจำนวนที่มากที่สุดขององค์ประกอบอื่น ๆ ดังนั้นองค์ประกอบของข้อความ “อินเตอร์เน็ตวันนี้ดีจัง!!!!” จะคำนวณได้เท่ากับ (4/6) * 4.33 = 2.89

จากนั้นจึงนำค่าที่ได้มาจำแนกโดยใช้อัลกอริทึม MaxEnt และ SVM โดยจากการศึกษาพบว่าการใช้จำนวนสัญลักษณ์พิเศษต่าง ๆ เพิ่มเข้าไปโดยมี n-gram เป็นฐานนั้นมีส่วนช่วยให้ความแม่นยำในการจำแนกข้อความส่อเสียดโดยใช้อัลกอริทึม MaxEnt ดียิ่งขึ้นจากร้อยละ 93.28 เป็นร้อยละ 93.32 ในขณะที่ความแม่นยำในการจำแนกข้อความส่อเสียดโดยอัลกอริทึม SVM นั้นกลับลดลงจากร้อยละ 92.86 เป็นร้อยละ 92.84

- **การนับจำนวนไอคอนแสดงอารมณ์**

ไอคอนแสดงอารมณ์ก็ถือเป็นองค์ประกอบหนึ่งที่นิยมใช้ในการจำแนกข้อความส่อเสียดออกจากข้อความปกติ (Pt´cek et al., 2014) เช่น การนับจำนวนครั้งที่ปรากฏขึ้นของไอคอนแสดงอารมณ์ซึ่งมีค่าแสดงอารมณ์ในเชิงบวกและเชิงลบตัวอย่างเช่น ☺ และ ☹ เป็นต้น จากนั้นจึงนำไปสร้างตารางเพื่อกำหนดเป็นองค์ประกอบซึ่งใช้ในการจำแนกข้อความส่อเสียดดังตาราง 2.8 จากนั้นจึงนำค่าองค์ประกอบที่ได้ไปจำแนกข้อความส่อเสียดโดยใช้ MaxEnt และ SVM เช่นเดียวกับองค์ประกอบที่แล้ว

ตาราง 2.8 ตัวอย่างการสร้างตารางองค์ประกอบโดยใช้ไอคอนแสดงอารมณ์

ประโยค	☺	☹
พายุเข้าแบบนี้อากาศดีจัง ☹ ☹ ☹	0	3
วันนี้อากาศดีจัง ☺	1	0

- **การนับจำนวนของตัวอักษรหรือคำที่ใช้ตัวอักษรพิมพ์ใหญ่ทั้งหมด**

ประโยคที่ใช้ตัวอักษรพิมพ์ใหญ่ทั้งหมดเช่น ประโยค “THE WEATHER TODAY IS EXTREMELY GOOD” มีแนวโน้มที่จะเป็นข้อความส่อเสียดมากกว่าประโยคปกติ “The weather today is extremely good” Pt´cek et al (2014) ได้ศึกษาการจำแนกข้อความส่อเสียดออกจากข้อความปกติโดยใช้การนับจำนวนของคำที่ใช้ตัวพิมพ์ใหญ่นำมาเป็นองค์ประกอบดังตาราง 2.9

ตาราง 2.9 ตัวอย่างการสร้างตารางองค์ประกอบโดยใช้จำนวนคำที่ใช้พิมพ์ใหญ่ทั้งหมด

ประโยค	**จำนวนคำที่ใช้พิมพ์ใหญ่ทั้งหมด**
THE WEATHER IS EXTREMELY GOOD	5
The weather is nice today	0

หลังจากนั้นจึงนำค่าของแต่ละองค์ประกอบมาผ่านการจำแนกโดยใช้ MaxEnt และ SVM ซึ่งประสิทธิภาพของการใช้องค์ประกอบต่าง ๆ ร่วมในการจำแนกข้อความส่อเสียดสามารถสรุปได้ดังตารางที่ 2.10

ตาราง 2.10 ตัวอย่างของการใช้องค์ประกอบต่าง ๆ ร่วมในการจำแนกข้อความส่อเสียดด้วยเทคนิค MaxEnt และ SVM (Pt´cek, Habernal, & Hong, 2014)

	ความแม่นยำ (%)	
องค์ประกอบ (Feature)	**Maximum Entropy**	**Support Vector Machine**
Baseline (n-gram)	93.28	**92.86**
Baseline + สัญลักษณ์พิเศษ	93.32	92.84
Baseline + ไอคอนแสดงอารมณ์	**93.97**	91.66
Baseline + จำนวนตัวพิมพ์ใหญ่	93.96	91.54

ในปัจจุบันได้มีการศึกษามากมายเกี่ยวกับการวิเคราะห์ข้อความส่อเสียดเช่น การวิเคราะห์ข้อความส่อเสียดของภาษาเช็กและภาษาอังกฤษในทวิตเตอร์ (Gonzlez-Ibez et al., 2011; Pt´cek et al., 2014) โดยใช้วิธีการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning) โดยใช้ข้อมูลในรูปแบบ n-gram และคุณลักษณะเฉพาะของภาษาซึ่งรวมถึงเครื่องหมายเว้นวรรคตอนต่าง ๆ สัญลักษณ์ทางอารมณ์และตัวอักษรขึ้นต้นคำในการจำแนกข้อความส่อเสียดและทำการวัดผล ซึ่งพบว่าภาษาอังกฤษนั้นมีความ

แม่นยำถึงร้อยละ 93 ซึ่งมากกว่าภาษาเช็กที่มีความแม่นยำเพียงร้อยละ 55 ซึ่งมีเหตุผลมาจากความซับซ้อนของภาษา

นอกจากการวิเคราะห์ข้อความส่อเสียดโดยใช้คุณลักษณะของภาษาแล้วยังมีการวิเคราะห์โดยใช้คุณลักษณะอื่น ๆ ที่แตกต่างกันไปของผู้เผยแพร่ข้อความ ผู้อ่าน และสิ่งแวดล้อมเช่น ประวัติการเผยแพร่ข้อความ หัวข้อที่ผู้เผยแพร่ข้อความมีความรู้ ประวัติเกี่ยวกับความคิดเห็นของผู้เผยแพร่ข้อความ โปรไฟล์ของผู้เผยแพร่ข้อความ รวมถึงความสัมพันธ์ระหว่างผู้อ่านกับผู้เผยแพร่ข้อความมาใช้ในการวิเคราะห์ข้อความส่อเสียดซึ่งพบว่าคุณลักษณะดังกล่าวมีส่วนช่วยให้การวิเคราะห์ข้อความส่อเสียดมีประสิทธิภาพมากขึ้นถึงร้อยละ 10 ถ้าเทียบกับการใช้การวิเคราะห์ข้อความส่อเสียดโดยใช้ลักษณะของภาษาเพียงอย่างเดียว (Bamman & Smith, 2015) อย่างไรก็ตามองค์ประกอบดังกล่าวมักจะเป็นองค์ประกอบที่ขึ้นอยู่กับลักษณะเฉพาะของแต่ละภาษา อีกทั้งการสำรวจประวัติหรือโปรไฟล์ของผู้เผยแพร่ข้อความในเครือข่ายสังคมออนไลน์ทวิตเตอร์นั้นค่อนข้างทำได้ยากและไม่ชัดเจน เนื่องจากเครือข่ายสังคมออนไลน์ทวิตเตอร์นั้นไม่มีพื้นที่ให้เก็บประวัติส่วนตัวเหมือนเครือข่ายสังคมออนไลน์อื่น ๆ เช่น เฟสบุ๊ค

2.6.3 ข้อจำกัดของการศึกษาที่ผ่านมา

จากการศึกษาที่ผ่านมาพบว่าการจำแนกข้อความส่อเสียดนั้นมักจะมีการศึกษากันอย่างแพร่หลายในภาษาต่างประเทศต่าง ๆ เช่น ภาษาอังกฤษ (Bamman & Smith, 2015; Gonzlez-Ibez et al., 2011; Pt´cek et al., 2014) แต่ไม่มีการศึกษาในหัวข้อนี้มากนักในภาษาไทย ซึ่งส่งผลให้การวิเคราะห์ความคิดเห็นในภาษาไทยนั้นมีความผิดพลาดเกิดขึ้นในกรณีที่ข้อความที่ทำการวิเคราะห์นั้นเป็นข้อความส่อเสียด

ผู้วิจัยมีความเห็นว่าการใช้องค์ประกอบอื่น ๆ ในการจำแนกข้อความส่อเสียด เช่น การใช้สัญลักษณ์พิเศษต่าง ๆ ไอคอนแสดงอารมณ์ หรือการใช้ตัวอักษรพิมพ์ใหญ่ติดกัน มักจะขึ้นอยู่กับ ลักษณะและวัฒนธรรมของแต่ละภาษาและบุคคลเช่น ในภาษาไทย จะไม่ปรากฏลักษณะของการใช้ตัวอักษรพิมพ์ใหญ่ซึ่งจะส่งผลให้เทคนิคการจำแนกข้อความส่อเสียดโดยใช้องค์ประกอบอื่น ๆ ที่ไม่ใช่คำอาจไม่เหมาะสมกับการนำมาประยุกต์กับภาษาไทย ดังนั้น การใช้ความน่าจะเป็นของประโยคในการจำแนกข้อความส่อเสียดออกจากข้อความปกติ อาจเป็นอีกวิธีหนึ่งที่น่าสนใจศึกษา โดยความน่าจะเป็นที่ประโยคหนึ่ง ๆ จะเป็นประโยคส่อเสียดนั้นน่าจะต่ำกว่าประโยคทั่วไปเนื่องจากอัตราส่วนของประโยคส่อเสียดจะค่อนข้างน้อยเปรียบเทียบกับจำนวนของประโยคทั่ว ๆ ไป (Hsu & Jain, 2015) อีกทั้งคำที่ปรากฏร่วมกันในข้อความส่อเสียดนั้นจะมีแนวโน้มว่าเป็นคำที่ไม่เป็นไปตามธรรมชาติ เช่น การใช้คำว่า ความเร็วแสงกับอินเตอร์เน็ต ในข้อความ “อินเตอร์เน็ตยี่ห้อนี้เร็วมากเร็วกว่าความเร็วแสงอีก” ดังนั้นหากข้อความไหนมีความน่าจะเป็นซึ่งคำนวณ

จากองค์ประกอบและลำดับของคำที่ใช้ในข้อความในระดับที่ต่ำ นั่นหมายความว่าข้อความนั้นมีโอกาสเกิดที่ต่ำ และมีความเป็นไปได้ที่ข้อความดังกล่าวจะเป็นข้อความเชิงส่อเสียด

ผู้วิจัยจึงมีความสนใจที่จะศึกษาประสิทธิภาพของการประยุกต์ใช้ความน่าจะเป็นของข้อความในการจำแนกข้อความส่อเสียดออกจากข้อความปกติรวมถึงศึกษาข้อจำกัดและปัจจัยที่เกี่ยวข้อง โดยมีสมมุติฐานที่ว่า หากข้อความนั้นมีความน่าจะเป็นของข้อความน้อย ซึ่งเกิดจากข้อความนั้นประกอบด้วยลำดับของคำที่ผิดปกติ ข้อความนั้นจะมีแนวโน้มที่จะเป็นข้อความส่อเสียดมากกว่าข้อความอื่น ๆ ที่มีความน่าจะเป็นของข้อความมากกว่า

บทที่ 3
ระเบียบวิธีวิจัย

วัตถุประสงค์หลักของการศึกษาในครั้งนี้เพื่อศึกษา และพัฒนาตัวแบบในการจำแนกข้อความส่อเสียดออกจากข้อความปกติโดยใช้ความน่าจะเป็นของข้อความ

รายละเอียดที่นำเสนอในบทของระเบียบวิจัยนี้จึงเป็นการดำเนินการเพื่อตอบวัตถุประสงค์หลักข้างต้น ได้แก่แบบจำลอง N-gram โดยใช้ความน่าจะเป็นของข้อความ ข้อมูลที่ใช้ในการทดลอง วิธีการเก็บข้อมูล เครื่องมือที่ใช้ในการทดลอง แผนการทดลอง การวิเคราะห์ผลการทดลอง ประเด็นของความเชื่อถือได้ (Reliability) และความถูกต้อง (Validity) ของข้อมูล

3.1 การคำนวณความน่าจะเป็นของข้อความโดยใช้แบบจำลอง N-gram

เนื่องจากข้อความในทวิตเตอร์นั้นมีความยาวจำกัด ดังนั้นเพื่อที่จะลดรูปแบบข้อมูลให้เหมาะสมที่จะใช้ในการคำนวณหาความน่าจะเป็นของข้อความ ผู้วิจัยจึงตัดสินใจที่จะแปลงข้อมูลให้อยู่ในลักษณะ Bigram ซึ่งแบบจำลอง N-gram เมื่อ N มีค่าเท่ากับ 2 (Bigram) สามารถสร้างได้โดยใช้คำศัพท์จากการตัดคำซึ่งอยู่ติดกันในข้อความโดยเพิ่มสัญลักษณ์เริ่มต้นและจบข้อความเข้าไป จากนั้นจึงนำมาสร้างเป็นตาราง Unigram ยกตัวอย่างเช่น ในข้อความ “เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้าเย่”[10] และข้อความ “โห้เน็ตXเพื่อนๆมันเยี่ยมจริงจริง” สามารถนำมาสร้างเป็นตาราง Unigram โดยใช้การตัดคำแบบยาวที่สุด (Longest Matching) ได้ดังตารางที่ 3.1 โดยจะทำการตัดคำที่ LexTo ไม่รู้จักออก เนื่องจากคำที่ LexTo ไม่รู้จัก อาจส่งผลให้ความน่าจะเป็นของข้อความต่ำกว่าที่ควรจะเป็น โดยในการศึกษาครั้งนี้ หน่วย (unit) ของ gram ที่ใช้ คือคำยาวที่สุดที่รู้จักโดย LexTo ยกตัวอย่างเช่น ข้อความ “เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้าเย่” จะสามารถตัดคำโดยใช้ LexTo ได้เป็น “<s> | เน็ต | X | ยัง | ใช้ไม่ได้ | เลย | เริ่ด | จ้า | </s>” แทนที่จะเป็น “<s> | เน็ต | X | ยัง | ใช้ไม่ได้ | เลย | เริ่ด | จ้า | **เย่** | </s> โดยคำว่า ‘เย่’ จะถูกตัดออกไป

[10] ข้อมูลที่นำมาใช้มีวัตถุประสงค์เพื่อใช้ในการศึกษาการจำแนกข้อความส่อเสียดออกจากข้อความปกติเท่านั้น ผู้วิจัยจึงขอทำการแทนชื่อผู้ให้บริการอินเตอร์เน็ตที่ใช้ในการศึกษาด้วยสัญลักษณ์ X

ตาราง 3.1 ตัวอย่างโครงสร้างข้อมูลในลักษณะ Unigram (N-gram, n = 1) จากข้อความ

Sentence	Unigram
เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้า	<s> | เน็ต | X | ยัง | ใช้ไม่ได้ | เลย | เริ่ด | จ้า | </s>
โห้เน็ตXเพื่อนๆมันเยี่ยมจริง จริง	<s> | โห้ | เน็ต | X | เพื่อนๆ | มัน | เยี่ยม | จริง | จริง | </s>
.....	.....

จากนั้นจึงแปลงข้อมูลให้อยู่ในลักษณะ Bigram (N-gram, n = 2) เพื่อนำคำที่ปรากฏไปใช้ในการคำนวณความน่าจะเป็นของคู่ของคำแต่ละคู่ เพื่อนำไปประกอบการคำนวณความน่าจะเป็นของข้อความต่อไป ดังโครงสร้างข้อมูลในลักษณะ Bigram ซึ่งแสดงในตารางที่ 3.2

ตาราง 3.2 ตัวอย่างโครงสร้างข้อมูลในลักษณะ Bigram จากข้อความ

Sentence	Bigram
เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้า	“<s>” , “เน็ต” | “เน็ต” , “X” | “X” , “ยัง” | “ยัง” , “ใช้” | “ใช้” , “ไม่ได้” | “ไม่ได้” , “เลย” | “เลย” , “เริ่ด” | “เริ่ด” , “จ้า” | “จ้า” , “</s>”
โห้เน็ตXเพื่อนๆมันเยี่ยมจริง จริง	“<s>” , “โห้” | “โห้” , “เน็ต” | “เน็ต” , “X” | “X” , “เพื่อนๆ” | “เพื่อน” , “มัน” | “มัน” , “เยี่ยม” | “เยี่ยม” , “จริง” | “จริง” , “จริง” | “จริง” , “</s>”
.....	.....

จากนั้นจึงสร้างตารางความน่าจะเป็นของข้อความโดยใช้สมการที่ 2.5 เพื่อหาความน่าจะเป็นของคู่คำศัพท์แต่ละคู่

$$P\left(w_i \mid w_{i-1}\right) = \frac{Count\left(w_{i-1}, w_i\right)}{Count\left(w_{i-1}\right)} \qquad \text{สมการที่ 2.5}$$

โดยกำหนดให้ P คือค่าความน่าจะเป็นของลำดับคำที่สนใจและ w_i คือคำศัพท์ตำแหน่งที่ i ของข้อความเช่น “<s>เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้าเย่</s>” เมื่อ i มีค่าเป็น 2 ดังนั้น w_2 จะมีค่าเท่ากับ “เน็ต” โดยสามารถเขียนในรูปแบบตารางได้ดังตารางที่ 3.3

ตาราง 3.3 ตัวอย่างแสดงตำแหน่งคำศัพท์ในข้อความ

i	1	2	3	4	5	6	7	8	9	10
w_i	<s>	เน็ต	X	ยัง	ใช้	ไม่ได้	เลย	เริ่ด	จ้า	</s>

และสามารถหาความน่าจะเป็นของคู่แรก หรือ “<s>” , “เน็ต” หรือ ความน่าจะเป็นที่คำว่า “เน็ต” จะปรากฏเป็นคำแรกในข้อความ โดยสามารถคำนวณได้โดย

$$P(\text{เน็ต} \mid \langle s \rangle) = \frac{Count(\langle s \rangle, \text{เน็ต})}{Count(\langle s \rangle)}$$

ซึ่งก็คือจำนวนคำที่ขึ้นต้นด้วยสัญลักษณ์เริ่มต้น “<s>” และตามด้วย “เน็ต” ทั้งหมด หารด้วยจำนวนของสัญลักษณ์เริ่มต้น “<s>” ทั้งหมดที่มี โดยหาความน่าจะเป็นของทุกคู่ที่เกิดขึ้นในแบบจำลอง Unigram จากนั้นจึงนำมาคำนวณหาความน่าจะเป็นของข้อความโดยนำความน่าจะเป็นของแต่ละคู่คำในข้อความมาคูณกัน ยกตัวอย่างเช่น ข้อความ “เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้าเย่” สามารถหาความน่าจะเป็นของข้อความได้ โดยกำหนดให้

S = “เน็ตXยังใช้ไม่ได้เลยเริ่ดจ้า”

ดังนั้นความน่าจะเป็นของข้อความ S สามารถหาได้โดย

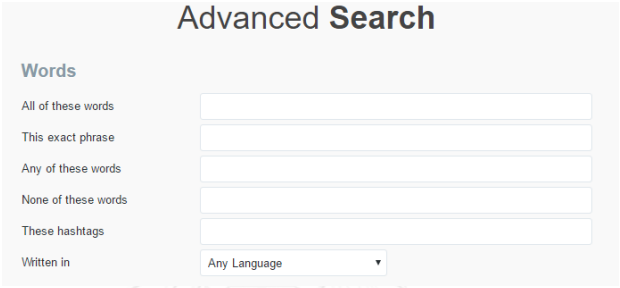
P(S) = P(เน็ต | <s>) * P(X | เน็ต) * P(ยัง | X) * P(ใช้ไม่ได้ | ยัง) *
P(เลย | ใช้ไม่ได้) * P(เริ่ด | เลย) * P(จ้า | เริ่ด) * P(</s> | จ้า)

จากนั้นจึงทำการหาความน่าจะเป็นในแต่ละข้อความของชุดข้อมูลนำมาเปรียบเทียบและหาความสัมพันธ์ระหว่างคะแนนความน่าจะเป็นที่คำนวณได้โดยเครื่องและคะแนนความน่าจะเป็นจากการประเมินโดยบุคคล

3.2 ข้อมูลที่ใช้ในการทดลอง

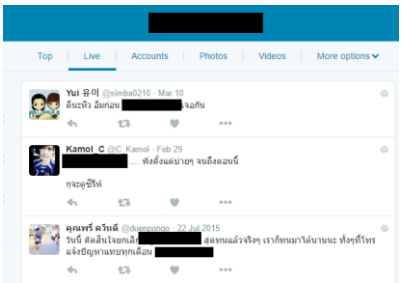
โดยการศึกษาครั้งนี้ผู้วิจัยจะใช้ข้อมูลความคิดเห็นบนเครือข่ายสังคมออนไลน์ทวิตเตอร์เป็นหลัก โดยการใช้ API ในการติดต่อขอข้อมูลย้อนหลังผ่านทาง Advanced Search ดังรูป 3.1 โดยผู้วิจัยจะเก็บข้อความทวีตเกี่ยวกับการใช้บริการอินเตอร์เน็ต ระหว่างวันที่ 25 มกราคม 2553 ถึง 9 มิถุนายน 2559 จำนวน 4,027 ข้อความ[11] ซึ่งประกอบไปด้วยข้อความปกติจำนวน 3,913 ข้อความ ข้อความที่ผู้วิจัยไม่แน่ใจว่าเป็นข้อความส่อเสียดหรือไม่ 6 ข้อความ และ ข้อความส่อเสียดจำนวน 108 ข้อความซึ่งเรียงตามลำดับเวลามาเป็นข้อมูลที่ใช้ในการศึกษาครั้งนี้

รูป 3.1 ตัวอย่างหน้าจอ Twitter Advanced Search



การรวบรวมข้อมูลทำโดยการค้นหาคีย์เวิร์ดผ่านทาง Advanced Search ของเครือข่ายสังคมออนไลน์ทวิตเตอร์โดยค้นหาชื่อผู้ให้บริการอินเตอร์เน็ต X ในประเทศไทย ควบคู่ไปกับการค้นหาโดยใช้แฮชแทกที่เกี่ยวข้อง โดยข้อความที่ใช้ในการค้นหาคือ “Xอินเตอร์เน็ต” “Xอินเตอร์เน็ท” “อินเตอร์เน็ตX” “อินเตอร์เน็ทX” “เน็ตX” “เน็ทX” รวมถึงแฮชแทก #ประชด และกำหนดให้ค้นหาข้อความที่อยู่ในภาษาไทยเท่านั้น ซึ่งให้ผลการค้นหาดังรูป 3.2 และตามรายละเอียดดังตารางที่ 3.4

รูป 3.2 ตัวอย่างผลลัพธ์การค้นหาข้อมูลผ่านทาง Advanced Search ด้วยแฮชแทก #X



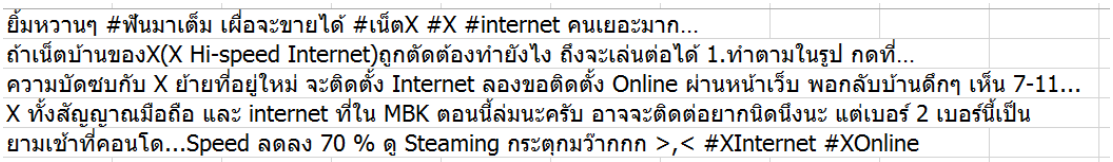
[11] เหตุผลที่ต้องรวบรวมข้อมูลย้อนหลังไปถึง 7 ปี เพราะ หากพิจารณาถึงข้อความที่มีแฮชแทก #ประชด (รายละเอียดดังตาราง ก.1 ในภาคผนวก ก.) จะพบว่าจำนวนข้อความส่อเสียดไม่ได้เพิ่มขึ้นตามเวลา กล่าวคือ การใช้ข้อความส่อเสียดไม่ได้มีแนวโน้มเพิ่มขึ้น แต่มีจำนวนมากน้อยตามคุณภาพของการบริการของอินเทอร์เน็ท X เหตุผลที่ผู้วิจัยรวบรวมข้อมูลย้อนหลังไปถึง 7 ปี คือ เพื่อให้ได้ข้อความส่อเสียดในปริมาณที่มากเพียงพอสำหรับการศึกษา

ตาราง 3.4 รายละเอียดการเก็บข้อมูลโดยการค้นหาด้วย Advanced Search

คีย์เวิร์ด	**จำนวน ข้อความปกติ**	**จำนวน ข้อความที่ไม่แน่ใจ**	**จำนวน ข้อความส่อเสียด**	**รวม**
Xอินเตอร์เน็ต	632	2	25	659
Xอินเตอร์เน็ท	75	0	0	75
อินเตอร์เน็ตX	431	1	21	453
อินเตอร์เน็ทX	40	1	1	42
เน็ตX	1,576	1	34	1,611
เน็ทX	1,159	1	27	1,187
--- รวม ---	**3,913**	**6**	**108**	**4,027**

หลังจากที่ได้ข้อมูลความคิดเห็นของผู้ใช้งานอินเทอร์เน็ตของผู้ให้บริการที่เป็นเป้าหมายแล้ว ผู้วิจัยจึงกรองข้อมูลที่ได้จากการค้นหาโดยการอ่านและตัดสินใจว่าข้อความข้างต้นเป็นข้อความความคิดเห็นที่เกี่ยวเนื่องกับผู้ให้บริการอินเตอร์เน็ตที่สนใจหรือไม่ ซึ่งหากเป็นข้อมูลที่ไม่เกี่ยวข้องกับผู้ให้บริการอินเตอร์เน็ตที่สนใจจะทำการตัดทวีตดังกล่าวทิ้ง โดยจำนวนข้อความหลังการกรองที่ทั้งสิ้น 3,971 ข้อความ โดยหลังจากกรองข้อมูลความคิดเห็นที่ได้รับจากการใช้ Advanced Search เสร็จเรียบร้อย จึงนำข้อมูลความคิดเห็นของผู้ใช้บริการอินเตอร์เน็ตไปเก็บรวบรวมไว้ในตาราง Microsoft Excel ดังรูปที่ 3.3 เพื่อที่นำไปใช้ในขั้นตอนการประมวลผลต่อไป

รูป 3.3 ตัวอย่างข้อมูลความคิดเห็นผู้ใช้บริการอินเตอร์เน็ตจากการกรองโดยผู้วิจัย



3.3 การจำแนกข้อความส่อเสียดโดยบุคคล

อย่างไรก็ตามการจำแนกข้อความส่อเสียดนั้นผู้อ่านอาจไม่สามารถทราบเจตนาของผู้เขียนได้ถูกต้อง 100% และในทางปฏิบัตินั้นผู้เขียนข้อความก็ไม่จำเป็นที่จะต้องระบุว่าข้อความที่ตนเองเขียนนั้นเป็นข้อความส่อเสียดหรือไม่ ดังนั้นผู้วิจัยเห็นว่าในการศึกษาครั้งนี้ควรกำหนดให้บุคคลอื่นที่มีประสบการณ์การใช้เครือข่ายสังคมออนไลน์ทวิตเตอร์จำนวน 5 คน มาทำการตัดสินว่าข้อความที่ได้รวบรวมมาแต่ละข้อความนั้นเป็นข้อความส่อเสียด ไม่เป็นข้อความส่อเสียด หรือไม่แน่ใจว่าเป็นข้อความส่อเสียดหรือไม่ โดยให้แต่ละคน

อ่านข้อความที่ได้รวบรวมไว้และทำเครื่องหมายในแต่ละข้อความเพื่อแสดงให้เห็นว่าข้อความดังกล่าวเป็นข้อความส่อเสียดหรือไม่ จากนั้นนำผลที่ได้มาสรุปรวมกันโดย

หากบุคคลที่ทำให้คะแนน**เห็นว่าข้อความดังกล่าวเป็นข้อความส่อเสียด** บุคคลนั้นจะให้คะแนนความน่าจะเป็นของข้อความนั้นด้วยเลข 0

หากบุคคลที่ให้คะแนน**ไม่แน่ใจว่าข้อความดังกล่าวเป็นข้อความส่อเสียด** บุคคลนั้นจะให้คะแนนความน่าจะเป็นของข้อความนั้นด้วยเลข 0.5

และหากบุคคลที่ให้คะแนน**เห็นว่าข้อความนั้นเป็นข้อความปกติ** บุคคลนั้นจะให้คะแนนความน่าจะเป็นของข้อความด้วยเลข 1

ซึ่งจะนำค่าน้ำหนักเหล่านั้นมาศึกษาความสามารถในของการจำแนกข้อความส่อเสียดออกจากข้อความปกติของเทคนิคที่ผู้วิจัยศึกษา

โดยบุคคลที่มีหน้าที่ในการทำเครื่องหมายจำแนกข้อความส่อเสียดนั้นมีคุณสมบัติดังตารางที่ 3.5 และตัวอย่างการสรุปคะแนนความน่าจะเป็นของข้อความที่ประเมินโดยบุคคลแสดงได้ดังตารางที่ 3.6 โดยจะแบ่งการให้คะแนนความน่าจะเป็นของข้อความออกเป็นครั้ง ๆ โดยในแต่ละครั้งจะให้บุคคลอื่นให้คะแนนความน่าจะเป็นทีละ 100 ข้อความเพื่อให้บุคคลที่ทำการให้คะแนนไม่เกิดความเอียงเอนในกรณีที่ต้องให้คะแนนความน่าจะเป็นจำนวนมากในครั้งเดียว ซึ่งบุคคลทั้ง 5 คนนั้นจะเป็นบุคคลเดียวกันในการให้คะแนนแต่ละครั้ง โดยหลังการให้คะแนนความน่าจะเป็นในแต่ละครั้ง ผู้วิจัยจะบันทึกเวลาที่ใช้ในการให้คะแนนความน่าจะเป็นของแต่ละบุคคลในแต่ละครั้ง เพื่อใช้ในการอ้างอิงในการศึกษาครั้งนี้

ตาราง 3.5 คุณสมบัติบุคคลที่ทำเครื่องหมายจำแนกข้อความส่อเสียด

	บุคคลที่ 1	**บุคคลที่ 2**	**บุคคลที่ 3**	**บุคคลที่ 4**	**บุคคลที่ 5**
เพศ	ชาย	ชาย	ชาย	หญิง	หญิง
อายุ	24	24	24	23	25
ประสบการณ์การใช้เครือข่ายสังคมออนไลน์	Facebook, Twitter	Facebook, Twitter	Facebook, Twitter	Facebook, Twitter	Facebook, Twitter
จำนวนปีที่ใช้เครือข่ายสังคมออนไลน์	6 ปี	7 ปี	5 ปี	5 ปี	6 ปี

	บุคคลที่ 1	บุคคลที่ 2	บุคคลที่ 3	บุคคลที่ 4	บุคคลที่ 5
เวลาเฉลี่ยที่ใช้เครือข่ายสังคมออนไลน์ต่อวัน	3 ชั่วโมง	4 ชั่วโมง	3 ชั่วโมง	3 ชั่วโมง	3 ชั่วโมง
การศึกษา	ปริญญาตรี สาขาวิทยาการคอมพิวเตอร์	ปริญญาตรี สาขาวิทยาการคอมพิวเตอร์	ปริญญาตรี สาขา วิศวกรรมคอมพิวเตอร์	ปริญญาตรี สาขาการจัดการและการท่องเที่ยว	ปริญญาตรี สาขาจุลชีววิทยา
อาชีพ	โปรแกรมเมอร์	พนักงานขาย	โปรแกรมเมอร์	อาชีพอิสระ	อาจารย์สอนพิเศษ

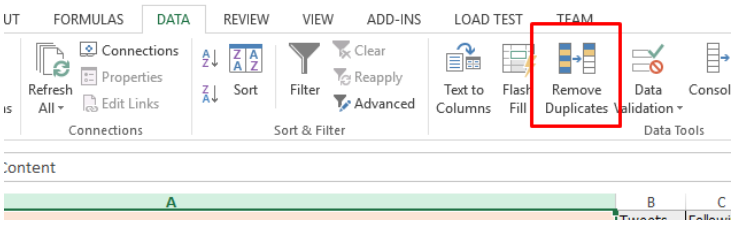
ตาราง 3.6 ตัวอย่างการสรุปคะแนนความน่าจะเป็นของข้อความ

	บุคคลที่ 1	**บุคคลที่ 2**	**บุคคลที่ 3**	**บุคคลที่ 4**	**บุคคลที่ 5**	**ค่าคะแนนเฉลี่ย**
อินเตอร์เน็ตความเร็วสูง สูงกว่าเต่านิดนึง	0	0	0	0	0	0
วันนี้อากาศร้อนจัง	1	1	1	1	1	1
เน็ตX เน็ตความเร็วแสง	0	0	0	0	0	0
เน็ตX เร็วจริงๆ 5555 ---*	0	0.5	0	0.5	0	0.2
สัญญาณXเจ๋งมากกกกกกกก!!!	0	0.5	0	0	0.5	0.2

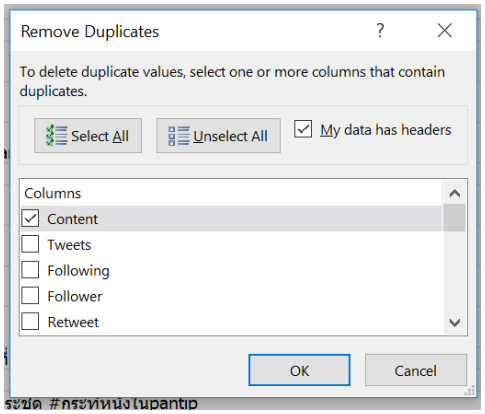
3.4 การประมวลข้อมูลสำหรับการจำแนกด้วยเทคนิคที่พัฒนาขึ้น

เนื่องจากข้อมูลที่ได้รวบรวมมาผ่านการค้นหาโดยใช้ข้อความการค้นหาที่หลากหลาย อาจส่งผลให้ข้อมูลบางส่วนมีความซ้ำซ้อนกันเกิดขึ้น จึงต้องมีการตัดข้อมูลที่เกิดขึ้นซ้ำกันโดยใช้เครื่องมือ Remove Duplicates ของ Microsoft Excel ดังรูปที่ 3.3 และเลือกเฉพาะแถวที่ต้องการลบข้อมูลซ้ำซ้อน ดังรูปที่ 3.4 โดยจำนวนข้อความหลังจากการตัดข้อความซ้ำมีจำนวน 3,840 ข้อความ

รูป 3.3 การใช้งานฟังก์ชัน Remove Duplicates



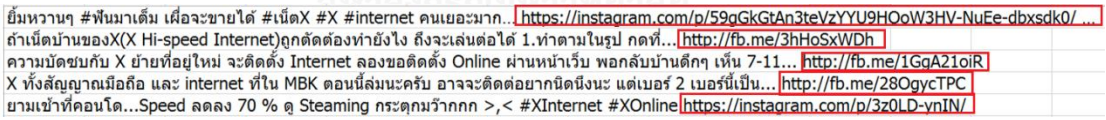
รูป 3.4 การลบข้อมูลซ้ำซ้อน



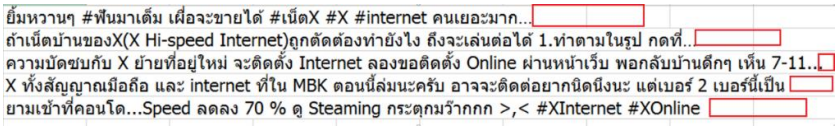
จากนั้นจึงตัดเฉพาะส่วนของข้อความทวีตซึ่งขึ้นต้นด้วย URL (ส่วนที่ขึ้นต้นด้วย http://) ทิ้ง เนื่องจากข้อความทวีตซึ่งประกอบไปด้วย URL นั้นไม่สามารถนำมาประมวลผลได้เนื่องจากส่วนของ URL เช่น http://www.google.co.th ไม่ใช่ข้อความที่แสดงความคิดเห็นดังนั้นจึงไม่น่าจะเกี่ยวข้องกับการแสดงความคิดเห็นเชิงส่อเสียด ดังรูปที่ 3.5

รูป 3.5 ตัวอย่างข้อความทวีตก่อนและหลังตัด URL

ก่อน



หลัง

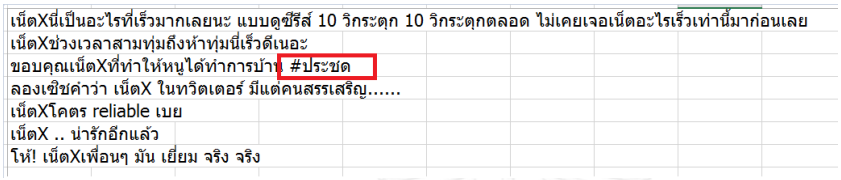


ขั้นตอนถัดไปคือการตัดคำที่ขึ้นต้นด้วยเครื่องหมายแฮชแทกออกจากข้อความ เนื่องจากคำเหล่านั้นมักเป็นคำที่ไว้ใช้สำหรับแสดงหมวดหมู่ของข้อความเช่น “อินเตอร์เน็ทวันนี้ไม่ดีเลย #เน็ทABC” แฮชแท็ก #เน็ทABC แสดงถึงว่าพูดเขียนข้อความกล่าวถึงอินเตอร์เน็ต ยี่ห้อ ABC ซึ่งไม่ใช่สาระสำคัญที่สามารถใช้ได้ในการจำแนกข้อความส่อเสียด โดยจะตัดเฉพาะแฮชแทกที่อยู่ตำแหน่งสุดท้ายของข้อความเท่านั้น เนื่องจาก

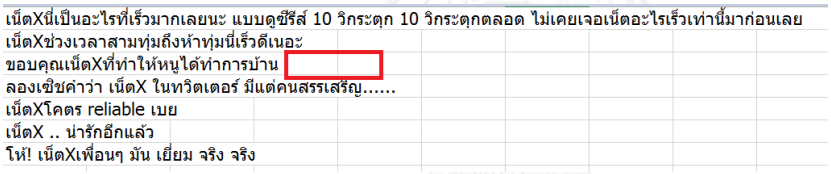
ตามสถิติข้อมูลที่ผู้วิจัยได้รวบรวมมานั้น ส่วนมากแฮชแท็กที่อยู่ท้ายข้อความนั้นมักจะแสดงถึงหมวดหมู่ของข้อความเท่านั้นและไม่ได้เป็นส่วนนึงของข้อความโดยตรง โดยสามารถตัดแฮชแท็กออก ได้ดังรูปที่ 3.6 โดยหากแฮชแท็ก อยู่ระหว่างข้อความ จะทำการตัดเฉพาะเครื่องหมายแฮชแท็กออก เช่น “เน็ต #X ช้าจังเลยวันนี้ #XInternet” แฮชแท็ก #XInternet จะถูกตัดออก ส่วนแฮชแท็ก #X นั้นจะลบเฉพาะเครื่องหมายแฮชแท็ก ซึ่งให้ผลลัพธ์คือ “เน็ต X ช้าจังเลยวันนี้”

รูป 3.6 ตัวอย่างข้อความทวีตก่อนและหลังตัดแฮชแทก

ก่อน



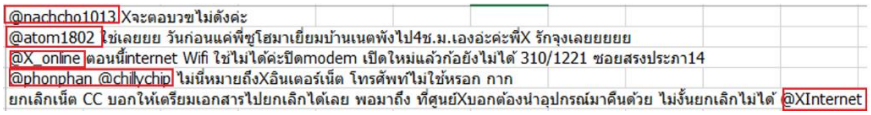
หลัง



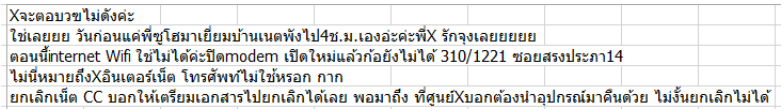
จากนั้นจึงตัดส่วนที่ขึ้นต้นด้วยเครื่องหมาย At (@) ทิ้ง เนื่องจากส่วนที่ขึ้นต้นด้วยเครื่องหมาย @ นั้นจะแสดงถึงรายชื่อบุคคลที่ทำการติดต่อสื่อสารกับผู้เขียนข้อความโดยตรง ซึ่งชื่อบุคคลนั้นไม่สามารถนำมาเป็นองค์ประกอบในการจำแนกข้อความส่อเสียดได้ ดังรูปที่ 3.7

รูป 3.7 ตัวอย่างข้อความทวีตก่อนและหลังตัดเครื่องหมาย At (@)

ก่อน



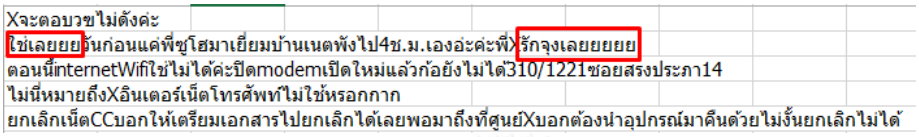
หลัง



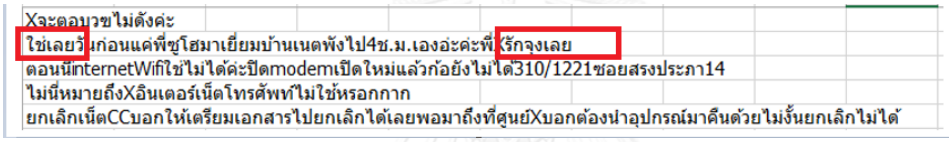
จากนั้นจะเป็นการตัดตัวอักษรของคำที่มีจำนวนตัวอักษรท้ายคำซ้ำกันเกิน 3 ตัว (Rajadesingan et al., 2015) เช่น “ช้ามากกกกกกกกกกกกกก” จะถูกตัดเป็น “ช้ามาก” หรือตัวเลขเช่น “55555555555” จะถูกตัดทิ้ง ดังรูป 3.8

รูป 3.8 ตัวอย่างข้อความทวีตก่อนและหลังตัดคำซ้ำกัน

ก่อน



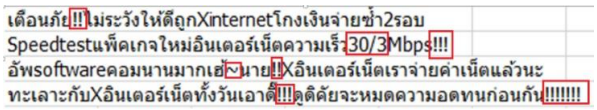
หลัง



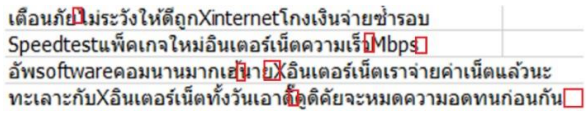
ขั้นตอนถัดไปคือการตัดตัวอักษรหรืออักขระพิเศษต่าง ๆ เช่นเครื่องหมายอัศเจรีย์ (!) หรือเครื่องหมายปรัศณี (?) ออกจากข้อความ เนื่องจากลักษณะของภาษาไทยมักไม่ค่อยปรากฏการใช้สัญลักษณ์พิเศษ โดยตัวอย่างการตัดสัญลักษณ์พิเศษปรากฏดังรูปที่ 3.9

รูป 3.9 ตัวอย่างข้อความทวีตก่อนและหลังตัดอักขระพิเศษ

ก่อน



หลัง



ขั้นตอนต่อไปของการประมวลผลเบื้องต้น เนื่องจากภาษาไทยเป็นภาษาที่ไม่ตัดคำ ทำให้ต้องมีการตัดคำออกเป็นส่วน ๆ ออกจากข้อความก่อนซึ่งในกระบวนการวิจัยนี้ใช้วิธีการตัดคำแบบ Longest Matching ซึ่งคือการตัดคำโดยอาศัยคำที่ยาวที่สุดซึ่งปรากฏอยู่ในพจนานุกรมก่อนดังรายละเอียดที่ได้กล่าวไว้ในบทที่ 2 หัวข้อที่ 2.1.3

ซึ่งกระบวนการตัดคำนี้จะใช้ API ของ NECTEC ซึ่งใช้ชื่อว่า LexTo – Thai Lexeme Tokenizer [12] ซึ่งเป็นการตัดคำโดยใช้เทคนิคตัดคำที่ยาวที่สุดโดยเทียบจากพจนานุกรมดังที่ได้กล่าวไว้ในบทที่ 2 โดยโปรแกรมจะเก็บรายการข้อความที่ต้องการตัดคำทั้งหมด จากนั้นจึงส่งข้อความดังกล่าวทีละข้อความไปยัง LexTo เพื่อทำการตัดคำ และโปรแกรมจะนำข้อความที่ผ่านการตัดคำแล้วมาบันทึกเก็บไว้ จากนั้นผู้วิจัยจะคัดลอกข้อความทั้งหมดที่ตัดคำแล้วมาเก็บในตาราง Excel ดังรูปที่ 3.10 และ 3.11 เพื่อใช้ในการประมวลผลและสร้างแบบจำลอง N-gram ต่อไป

รูป 3.10 ตัวอย่างการใช้ LexTo ในการตัดคำจากข้อความ

[12] http://www.sansarn.com/lexto/

รูป 3.11 ตัวอย่างการเก็บข้อมูลลงใน Excel

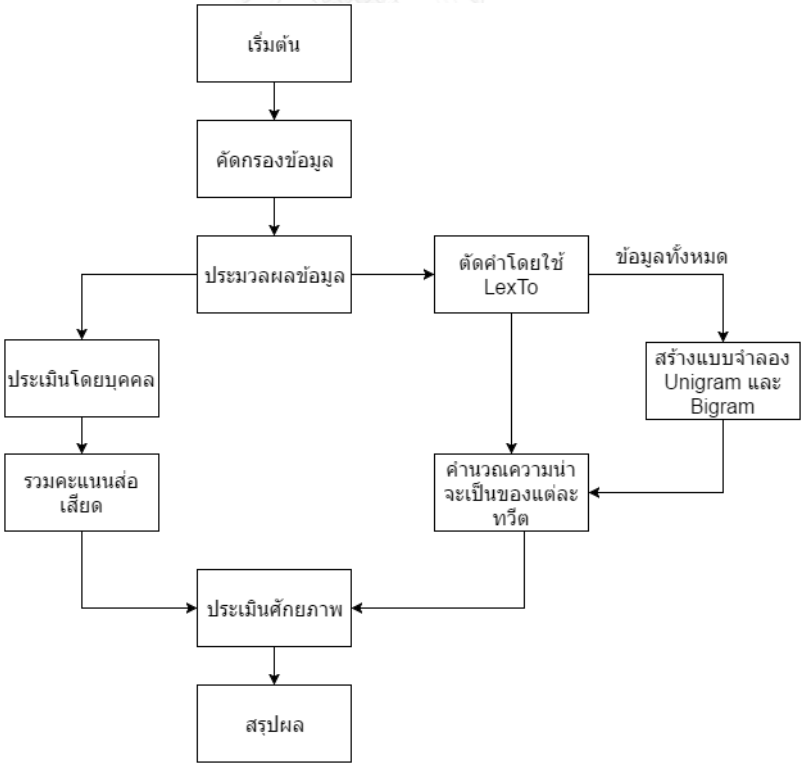
Content
เตือนภัยไม่ระวังให้ดีถูกXinternetโกงเงินจ่ายซ้ำรอบ
Speedtestแพ็คเกจใหม่อินเตอร์เน็ตความเร็วMbps
อัพsoftwareคอมนานมากเฮ้นายXอินเตอร์เน็ตเราจ่ายค่าเน็ตแล้วนะ
ทะเลาะกับXอินเตอร์เน็ตทั้งวันเอาดิ้ดูดิคัยจะหมดความอดทนก่อนกัน

Tokenize
เตือนภัย | ไม่ | ระวัง | ให้ | ดี | ถูก | Xinternet | โกงเงิน | จ่าย | ซ้ำ | รอบ |
Speedtest | แพ็คเกจ | ใหม่ | อินเตอร์เน็ต | ความเร็ว | Mbps |
อัพ | software | คอม | นาน | มาก | เฮ้ | นาย | X | อินเตอร์เน็ต | เรา | จ่าย | ค่า | เน็ต | แล้ว | นะ
ทะเลาะ | กับ | X | อินเตอร์เน็ต | ทั้งวัน | เอา | ดิ้ | ดู | ดิ | คัย | จะ | หมด | ความอดทน | ก่อน | กัน |

3.5 ขั้นตอนการทดลอง

ในการทดลองนี้ ผู้วิจัยจะแบ่งการทดลองออกเป็นสองส่วนและนำมาเปรียบเทียบกันระหว่างผลสรุปการให้คะแนนความน่าจะเป็นโดยตัวบุคคลและการคำนวณความน่าจะเป็นของข้อความโดยเครื่อง เพื่อให้ทราบว่าเทคนิคที่ศึกษามีศักยภาพในการนำมาประยุกต์กับการจำแนกข้อความส่อเสียดมากน้อยเพียงใด โดยสามารถสรุปรูปแบบการทำงานได้ดังรูปที่ 3.12

รูป 3.12 ขั้นตอนการทำงานของการจำแนกข้อความส่อเสียด



ในส่วนของการประเมินโดยบุคคลนั้น จะดำเนินการโดยบุคคลทั้งห้าคนประเมินข้อความแต่ละข้อความว่าข้อความดังกล่าวเป็นข้อความเชิงส่อเสียดหรือไม่และสรุปรวมคะแนนความน่าจะเป็นในแต่ละข้อความโดยนำคะแนนของบุคคลที่ทำการประเมินทั้งห้ามาเฉลี่ยรวมกัน

ในส่วนที่สองจะเป็นส่วนที่เกี่ยวข้องกับเทคนิคการจำแนกข้อความส่อเสียดที่ได้พัฒนาขึ้นโดยอันดับแรกจะประมวลผลข้อมูลเบื้องต้นก่อนตามที่ระบุไว้ในหัวข้อ 3.4

หลังจากประมวลผลเบื้องต้นแล้วจึงทำการคำนวณความน่าจะเป็นของข้อความแต่ละข้อความโดยใช้ N-gram ในระดับคำ โดยเริ่มจากการสร้าง Unigram โดยการใช้เครื่องมือ LexTo ในการตัดคำ จากนั้นจึงนำ Unigram ที่ได้มาสร้างเป็น Bigram เพื่อหาความน่าจะเป็นของข้อความดังรายละเอียดในสมการที่ 2.4

ในขั้นตอนสุดท้ายจะเป็นการสรุปผลการทดลองเพื่อสรุปประสิทธิภาพของการจำแนกข้อความส่อเสียดโดยเทคนิคที่ศึกษา โดยเปรียบเทียบกันระหว่างผลสรุปการให้คะแนนความน่าจะเป็นโดยบุคคลและการคำนวณความน่าจะเป็นโดยเครื่องว่ามีความสอดคล้องกันมากน้อยเพียงใด โดยหากผู้เชี่ยวชาญเห็นว่าข้อความดังกล่าวเป็นข้อความส่อเสียด คะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลจะเข้าใกล้ 0 แต่หากผู้เชี่ยวชาญเห็นว่าข้อความดังกล่าวเป็นข้อความปกติ คะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลนั้นจะเข้าใกล้ 1 และเช่นเดียวกันกับการคำนวณความน่าจะเป็นโดยเครื่อง หากความน่าจะเป็นของข้อความเข้าใกล้ 0 ข้อความนั้นจะมีแนวโน้มที่จะเป็นข้อความเชิงส่อเสียด หากความน่าจะเป็นของข้อความเข้าใกล้ 1 ข้อความนั้นจะมีแนวโน้มที่จะเป็นข้อความปกติ และเพื่อให้ทราบถึงความสัมพันธ์ดังกล่าวผู้วิจัยจึงเลือกใช้การทดสอบทางสถิติด้วย Pearson Correlation Test

3.6 การวิเคราะห์ผลการทดลอง

ในการวิเคราะห์ประสิทธิภาพของการจำแนกข้อความส่อเสียดออกจากข้อความปกตินั้นจะใช้การวัดความสัมพันธ์ระหว่างผลสรุปของคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลและคะแนนความน่าจะเป็นที่ได้จากการคำนวณโดยเครื่องว่าความสัมพันธ์ดังกล่าวเป็นไปในทิศทางเดียวกันหรือไม่ ซึ่งจะใช้สถิติ Pearson Correlation Test ในการหาความสัมพันธ์เนื่องจากตัวแปลที่ต้องการวิเคราะห์นั้นเป็นตัวแปลชนิดอันตรภาค (Interval Variable) และตัวแปลชนิดอัตราส่วน (Ratio Variable) โดยผลลัพธ์ที่ได้นั้นจะเรียกว่า “สัมประสิทธิ์สหสัมพันธ์” หรือค่า r ซึ่งมีค่าอยู่ระหว่าง -1.00 ถึง 1.00 ซึ่งแสดงถึงความสัมพันธ์ระหว่างตัวแปลสองตัวโดยช่วงของสัมประสิทธิ์สหสัมพันธ์มีความหมายดังนี้

ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าติดลบ (r<0) แสดงว่าตัวแปล 2 ตัวมีความสัมพันธ์กันในทิศทางตรงกันข้าม

ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นบวก (r>0) แสดงว่าตัวแปล 2 ตัวมีความสัมพันธ์กันในทิศทางเดียวกัน

ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นศูนย์ (r=0) แสดงว่าตัวแปล 2 ตัวไม่มีความสัมพันธ์กัน

จากความหมายของสัมประสิทธิ์สหสัมพันธ์ที่กล่าวมาข้างต้น จะสามารถตั้งสมมุติฐานในการวิเคราะห์ผลการทดลองได้ดังนี้

H_0: ผลสรุปคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลไม่มีความสัมพันธ์กับความน่าจะเป็นของข้อความที่ได้จากการคำนวณโดยเครื่อง (r = 0)

H_1: ผลสรุปคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลมีความสัมพันธ์กับความน่าจะเป็นของข้อความที่ได้จากการคำนวณโดยเครื่องในทิศทางเดียวกัน (r>0)

ซึ่งหากตัวเลขระหว่างค่าความสัมพันธ์ของผลสรุปคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยบุคคลและคะแนนความน่าจะเป็นที่ได้จากการคำนวณโดยเครื่องนั้นเป็นไปในทิศทางเดียวกัน สมมุติฐาน H_0 จะถูกปฏิเสธ และยอมรับสมมุติฐาน h_1 ซึ่งสรุปได้ว่าตัวแปลสองตัวมีความสัมพันธ์ไปในทิศทางเดียวกัน หรือก็คือข้อความที่มีความน่าจะเป็นที่คำนวณได้โดยเครื่องต่ำ มักจะเป็นข้อความเชิงส่อเสียด

3.7 ประเด็นของความเชื่อถือได้ (Reliability) และความถูกต้อง (Validity) ของข้อมูล

เพื่อให้แน่ใจว่าข้อมูลในทุกขั้นตอนมีความน่าเชื่อถือและความถูกต้อง ผู้วิจัยจะพัฒนาโปรแกรมที่บันทึกผลลัพธ์ที่เกิดจากขั้นตอนทุกขั้นตอนโดยเริ่มตั้งแต่การรวบรวมข้อมูล การประมวลผลข้อมูลเบื้องต้นในแต่ละขั้นตอน และผลลัพธ์ที่ได้จากการจำแนกข้อความส่อเสียดออกจากข้อความปกติ เพื่อให้สามารถตรวจสอบย้อนกลับ และเพื่อยืนยันความถูกต้องของกระบวนการได้

บทที่ 4
ผลการทดลอง

ในบทนี้จะเป็นการนำเสนอผลการวิเคราะห์ข้อมูลซึ่งผู้วิจัยได้เก็บรวบรวมข้อมูลและทำการวิเคราะห์ตามกระบวนการต่าง ๆ ที่ได้นำเสนอในบทที่ 3

4.1 ผลการทดลอง

เริ่มแรกผู้วิจัยเก็บรวมรวมความคิดเห็นของผู้ใช้บริการอินเตอร์เน็ต X ในประเทศไทยเป็นจำนวน 3,971 ข้อความ ซึ่งตัวอย่างข้อความความคิดเห็นแสดงดังรูป 4.1

รูป 4.1 ตัวอย่างข้อความความคิดเห็น

เดะนี้เน็ทXเป็นไรอะ แย่ลงปะนิ
เน็ทXพอเปลี่ยนเป็นสายไฟเบอร์แล้วไม่เห็นแรงขึ้นเลย อะไรของหร่อน
คืออยากบอกว่าเน็ทXอาการสาหัสมากกกกกกกกกกก
อะไรนะเน็ทXล่มทั้งระบบ
เน็ทXย่านชานเมือง เน่ากันรีเปล่าครับ
เน็ทXกากมากกกกกกกก
รถหวอวิ่งเยอะจัง หรือไฟไหม้ชุมสายเน็ทX
เบื่อเน็ทXกระตุกอยู่ได้ ล่าไยมากก

จากนั้นผู้วิจัยตัดข้อความที่ซ้ำกันจากชุดข้อมูลข้างต้นทำให้เหลือข้อความทั้งสิ้นจำนวน 3,844 ข้อความ และทำการประมวลผลข้อมูลโดยเริ่มจากตัด URL แฮชแทก เครื่องหมาย @ ตัวอักษรหรือตัวเลขที่ซ้ำกันตั้งแต่ 3 ตัวขึ้นไป อักขระพิเศษต่าง ๆ เช่น เครื่องหมายอัศเจรีย์ หรือเครื่องหมายปรัศนี และคำสันธานที่มีอยู่ในข้อความออกโดยใช้ฐานข้อมูลคำศัพท์จาก LEXiTRON ดังแสดงด้วยกรอบแดงในรูป 4.2 และ 4.3 ตามลำดับ

รูป 4.2 ตัวอย่างข้อความความคิดเห็นก่อนประมวลผลข้อมูล

เดะนี้เน็ทXเป็นไรอะ แย่ลงปะนิ
เน็ทXพอเปลี่ยนเป็นสายไฟเบอร์แล้วไม่เห็นแรงขึ้นเลย อะไรของหร่อน
คืออยากบอกว่าเน็ทXอาการสาหัสมากกกกกกกกกก
อะไรนะเน็ทXล่มทั้งระบบ
เน็ทXย่านชานเมือง เน่ากันรีเปล่าครับ
เน็ทXกากมากกกกกกกก
รถหวอวิ่งเยอะจัง หรือไฟไหม้ชุมสายเน็ทX
เบื่อเน็ทXกระตุกอยู่ได้ ล่าไยมากก

รูป 4.3 ตัวอย่างข้อความความคิดเห็นหลังประมวลผลข้อมูล

เดะนี้เน็ทXเป็นไรอะ แย่ลงปะนิ
เน็ทXพอเปลี่ยนเป็นสายไฟเบอร์แล้วไม่เห็นแรงขึ้นเลย อะไรของพร่อน
คืออยากบอกว่าเน็ทXอาการสาหัสมาก
อะไรนะเน็ทXล่มทั้งระบบ
เน็ทXย่านชานเมือง เน่ากันรีเปล่าครับ
เน็ทXกากมาก
รถหวอวิ่งเยอะจัง หรือไฟไหม้ชุมสายเน็ทX
เบื่อเน็ทXกระตุกอยู่ได้ ล่าไยมาก

จากนั้นผู้วิจัยแบ่งการทดลองออกเป็นสองชุด เพื่อเปรียบเทียบการให้คะแนนความน่าจะเป็นของข้อความโดยบุคคลและโดยเครื่อง

ในการทดลองชุดแรก ผู้วิจัยกำหนดให้บุคคลที่มีประสบการณ์การใช้เครือข่ายสังคมออนไลน์ทวิตเตอร์จำนวน 5 คน ทำการตัดสินว่าข้อความที่ได้รวบรวมมาเป็นข้อความส่อเสียด ไม่เป็นข้อความส่อเสียด หรือไม่แน่ใจว่าเป็นข้อความส่อเสียด ซึ่งเกณฑ์การให้คะแนนในแต่ละข้อความแสดงในตารางที่ 4.1 โดยเริ่มให้คะแนนวันละ 100 ข้อความ เริ่มตั้งแต่วันที่ 10 มิถุนายน 2559 ถึงวันที่ 19 กรกฎาคม 2559 รวมทั้งสิ้น 40 วัน สรุปผลการให้คะแนนได้ดังตารางที่ 4.2 โดย N (Normal) ในตารางแสดงถึงจำนวนข้อความปกติ S (Sarcasm) แสดงถึงจำนวนข้อความส่อเสียด และ U (Uncertain) แสดงถึงจำนวนข้อความที่ไม่แน่ใจ ตามลำดับ

ตาราง 4.1 เงื่อนไขการให้คะแนนความน่าจะเป็นของข้อความโดยบุคคล

เงื่อนไข	**คะแนน**
ข้อความดังกล่าวเป็นข้อความส่อเสียด	0
ไม่แน่ใจว่าข้อความดังกล่าวเป็นข้อความส่อเสียดหรือไม่	0.5
ข้อความดังกล่าวไม่เป็นข้อความส่อเสียด	1

ตาราง 4.2 ผลสรุปการให้คะแนนความน่าจะเป็นของข้อความโดยบุคคลทั้ง 5 เป็นเวลา 40 วัน

วัน	**บุคคลที่ 1**			**บุคคลที่ 2**			**บุคคลที่ 3**			**บุคคลที่ 4**			**บุคคลที่ 5**		
	N	S	U	N	S	U	N	S	U	N	S	U	N	S	U
1	98	1	1	98	1	1	97	2	1	97	2	1	89	11	0
2	100	0	0	100	0	0	100	0	0	96	3	1	94	6	0
3	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
4	98	2	0	98	2	0	96	4	0	96	2	2	97	2	1
5	100	0	0	100	0	0	100	0	0	100	0	0	99	1	0
6	100	0	0	100	0	0	98	0	2	95	4	1	98	0	2

วัน	บุคคลที่ 1			บุคคลที่ 2			บุคคลที่ 3			บุคคลที่ 4			บุคคลที่ 5		
	N	S	U	N	S	U	N	S	U	N	S	U	N	S	U
7	100	0	0	100	0	0	100	0	0	99	1	0	99	1	0
8	100	0	0	100	0	0	100	0	0	98	2	0	99	1	0
9	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
10	100	0	0	100	0	0	100	0	0	95	4	1	97	3	0
11	100	0	0	100	0	0	100	0	0	99	1	0	100	0	0
12	99	0	1	99	0	1	98	1	1	93	3	4	100	0	0
13	99	1	0	99	1	0	100	0	0	99	0	1	99	1	0
14	100	0	0	100	0	0	99	1	0	93	5	2	96	2	2
15	98	2	0	98	2	0	99	1	0	98	2	0	99	1	0
16	100	0	0	100	0	0	100	0	0	99	1	0	100	0	0
17	99	1	0	99	1	0	99	0	1	95	4	1	96	0	4
18	99	1	0	99	1	0	100	0	0	99	1	0	99	0	1
19	100	0	0	100	0	0	100	0	0	100	0	0	99	1	0
20	99	1	0	99	1	0	100	0	0	99	1	0	100	0	0
21	100	0	0	100	0	0	100	0	0	99	1	0	100	0	0
22	100	0	0	100	0	0	100	0	0	99	1	0	99	1	0
23	99	0	1	99	0	1	100	0	0	98	1	1	99	0	1
24	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
25	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
26	99	1	0	99	1	0	99	1	0	98	1	1	99	1	0
27	100	0	0	100	0	0	100	0	0	99	1	0	100	0	0
28	97	3	0	97	3	0	96	3	1	96	2	2	99	0	1
29	94	6	0	94	6	0	94	6	0	95	5	0	98	2	0
30	96	2	2	96	2	2	95	2	3	96	3	1	97	1	2
31	99	1	0	99	1	0	99	1	0	99	1	0	100	0	0
32	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
33	98	2	0	98	2	0	99	1	0	99	1	0	96	4	0
34	99	1	0	99	1	0	99	1	0	99	1	0	99	1	0
35	98	1	1	98	1	1	98	1	1	98	1	1	95	2	3

วัน	บุคคลที่ 1			บุคคลที่ 2			บุคคลที่ 3			บุคคลที่ 4			บุคคลที่ 5		
	N	S	U	N	S	U	N	S	U	N	S	U	N	S	U
36	95	5	0	95	5	0	95	5	0	95	5	0	99	1	0
37	96	3	1	96	3	1	96	3	1	96	3	1	99	0	1
38	97	3	0	97	3	0	96	1	3	96	1	3	98	2	0
39	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
40	31	69	0	31	69	0	32	68	0	32	68	0	32	68	0
41	26	1	0	27	0	0	27	0	0	27	0	0	26	1	0
	3913	**107**	**7**	**3914**	**106**	**7**	**3911**	**102**	**14**	**3871**	**132**	**24**	**3895**	**114**	**18**

จากนั้นผู้วิจัยสรุปผลคะแนนความน่าจะเป็นของข้อความของแต่ละข้อความโดยนำคะแนนความน่าจะเป็นของข้อความของแต่ละบุคคลมาหาค่าเฉลี่ยซึ่งแสดงได้ดังตัวอย่างในตารางที่ 4.3

ตาราง 4.3 ตัวอย่างการให้คะแนนความน่าจะเป็นของข้อความโดยบุคคล

ข้อความ	บุคคลที่ 1	บุคคลที่ 2	บุคคลที่ 3	บุคคลที่ 4	บุคคลที่ 5	ค่าเฉลี่ย
เป็นคนเกลียดการรอครับ โดยเฉพาะไอ้วงกลมที่หมุนติ้วๆ บนด้านบนไอโฟนเนี่ย จะโหลดอีกนานแค่ไหน เน็ทXจะกากไปไหน โถ่	1	1	1	1	1	1
เน็ทXช้ามาก	1	1	1	1	1	1
เน็ทXเป็นอะไรเนี่ย	1	1	1	1	1	1
ไม่ทัน ขอบคุณเน็ทXค่ะ ที่แรงขนาดนี้	0	0	0	0	0	0
เน็ทXตอนกลางคืนนี่แบบบ	1	1	1	1	1	1
เน็ทXนี่อย่าใช้เด็ดขาด ผูกนูน่นี่แล้วแม่งระบบมันตอบสนองต่อความต้องการลูกค้าได้ช้ามาก	1	1	1	1	1	1
โอเคเราจะไปอาบน้ำละ เราจะไว้วางใจเน็ทX หึ	0.5	0.5	0	1	0	0.4
เน็ทXนี่ย้อนแย้งอ่ะ เฮ้ออ	1	1	1	1	1	1

ในการทดลองส่วนที่สองผู้วิจัยจะนำข้อความทวีตไปตัดคำโดยใช้เครื่องมือตัดคำ LexTo ซึ่งเป็นการตัดคำโดยใช้วิธีตัดคำที่ยาวที่สุด (Longest Matching) ซึ่งให้ผลลัพธ์ดังรูปที่ 4.4 โดยสัญลักษณ์ <s> แสดงถึงการเริ่มต้นข้อความ </s> แสดงถึงการจบข้อความ และ | แสดงถึงการแบ่งคำ

ในส่วนนี้ผู้วิจัยใช้วิธีการตัดคำที่ไม่สามารถหาความหมายได้ออกก่อนนำไปคำนวณความน่าจะเป็นของข้อความเนื่องจากตามธรรมชาติของภาษาแล้ว คำที่ไม่ปรากฏความหมายอยู่ในพจนานุกรมมักจะเป็นภาษาที่ไม่เป็นทางการ หรือเป็นคำที่เขียนผิด ผู้วิจัยเห็นว่าควรตัดคำที่ไม่มีความหมายออกในการคำนวณ

ความน่าจะเป็นของข้อความ เนื่องจากคำที่ไม่มีความหมายนั้นอาจส่งผลต่อการคำนวณความน่าจะเป็นของข้อความได้

รูป 4.4 ตัวอย่างข้อความหลังตัดคำด้วยโปรแกรม LexTo

ข้อความ
<s>|เป็น|คน|เกลียด|การ|รอ|ครับ|โดยเฉพาะ|ไอ้|วงกลม|ที่|หมุน|ติ้ว|ๆ|บน|ด้านบน|ไอโฟน|เนี่ย|จะ|โหลด|อีก|นาน|แค่|ไหน|เน็ทX|จะ|กาก|ไป|ไหน|โถ่|</s>
<s>|เน็ทX|ช้า|มาก|</s>
<s>|เน็ทX|เป็น|เชี่|ยอะ|ไร|เนี่ย|</s>
<s>|ไม่ทัน|ขอบคุณ|เน็ทX|ค่ะ|ที่|แรง|ขนาด|นี้|</s>
<s>|เน็ทX|ตอนกลางคืน|นี่|แบบ|บ|</s>
<s>|เน็ทX|นี่|อย่า|ใช้|เด็ดขาด|ผูก|นูน่|นี่|แล้ว|แม่ง|ระบบ|
มัน|ตอบสนอง|ต่อ|ความต้องการ|ลูกค้า|ได้|ช้า|มาก|</s>
<s>|โอเค|เรา|จะ|ไป|อาบน้ำ|ละ|เรา|จะ|ไว้วางใจ|เน็ทX|หื|</s>
<s>|เน็ทX|นี่|ย้อน|แย้ง|อ่ะ|เฮ้อ|อ|</s>

จากนั้นผู้วิจัยใช้เทคนิคคำนวณความน่าจะเป็นของข้อความซึ่งได้กล่าวไว้ในบทที่ 2 และ 3 เพื่อหาความน่าจะเป็นของข้อความแต่ละข้อความ โดยค่าความน่าจะเป็นที่คำนวณได้มีค่าอยู่ระหว่าง 1.18×10^{-83} และ 1.16×10^{-2} แต่เนื่องจากค่าความน่าจะเป็นของข้อความมีค่าเข้าใกล้ 0 ดังนั้นผู้วิจัยจึง Take Log เพื่อลดความไม่เสถียรทางตัวเลข (Numerical Instability) โดยค่าความน่าจะเป็นของข้อความหลังใช้ฟังก์ชันลอการิทึมมีค่าอยู่ระหว่าง -82.93 และ -1.93 ตามลำดับ โดยแสดงตัวอย่างผลลัพธ์ได้ดังตาราง 4.4

ตาราง 4.4 ตัวอย่างค่าความน่าจะเป็นของแต่ละข้อความที่ได้จากการคำนวณโดยเครื่อง

ข้อความ	Log Probability
<s>|เป็น|คน|เกลียด|การ|รอ|ครับ|โดยเฉพาะ|ไอ้|วงกลม|ที่|หมุน|ติ้ว|ๆ|บน|ด้านบน|ไอโฟน|เนี่ย|จะ|โหลด|อีก|นาน|แค่|ไหน|เน็ทX|จะ|กาก|ไป|ไหน|โถ่|</s>	-48.82772601
<s>|เน็ทX|ช้า|มาก|</s>	-4.240073128
<s>|เน็ทX|เป็น|เชี่|ยอะ|ไร|เนี่ย|</s>	-6.529818452
<s>|ไม่ทัน|ขอบคุณ|เน็ทX|ค่ะ|ที่|แรง|ขนาด|นี้|</s>	-16.97560548
<s>|เน็ทX|ตอนกลางคืน|นี่|แบบ|บ|</s>	-9.460385316
<s>|เน็ทX|นี่|อย่า|ใช้|เด็ดขาด|ผูก|นูน่|นี่|แล้ว|แม่ง|ระบบ| มัน|ตอบสนอง|ต่อ|ความต้องการ|ลูกค้า|ได้|ช้า|มาก|</s>	-31.02021222
<s>|โอเค|เรา|จะ|ไป|อาบน้ำ|ละ|เรา|จะ|ไว้วางใจ|เน็ทX|หื|</s>	-20.86553967
<s>|เน็ทX|นี่|ย้อน|แย้ง|อ่ะ|เฮ้อ|อ|</s>	-10.17140177

เพื่อตรวจสอบความสัมพันธ์เบื้องต้นระหว่างความน่าจะเป็นที่คำนวณได้โดยเครื่อง และการประเมินคะแนนความน่าจะเป็นโดยบุคคล ผู้วิจัยจึงนำข้อความ รวมถึงค่าความน่าจะเป็นที่คำนวณได้โดย

เครื่อง และผลสรุปคะแนนการประเมินความน่าจะเป็นโดยบุคคลมาเปรียบเทียบกัน ซึ่งพบว่า คะแนนความน่าจะเป็นโดยบุคคลนั้นมีความสัมพันธ์เป็นไปในทิศทางเดียวกันกับค่าความน่าจะเป็นที่คำนวณได้โดยเครื่อง กล่าวคือเมื่อผลสรุปคะแนนค่าความน่าจะเป็นโดยบุคคลมีค่าน้อย ค่าความน่าจะเป็นที่คำนวณได้โดยเครื่องก็จะมีค่าน้อยเป็นไปในทิศทางเดียวกัน ดังตัวอย่างในตารางที่ 4.5

ตาราง 4.5 ตัวอย่างผลสรุปคะแนนความน่าจะเป็นโดยบุคคล และค่าความน่าจะเป็นที่คำนวณได้โดยเครื่อง

ข้อความ	**ผลสรุปคะแนนความน่าจะเป็นโดยบุคคล**	**ค่าความน่าจะเป็นที่คำนวณได้โดยเครื่อง**
เปิด Youtube ด้วยความละเอียด 360k กระตุกแล้ว กระตุกอีก ขอบคุณX อินเตอร์เน็ตไฮสปีดประทับใจไม่เคยลืม	0.0	-35.33031184
เน็ตXนี่เป็นอะไรที่เร็วมากเลยนะ แบบดูซีรีส์ 10 วิกตุ 10 วิกตุตลอด ไม่เคยเจอเน็ตอะไรเร็วเท่านี้มาก่อนเลย	0.0	-45.92939325
เน็ตXแรงมากค่ะ ขนาดเว็บXเองยังเข้าไม่ได้ แหกโค้งไปเลยค่ะ	0.6	-26.46861125
Xอินเตอร์เน็ต เจ้าแห่งความห่วยแตก ครองแชมป์ทุกวันทุกเดือนทุกปีทุกสมัย	1.0	-15.35249299

4.2 การทดสอบผลการทดลองทางสถิติ

ในการวิเคราะห์ผลการทดลอง ผู้วิจัยใช้สถิติ Pearson Correlation Test เพื่อหาความสัมพันธ์ระหว่างผลสรุปของความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยบุคคลและที่คำนวณได้โดยเครื่องว่าเป็นไปในทิศทางเดียวกันหรือไม่ โดยผู้วิจัยตั้งสมมุติฐานของการทดลองนี้ได้ดังนี้

H0: ผลสรุปความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยบุคคลไม่มีความสัมพันธ์หรือแปรผกผันกับความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง ($r<=0$)

H1: ผลสรุปความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยบุคคลมีความสัมพันธ์เป็นไปในทิศทางเดียวกันกับความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง ($r>0$)

ผู้วิจัยใช้โปรแกรม SPSS ในการทดสอบสมมุติฐานโดยกำหนดผลสรุปความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยบุคคลและความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่องเป็นตัวแปรที่ใช้ในการคำนวณหาความสัมพันธ์ และกำหนดช่วงความเชื่อมั่นที่ร้อยละ 95

ซึ่งผลลัพธ์ของการใช้สถิติ Pearson Correlation Test เพื่อคำนวณหาความสัมพันธ์ระหว่างผลสรุปค่าความน่าจะเป็นของข้อความที่ประเมินโดยบุคคล และความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง โดยผลการทดสอบแสดงได้ดังรูปที่ 4.5

รูป 4.5 ผลลัพธ์การใช้สถิติ Pearson Correlation Test

Correlations

Correlations

		Prob	Rate
Prob	Pearson Correlation	1	.035*
	Sig. (1-tailed)		.015
	N	3844	3844
Rate	Pearson Correlation	.035*	1
	Sig. (1-tailed)	.015	
	N	3844	3844

*. Correlation is significant at the 0.05 level (1-tailed).

โดยจากผลการทดลอง ค่า Sig. (1-tailed) หรือ p-value (เนื่องจากเป็นการทดสอบสมมุติฐานแบบทางเดียว (1-tailed test)) มีค่าเท่ากับ 0.015 ซึ่งมีค่าน้อยกว่า 0.05 จึงสามารถสรุปผลได้ว่า จากหลักฐานข้อมูลที่ได้รวบรวมมา มีความเพียงพอที่จะปฏิเสธ H0 ได้ หรือก็คือ ผลสรุปความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยบุคคลมีความสัมพันธ์เป็นไปในทิศทางเดียวกันกับความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง

โดยหลังจากสรุปผลการทดสอบความสัมพันธ์ระหว่างผลสรุปความน่าจะเป็นของข้อความที่ประเมินโดยบุคคลและความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง หากสังเกตที่ค่าสัมประสิทธิ์สหสัมพันธ์หรือค่า r ซึ่งแสดงถึงระดับความสัมพันธ์ โดยจะเห็นว่าจากค่าสัมประสิทธิ์สหสัมพันธ์ ในรูป 4.5 ซึ่งมีค่าคือ 0.035 ซึ่งค่อนข้างน้อย อย่างไรก็ตาม เนื่องจากข้อมูลข้างต้นไม่ใช่ข้อมูลเชิงปริมาณแท้จริง ซึ่งอาจส่งผลให้ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าน้อย จึงอาจสรุปได้ว่าคะแนนความน่าจะเป็นของข้อความซึ่งประเมินโดยบุคคลและความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่องมีความสัมพันธ์กันในทิศทางเดียวกัน แต่มีความสัมพันธ์ที่ไม่ชัดเจนเท่าใดนัก

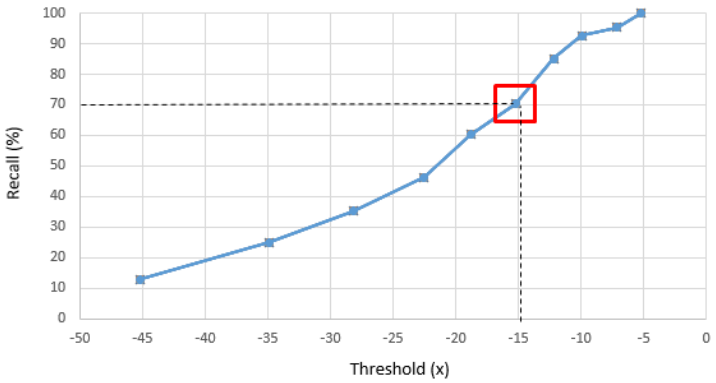
4.3 การทดสอบประสิทธิภาพของเทคนิค

4.3.1 การกำหนดเส้นแบ่งค่าความน่าจะเป็น

ผู้วิจัยได้จัดเรียงลำดับข้อความตามความน่าจะเป็นที่คำนวณได้เพื่อหาค่าเส้นแบ่งความน่าจะเป็นของข้อความ (threshold) เพื่อใช้จำแนกข้อความส่อเสียดออกจากข้อความปกติ โดยกราฟที่ 4.1 แสดงค่าความระลึกที่เส้นแบ่งค่าความน่าจะเป็นของข้อความแต่ละช่วงโดยหากความน่าจะเป็นของข้อความมีค่าน้อยกว่าค่าเส้นแบ่ง ข้อความนั้นจะถูกจำแนกเป็นข้อความส่อเสียด ผู้วิจัยพบว่าหากกำหนดเส้นแบ่งค่าความ

น่าจะเป็นไว้ที่ -14.87 จะสามารถครอบคลุมข้อความส่อเสียดในชุดข้อมูล[13] ได้ร้อยละ 70.37 ดังแสดงในตารางที่ 4.6

กราฟที่ 4.1 ความสัมพันธ์ระหว่างค่าความระลึกและเส้นแบ่งในแต่ละช่วง



ตาราง 4.6 ผลสรุปจำนวนข้อความส่อเสียดในแต่ละเปอร์เซ็นไทล์

เปอร์เซ็นไทล์	จำนวนข้อความส่อเสียด	จำนวนข้อความส่อเสียดสะสม	เปอร์เซ็นต์สะสม (%)	ขอบบน (ความน่าจะเป็น)	ขอบล่าง (ความน่าจะเป็น)
10	14	14	12.96	-74.709	-42.970
20	13	27	25.00	-42.483	-34.204
30	12	39	36.11	-32.922	-26.469
40	10	49	45.37	-26.330	-21.804
50	11	60	55.56	-21.047	-18.142
60	16	76	70.37	-17.343	-14.872
70	15	91	84.26	-14.273	-11.431
80	8	99	91.67	-10.991	-9.870
90	5	104	96.30	-8.654	-6.569
100	4	108	100.00	-6.041	-5.231

[13] ข้อความส่อเสียดจำแนกโดยผู้วิจัยโดยใช้เกณฑ์การจำแนกโดยดูจากแฮชแทก #ประชด และวิจารณญาณของผู้วิจัย เป็นหลัก

4.3.2 การทดสอบประสิทธิภาพของเทคนิคกับข้อมูลทดสอบ

เพื่อศึกษาถึงประสิทธิภาพของเทคนิคการจำแนกข้อความส่อเสียดโดยใช้ความน่าจะเป็นของประโยค ผู้วิจัยได้รวบรวมข้อมูลเพิ่มเติมจากทวิตเตอร์โดยใช้คีย์เวิร์ดชุดเดียวกับที่ใช้ในการเก็บรวบรวมข้อมูลในครั้งแรก โดยรวบรวมข้อมูลตั้งแต่วันที่ 15 เมษายน 2560 ถึงวันที่ 16 เมษายน 2560 จำนวน 59 ข้อความเพื่อใช้ในการทดสอบเทคนิคที่นำเสนอและค่าเส้นแบ่ง โดยข้อความดังกล่าวประกอบด้วยข้อความปกติจำนวน 52 ข้อความ และข้อความเชิงส่อเสียดจำนวน 7 ข้อความ[14] และคำนวณความน่าจะเป็นของชุดข้อความดังกล่าว พบว่า ความน่าจะเป็นของข้อความในรูปแบบลอการิทึมของข้อความส่อเสียดในชุดข้อมูลดังกล่าวอยู่ในช่วง -24.39 ถึง -6.98 และหากนำเส้นแบ่งที่ -14.87 จะครอบคลุมชุดข้อมูลส่อเสียดได้ทั้งสิ้น 5 จาก 7 ข้อความ หรือครอบคลุมข้อความส่อเสียดได้ร้อยละ 71.43

[14] ข้อความส่อเสียดจำแนกโดยผู้วิจัยโดยใช้เกณฑ์การจำแนกโดยดูจากแฮชแทก #ประชด และวิจารณญาณของผู้วิจัย เป็นหลัก

บทที่ 5
สรุปผลการศึกษา และข้อเสนอแนะ

การวิจัยครั้งนี้มีจุดประสงค์เพื่อศึกษาและพัฒนาตัวแบบสำหรับจำแนกข้อความส่อเสียดออกจากข้อความปกติในภาษาไทยโดยใช้ข้อมูลความคิดเห็นของผู้บริโภคเกี่ยวกับการบริการอินเตอร์เน็ตในประเทศไทย

ข้อมูลที่ใช้ในการวิจัยได้รวบรวมจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ระหว่างวันที่ 25 มกราคม 2553 ถึง 9 มิถุนายน 2559 โดยเก็บรวบรวมข้อมูลทั้งสิ้น 4,027 ข้อความ ซึ่งประกอบไปด้วยข้อความปกติจำนวน 3,913 ข้อความ และข้อความที่เป็นข้อความเชิงส่อเสียดทั้งสิ้น 108 ข้อความ จากนั้นผู้วิจัยได้ตัดข้อความที่มีความซ้ำซ้อนและประมวลผลข้อมูลโดยเริ่มจากการตัด URL แฮชแท็ก เครื่องหมาย @ ตัวอักษรหรือตัวเลขที่ปรากฏติดกันมากกว่า 3 ตัวขึ้นไป รวมถึงอักขระพิเศษต่าง ๆ ตามลำดับ ซึ่งส่งผลให้ชุดข้อมูลในการทดลองเหลืออยู่ทั้งสิ้น 3,844 ข้อความ

การทดลองแบ่งออกเป็นสองชุด โดยเปรียบเทียบการให้คะแนนความน่าจะเป็นของข้อความโดยมนุษย์และคะแนนความน่าจะเป็นที่คำนวณได้จากเครื่อง โดยในส่วนของการให้คะแนนความน่าจะเป็นของข้อความโดยมนุษย์นั้น ข้อความแต่ละข้อความจะถูกให้คะแนนโดยบุคคลที่มีประสบการณ์การใช้เครือข่ายสังคมออนไลน์จำนวน 5 คน ว่าข้อความแต่ละข้อความเป็นข้อความส่อเสียดหรือไม่ โดยถ้าผู้ให้คะแนนเห็นว่าข้อความดังกล่าวเป็นข้อความส่อเสียด ผู้ให้คะแนนจะให้คะแนนข้อความนั้นเป็น 0 คะแนน ถ้าผู้ให้คะแนนเห็นว่าข้อความนั้นเป็นข้อความปกติ ผู้ให้คะแนนจะให้คะแนนข้อความนั้นเป็น 1 คะแนน และถ้าผู้ให้คะแนนไม่แน่ใจว่าข้อความนั้นเป็นข้อความส่อเสียดหรือข้อความปกติ ผู้ให้คะแนนจะให้คะแนนข้อความนั้นเป็น 0.5 คะแนน หลังจากนั้นจึงนำคะแนนของบุคคลทั้ง 5 คนมาหาค่าเฉลี่ยเป็นคะแนนความน่าจะเป็นของข้อความ

ในส่วนของการทดลองโดยเครื่อง ข้อความแต่ละข้อความจะถูกตัดคำโดยใช้เครื่องมือ Lexto ซึ่งเป็นเครื่องมือการตัดคำตามคำที่ยาวที่สุดที่ปรากฏอยู่ในพจนานุกรม ซึ่งในขั้นตอนนี้ คำที่ไม่รู้จักหรือไม่มีความหมายจะถูกตัดออกจากข้อความ จากนั้นจึงแปลงชุดข้อความที่ถูกตัดคำแล้วให้อยู่ในรูปแบบ bi-gram เพื่อคำนวณหาคะแนนความน่าจะเป็นของข้อความด้วยเทคนิคที่นำเสนอและนำคะแนนความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่องมาเปรียบเทียบกับคะแนนความน่าจะเป็นของข้อความที่ได้จากการให้คะแนนโดยมนุษย์ เพื่อศึกษาถึงประสิทธิภาพของเทคนิคที่ใช้ว่ามีความสามารถในการจำแนกข้อความส่อเสียดออกจากข้อความปกติได้ดีมากน้อยเพียงใดหากเทียบกับการจำแนกข้อความส่อเสียดโดยมนุษย์

5.1 สรุปผลการศึกษา

จากผลการทดสอบผลการทดลองทางสถิติโดยใช้การทดสอบสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's Correlation Test) พบว่า คะแนนความน่าจะเป็นของข้อความที่ได้จากการประเมินโดยมนุษย์มีความสัมพันธ์เป็นไปในทิศทางเดียวกันกับคะแนนความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่องที่ช่วงความเชื่อมั่น 95% และหากพิจารณาที่ค่า r จะพบว่าค่า r ของผลการทดสอบคือ 0.035 อย่างไรก็ตาม สืบเนื่องจากข้อมูลข้างต้นไม่ใช่ข้อมูลเชิงปริมาณที่แท้จริง ส่งผลให้ค่า r ค่อนข้างน้อย แต่ก็สามารถสรุปได้ว่าคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยมนุษย์และคะแนนความน่าจะเป็นที่ได้จากการคำนวณโดยเครื่องมีความสัมพันธ์เป็นไปในทิศทางเดียวกัน แต่มีความสัมพันธ์ที่ไม่ชัดเจนเท่าใดนัก

เพื่อทดสอบถึงประสิทธิภาพในการจำแนกข้อความส่อเสียดจากเทคนิคที่ใช้ ผู้วิจัยได้จัดเรียงลำดับข้อความตามคะแนนความน่าจะเป็นที่คำนวณได้โดยเครื่องเพื่อหาค่าเส้นแบ่งความน่าจะเป็นของข้อความที่เหมาะสม (threshold) เพื่อใช้จำแนกข้อความส่อเสียดออกจากข้อความปกติ และพบว่าหากกำหนดค่าเส้นแบ่งไว้ที่ -14.872 จะสามารถครอบคลุมจำนวนข้อความส่อเสียดในชุดข้อมูลได้ร้อยละ 70.37

นอกจากนี้ผู้วิจัยได้ทำการศึกษาเพิ่มเติมเพื่อศึกษาถึงประสิทธิภาพของเทคนิคการจำแนกข้อความส่อเสียดที่นำเสนอ โดยผู้วิจัยได้รวบรวมข้อความทวีตจากทวิตเตอร์โดยใช้คำค้นหา (Keyword) ชุดเดียวกันกับที่ใช้ในเก็บรวบรวมข้อความสำหรับการศึกษาในครั้งแรกและได้ข้อความใหม่จำนวน 59 ข้อความเพื่อใช้ในการทดสอบเทคนิคที่นำเสนอและค่าเส้นแบ่ง โดยข้อความดังกล่าวประกอบด้วยข้อความปกติจำนวน 52 ข้อความ และข้อความที่มีความหมายเชิงส่อเสียดจำนวน 7 ข้อความ เมื่อใช้เทคนิคที่นำเสนอในการจำแนกข้อความชุดนี้โดยกำหนดเส้นแบ่งไว้ที่ -14.872 พบว่าสามารถครอบคลุมข้อความส่อเสียดได้ถึง 5 ข้อความจากข้อความส่อเสียดทั้งหมด 7 ข้อความ หรือครอบคลุมได้ร้อยละ 71.43

จากผลการศึกษาเพื่อทดสอบประสิทธิภาพของเทคนิคที่ได้พัฒนาขึ้นสามารถสรุปได้ว่า เทคนิคที่ใช้มีความสามารถในการจำแนกข้อความส่อเสียดออกจากข้อความปกติได้ดีอยู่ในระดับหนึ่งเมื่อเทียบกับการจำแนกข้อความส่อเสียดโดยมนุษย์ โดยการศึกษาครั้งนี้ได้ทำให้ผู้วิจัยสามารถสร้างตัวแบบสำหรับการจำแนกข้อความส่อเสียดออกจากข้อความปกติโดยใช้ความน่าจะเป็นของข้อความ และทราบถึงปัจจัยต่าง ๆ ที่ส่งผลต่อการจำแนกข้อความส่อเสียด ซึ่งมีความสอดคล้องกับวัตถุประสงค์การวิจัยที่ได้กล่าวไว้ในบทที่ 1

5.2 ข้อจำกัดของการศึกษาและข้อเสนอแนะ

ผลการศึกษาวิจัยเรื่องการจำแนกข้อความส่อเสียดออกจากข้อความปกติโดยใช้ความน่าจะเป็นของทวีต ทำให้ทราบถึงเทคนิคการนำความน่าจะเป็นของข้อความมาใช้ในการจำแนกข้อความส่อเสียดออกจากข้อความปกติ ซึ่งเป็นประโยชน์ในการวิเคราะห์ความคิดเห็น โดยทำให้การวิเคราะห์ความคิดเห็นเป็นไปได้อย่างถูกต้องและแม่นยำมากยิ่งขึ้น อย่างไรก็ตาม ในการศึกษาครั้งนี้ยังมีข้อจำกัดคือ

1) ในการศึกษาครั้งนี้เป็นการศึกษากับข้อมูลความคิดเห็นต่อผู้ให้บริการโทรคมนาคมเพียงรายเดียว ซึ่งผู้วิจัยเห็นว่า หากมีการศึกษาเทคนิคนี้กับข้อมูลชุดอื่น ๆ เพิ่มเติม จะทำให้ทราบถึงปัจจัยที่มีผลต่อประสิทธิภาพของเทคนิคนี้มากยิ่งขึ้น

2) ค่าสัมประสิทธิ์สหสัมพันธ์ของการทดสอบทางสถิติที่ผู้วิจัยใช้ในการทดสอบมีค่าน้อย อาจเนื่องจากคะแนนความน่าจะเป็นที่ได้จากการประเมินโดยมนุษย์ไม่ใช่ข้อมูลเชิงปริมาณที่แท้จริง เพราะผู้วิจัยใช้จำนวนผู้ประเมินเพียง 5 คน ทำให้ขอบเขตของคะแนนที่ได้จากการประเมินโดยบุคคลมีค่าเป็นไปได้ทั้งหมดเพียง 11 ค่า ซึ่งก็คือ 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 เท่านั้น โดยในการศึกษาต่อเนื่อง ผู้วิจัยเห็นว่าหากเพิ่มจำนวนผู้ทำการประเมินคะแนนความน่าจะเป็นให้มากยิ่งขึ้น จะทำให้มาตราส่วนคะแนนเฉลี่ยให้มีความละเอียดมากขึ้น ซึ่งอาจส่งผลให้ผลสรุปที่ได้จากการศึกษามีความชัดเจนมากยิ่งขึ้น

3) จากผลการรวบรวมข้อมูลและการวิเคราะห์ผลการทดลอง ผู้วิจัยสังเกตว่า ความยาวของข้อความมีผลต่อความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่อง หรืออีกนัยหนึ่งคือ ยิ่งข้อความมีความยาวมากยิ่งขึ้น ความน่าจะเป็นที่คำนวณได้โดยเครื่องจะมีค่าน้อยกว่าข้อความที่สั้นกว่า โดยผู้วิจัยประสงค์จะศึกษาต่อเนื่องโดยนำชุดข้อมูลที่มีความยาวของข้อความใกล้เคียงกันมาทำการทดลองจำแนกข้อความส่อเสียดออกจากข้อความปกติโดยใช้ความน่าจะเป็นที่คำนวณได้โดยเครื่องและศึกษาว่าจะมีผลแตกต่างกับผลที่ได้จากการศึกษาครั้งนี้มากน้อยเพียงใด

4) เนื่องจากงานวิจัยชิ้นนี้ใช้ข้อมูลในรูปแบบ bigram ในการแบ่งข้อความออกเป็นหน่วยย่อยที่อยู่ติดกัน ซึ่งการใช้รูปแบบข้อมูลแบบ bigram ค่อนข้างมีข้อจำกัดในการคำนวณความน่าจะเป็นของข้อความ กล่าวคือ การคำนวณความน่าจะเป็นในชุดข้อมูลรูปแบบนี้จะคำนวณความน่าจะเป็นเฉพาะคำที่เกิดติดกันเท่านั้นโดยไม่สามารถข้ามคำได้ เช่น ข้อความ “ฉันนั่งตกปลา” หากใช้รูปแบบข้อมูลในรูปแบบ bigram จะสามารถแบ่งข้อความออกเป็นหน่วยย่อยได้ดังนี้ *{“<s>, ฉัน”, “ฉัน, นั่ง” , “นั่ง, ตกปลา” , “ตกปลา, </s>”}* โดยผู้วิจัยมีความเห็นว่า หากเปลี่ยนรูปแบบข้อมูลในการทดลองจาก bigram เป็น k-skip-2-gram อาจจะส่งผลให้ความน่าจะเป็นของข้อความที่คำนวณได้โดยเครื่องมีความยืดหยุ่นมากยิ่งขึ้น เนื่องจากการคำนวณความน่าจะเป็นในชุดข้อมูลรูปแบบ k-skip-2-gram จะอนุญาตให้ลำดับของคำที่สนใจในข้อความไม่จำเป็นต้องเกิดติดกันเสมอไป โดยอาจปรากฏอยู่ใกล้เคียงกัน ซึ่งอาจมีความเหมาะสมกับธรรมชาติของภาษาไทยมากกว่า

5) เนื่องจากภายในวันที่ 40 มีการประเมินข้อความเป็นข้อความส่อเสียดค่อนข้างมากเนื่องจากผู้วิจัยไม่ได้สลับที่ (shuffle) ข้อความซึ่งใช้แฮชแทก #ประชด ในการค้นหา ซึ่งส่งผลให้ข้อความซึ่งประกอบด้วยแฮชแทกดังกล่าว อาจปรากฏอยู่ภายในวันที่ 40 เป็นจำนวนมาก และอาจส่งผลให้เกิดความลำเอียง หรืออคติ (bias) ต่อการประเมินคะแนนความน่าจะเป็นโดยบุคคล โดยในอนาคต ผู้วิจัยประสงค์จะสลับที่ข้อความเพื่อเป็นการลดอคติที่อาจเกิดขึ้นในระหว่างการประเมิน

รายการอ้างอิง

วิรัช ศรเลิศล้ำวาณิช. 2536. "การตัดคำในระบบแปลภาษา (Word Segmentation for Thai in Machine)". ใน National Electronics and Computer Technology. NECTEC.

Apisuwankun, P., & Mongkolnavin, J. (2013). Opinion Strength Identification in Customer Review Summarizing System Using Association Rule Technique Paper presented at the The International Conference on E-Technologies and Business on the Web (EBW 2013).

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Paper presented at the LREC.

Bamman, D., & Smith, N. A. (2015). Contextualized Sarcasm Detection on Twitter. Paper presented at the Proceedings of the International AAAI Conference on Weblogs and Social Media, Oxford, UK.

Bo P., Lillian L., and Shivakumar V. 2002. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02), Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 79-86.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

Chamlertwat, W., Bhattarakosol, P., and Rungkasiri, T. (2011). Innovative marketing tool by applying opinion mining on the micro-blog. Chulalongkorn University.

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., and Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis. Journal of Universal Computer Science, 18(8), 973-992.

Charoenpornsawat, P. (1999). Feature-based Thai Word Segmentation. (Master Degree), Chulalongkorn University.

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. International Journal of Technology Enchancements and Emerging Engineering Research, 1(4), 131.

Ding, X., Liu, B., & Zhang, L. (2009). Entity discovery and assignment for opinion mining applications. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France.

Do BH., Wu AS., Maley J. and Biswal S., "Automatic Retrieval of Bone Fracture Knowledge Using Natural Language Processing," J Digit Imaging, vol. 26, no. 4, pp. 709–713, Oct. 2012.

El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, 40(4), 501-513.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82.

Fundin A. and Elg M., "Continuous learning using dissatisfaction feedback in new product development contexts," Int J Qual & Reliability Mgmt, vol. 27, no. 8, pp. 860–877, Sep. 2010.

Gonzlez-Ibez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: a closer look. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, Portland, Oregon.

Haruechaiyasak, C. (2010). Basic NLP Tools and Text Mining Applications

Haruechaiyasak, C., Kongthon, A., Palingoon, P., & Trakultaweekoon, K. (2013). S-Sense: A Sentiment Analysis Framework for Social Media Sensing.

Hsinchun, C., & Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems, 26(3), 1-34.

Hsu, V., & Jain, S. (2015). The Lowest Form of Wit: Identifying Sarcasm in Social Media the 14th Annual CS 229 Machine Learning. Arrillaga Center for Sports and Recreation (ACSR), 341 Galvez St, Stanford, CA 94305 Standford University.

Jindal, N., & Liu, B. (2006). Identifying comparative sentences in text documents. Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA.

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition: Prentice Hall PTR.

Jurafsky, D., & Martin, J. H. (2014). Speech and Language Processing: Prentice Hall.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527-541.

Osherenko, A. (2010). Opinion mining and lexical affect sensing. Augsburg University, Diss.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Brew, C. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. Biomedical Informatics Insights, 5(Suppl 1), 3-16.

Pt´cek, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. Information Processing & Management, 50(5), 693-707.

Pt´cek, T., Habernal, I., and Hong, J. (2014). Sarcasm Detection on Czech and English Twitter. Paper presented at the Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.

Rajadesingan, A., Zafarani, R., and Liu, H. (2015). Sarcasm Detection on Twitter: A Behavioral Modeling Approach. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15). ACM, New York, NY, USA, 97-106.

Rodrigo, M., João, F. V., Wilson, P. G. N., Document-level sentiment classification: An empirical comparison between SVM and ANN, Expert Systems with Applications, Volume 40, Issue 2, 2013, Pages 621-633, ISSN 0957-4174

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2015). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. Int J Med Inform.

Sibarani, E. M., Nadial, M., Panggabean, E., & Meryana, S. (2013). A Study of Parsing Process on Natural Language Processing in Bahasa Indonesia. Paper presented at the Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering.

Soo-Min K., and Eduard H. 2004. Determining the sentiment of opinions. In Proceedings of COLING-04, 1267-1373.

Suh, B., Lichan, H., Pirolli, P., & Chi, E. H. (2010, 20-22 Aug. 2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. Paper presented at the Social Computing (SocialCom), 2010 IEEE Second International Conference.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Theresa W., Janyce W., and Paul H.. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35(3):99-433.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter:

Tungthamthiti, P., Kiyoaki, S., and Masnizah, M., 2014, Recognition of Sarcasm in Tweets Based on Concept Level Sentiment Analysis and Supervised Learning Approaches, In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, 404–413, Phuket, Thailand

Wang, H., Can, D., Kazemzadeh, A., Fran, Bar, O., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. Paper presented at the Proceedings of the ACL 2012 System Demonstrations, Jeju Island, Korea.

Wei, J., Hay, H. H., & Rohini, K. S. (2009). OpinionMiner: a novel machine learning system for web opinion mining and extraction. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA.

What 140 Characters Reveal about Political Sentiment. Paper presented at the Fourth International AAAI Conference.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Systems with Applications, 39(11), 10283-10291.

รายการอ้างอิง

ภาคผนวก ก.

ตาราง ก.1 แสดงจำนวนข้อความส่อเสียดเกี่ยวกับผู้ให้บริการอินเตอร์เน็ท X ในแต่ละปี โดยค้นหาจาก Hashtag #ประชด

ปี	จำนวนข้อความส่อเสียด
2552	1
2553	0
2554	7
2555	26
2556	21
2557	7
2558	5
2559	1
2560	1
รวม	**69**

ประวัติผู้เขียนวิทยานิพนธ์

ข้าพเจ้าชื่อนายกษิดิ์เดช ทาแป้ง เกิดเมื่อวันที่ 24 เมษายน พ.ศ. 2535 ปัจจุบันอายุ 25 ปี ปัจจุบันทำงานอยู่ที่การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย สำนักงานใหญ่บางกรวย มีหน้าที่ในการวิเคราะห์ ออกแบบ และพัฒนาระบบสารสนเทศภายในฝ่ายสื่อสารองค์การ

ปัจจุบัน จบการศึกษาระดับปริญญาตรีวิทยาศาสตรบัณฑิตสาขาวิทยาการคอมพิวเตอร์จากมหาวิทยาลัยอัสสัมชัญ และปริญญาตรีศิลปศาสตร์บัณฑิตเอกภาษาอังกฤษจากมหาวิทยาลัยรามคำแหง

ในอดีต เคยรับหน้าที่ในตำแหน่งนักพัฒนาด้านความปลอดภัย บริษัทไซเบอร์เกม คอร์ปอเรชั่น จำกัด และนักพัฒนาในส่วนของการชำระเงินด้วยบัตรเครดิตของบริษัททูซีทูพี (ประเทศไทย)