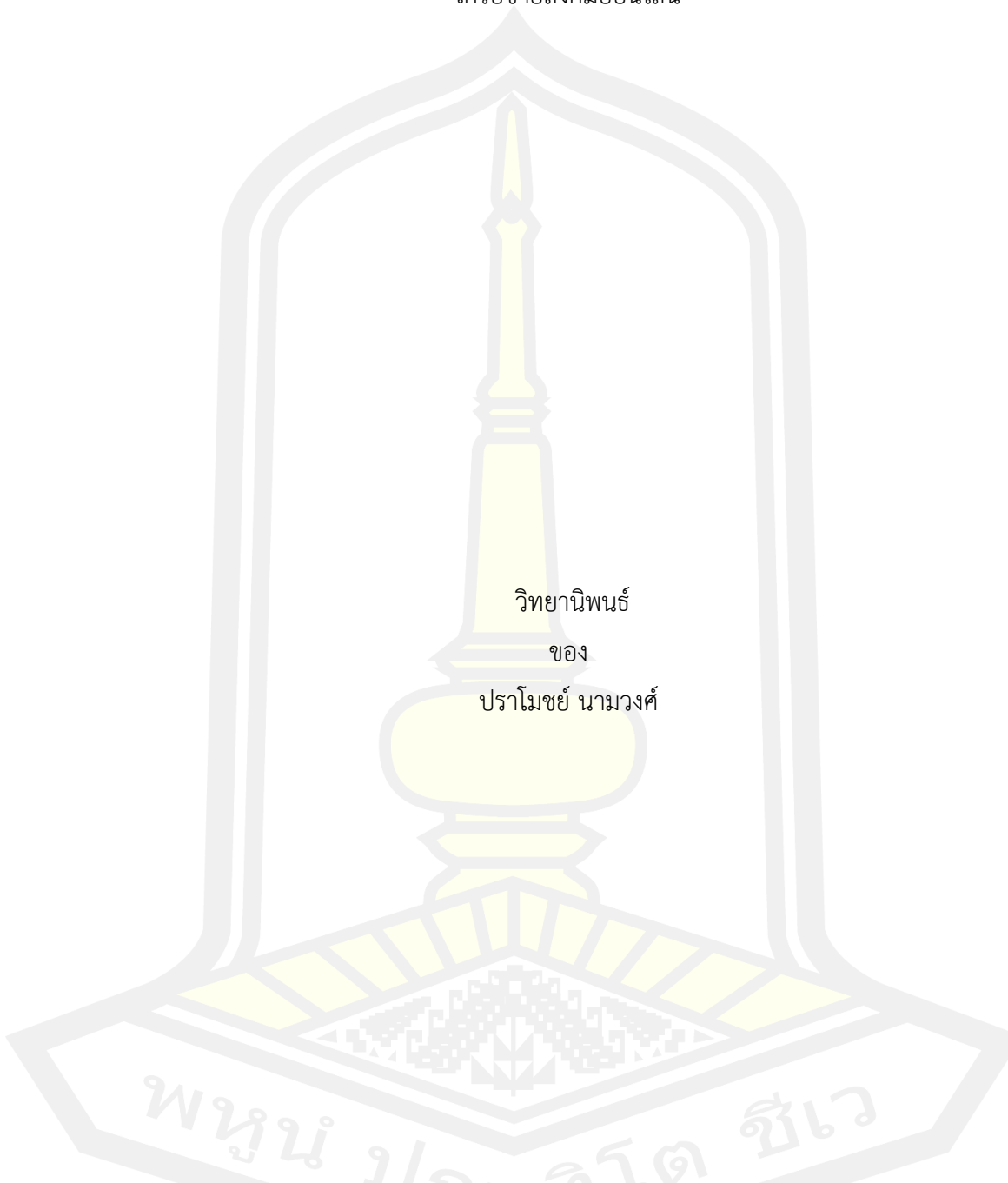
การตรวจจับข้อความประชดประชันด้วยคุณลักษณะร่วมระหว่างบริบทและเนื้อหาข้อความบนเครือข่ายสังคมออนไลน์

วิทยานิพนธ์
ของ
ปราโมชย์ นามวงศ์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
กุมภาพันธ์ 2566

การตรวจจับข้อความประชดประชันด้วยคุณลักษณะร่วมระหว่างบริบทและเนื้อหาข้อความบนเครือข่ายสังคมออนไลน์



วิทยานิพนธ์

ของ

ปราโมชย์ นามวงศ์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

กุมภาพันธ์ 2566

Sarcasm Messages Detection using Hybrid Features Extraction Deriving from Context and Content Sentences on Social Networks

Pramote Namwong

A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Computer Science)

February 2023

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายปราโมชย์ นามวงศ์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รศ. ดร. พนิดา ทรงรัมย์)

.....กรรมการ
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(รศ. ดร. สุชาติ คุ้มมะณี)

.....กรรมการ
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....
(รศ. ดร. จันทิมา พลพินิจ)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การตรวจจับข้อความประชดประชันด้วยคุณลักษณะร่วมระหว่างบริบทและเนื้อหาข้อความบนเครือข่ายสังคมออนไลน์
ผู้วิจัย ปราโมชย์ นามวงศ์
อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. พนิดา ทรงรัมย์
ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2566

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาการตรวจจับข้อความประชดประชันภาษาไทยบนเครือข่ายสังคมออนไลน์ และทำการศึกษาคุณลักษณะที่สกัดจากบริบทของข้อความและคุณลักษณะที่สกัดจากเนื้อหาของข้อความในการจำแนกข้อความประชดประชัน เทคนิคการเรียนรู้เชิงลึกและเทคนิคการเรียนรู้ของเครื่องถูกนำมาประยุกต์ใช้ในการจำแนกข้อความประชดประชัน จากการทดลองแสดงให้เห็นว่าการสกัดคุณลักษณะจากบริบทของข้อความรวมกับเนื้อหาของข้อความให้ค่าความถูกต้องมากที่สุด และวิธี LSTM ให้ค่าความถูกต้องมากที่สุดคือ 96.79% เมื่อกำหนดพารามิเตอร์ Bidirectional LSTM จำนวน 256 โหนด ใช้ Activation function เป็น ReLU Dropout layer เท่ากับ 0.2, Output activation เป็น Sigmoid, loss function แบบ binary cross entropy และ optimizer เป็น adam

คำสำคัญ : การเรียนรู้ของเครื่อง, คำประชดประชัน, การเรียนรู้เชิงลึก, หน่วยความจำระยะสั้นระยะยาว, ปัญญาประดิษฐ์

TITLE Sarcasm Messages Detection using Hybrid Features Extraction Deriving from Context and Content Sentences on Social Networks
AUTHOR Pramote Namwong
ADVISORS Associate Professor Panida Songram , Ph.D.
DEGREE Doctor of Philosophy **MAJOR** Computer Science
UNIVERSITY Mahasarakham University **YEAR** 2023

ABSTRACT

This research aims to develop the detection of sarcastic message in Thai language on social networks. Moreover, context-based and content-based features are extracted from messages and studied for detecting sarcastic messages. Deep leaning and machine learning techniques are applied to classify sarcastic message. From the experimental results, they show that the combination of context-based and content-based features gives the highest accuracy. LSTM gives the highest accuracy at 96.79% when using bidirectional LSTM with 256 nodes, ReLU is used as active function, dropout layer is 0.2, Sigmoid is used as output activation, loss function is binary cross entropy, and adam is used as optimizer.

Keyword : machine learning, sarcasm, deep learning, long short-term memory, artificial intelligence

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จลุล่วงได้ด้วยดี ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. พนิดา ทรงรัมย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำแนะนำและคำปรึกษาในการทำงานวิจัย ตลอดจนการตรวจสอบความถูกต้องของวิทยานิพนธ์

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วรารัตน์ สงฆ์แป้น ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมภูวิเศษ และผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล กรรมการสอบ และผู้ช่วยศาสตราจารย์ ดร.สุชาติ คุ้มมะณี กรรมการสอบ ที่กรุณาให้คำแนะนำตลอดจนคำแนะนำในการปรับปรุงแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์มากขึ้น

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.หทัยรัตน์ ควรรู้ดี คณบดีคณะบริหารธุรกิจและการจัดการ มหาวิทยาลัยราชภัฏอุบลราชธานี ที่กรุณาให้คำแนะนำ ให้กำลังใจ สนับสนุน ให้คำปรึกษา และให้เวลาในการทำวิยานิพนธ์ให้สำเร็จลุล่วงได้ได้ด้วยดี

ขอขอบพระคุณมหาวิทยาลัยราชภัฏอุบลราชธานีที่มอบทุนอุดหนุนการศึกษาและค่าใช้จ่ายในการทำวิจัยให้กับผู้วิจัยในการศึกษาระดับปริญญาเอกในครั้งนี้

ขอขอบพระคุณ ครู อาจารย์ ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยอย่างมากมายจนประสบความสำเร็จ

ท้ายที่สุดขอขอบพระคุณ บิดา มารดา ผู้ให้กำเนิด ญาติพี่น้อง ภรรยาและลูก ที่ส่งเสริมให้ผู้วิจัยมีเวลาในการดำเนินการวิจัย ทั้งยังให้กำลังใจแก่ผู้วิจัยจนประสบความสำเร็จใจชีวิต

ปราโมชย์ นามวงศ์

สารบัญ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

การเติบโตของเครือข่ายสังคมออนไลน์เป็นไปอย่างรวดเร็ว การสื่อสารผ่านข้อความมีความสะดวกและรวดเร็วมากขึ้น ทำให้จำนวนข้อความแสดงความคิดเห็น ทัศนคติ ผ่านทางเครือข่ายสังคมออนไลน์มีแนวโน้มว่าจะเพิ่มสูงมากขึ้น ซึ่งข้อความที่เกิดขึ้นจำนวนมากนั้น มีทั้งข้อความที่เป็นข้อเท็จจริงและความคิดเห็น ข้อความคิดเห็นเป็นข้อความที่แสดงออกถึงอารมณ์และความรู้สึกของผู้เขียน ข้อความคิดเห็นสามารถแบ่งออกเป็น 3 ขั้วหลักๆ คือ ข้อความคิดเห็นที่แสดงความคิดเห็นในเชิงบวก (Positive Opinions) และความคิดเชิงลบ (Negative Opinions) และความคิดเห็นที่เป็นกลาง (Neutral Opinions) ข้อความคิดเห็นที่เพิ่มขึ้นจำนวนมหาศาลบนเครือข่ายสังคมออนไลน์นี้ได้ถูกนำมาใช้ในการวิเคราะห์ด้านต่างๆ เช่น ผู้ลงสมัครเลือกตั้งสามารถนำข้อความคิดเห็นมาใช้ในการสำรวจคะแนนความนิยมของประชาชนต่อพรรคการเมืองของตนหรือเพื่อนำเสนอเรื่องที่ประชาชนให้ความสนใจ ผู้ให้บริการโรงแรมสามารถสำรวจความพึงพอใจต่อการให้บริการเพื่อใช้ในการปรับปรุงการให้บริการโดยใช้ข้อความคิดเห็น ผู้ผลิตสินค้าและบริการสามรถสำรวจจความพึงพอใจของผู้บริโภคต่อสินค้าและบริการผ่านข้อความคิดเห็น เพื่อนำผลการวิเคราะห์ไปใช้ในการปรับปรุงสินค้าและบริการให้ดียิ่งขึ้น หรือเพื่อ พัฒนาสินค้าให้ตรงตามความต้องการของผู้บริโภค นอกจานี้ข้อความคิดเห็นยังสามารถในในด้านการศึกษา โดยการสำรวจความพึงพอใจของผู้เรียนต่อการจัดการเรียนการสอน สิ่งสนับสนุนการเรียนรู้จากข้อความคิดเห็น เพื่อนำมาปรับปรุงในการจัดการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น เป็นต้น ปัจจุบันมีการนำเสนอเทคนิคทางด้านคอมพิวเตอร์เพื่อใช้ในการวิเคราะห์อารมณ์ความรู้สึกหรือทัศนคติ (Sentiment Analysis) จากข้อความคิดเห็น วิธีการที่ได้รับความนิยมมากที่สุด คือ การทำเหมืองความคิดเห็น (Opinion Mining) ซึ่งเป็นวิธีการในการสกัดเอาความรู้สึกจากข้อความจำนวนมหาศาล มาใช้ประโยชน์ นักวิจัยจำนวนมากทำการศึกษาในด้านนี้ เช่น Chan และ Chong [1] ได้ศึกษาการวิเคราะห์ความรู้สึกของนักลงทุนในตลาดหุ้น ซึ่งการวิเคราะห์ความรู้สึกของนักลงทุนที่แสดงออกมาผ่านข้อความนั้นจะเป็นประโยชน์ต่อการวิเคราะห์แนวโน้มในตลาดหุ้น Tartir และ Abdul-Nabi [2] ได้นำเสนอการค้นหาวิเคราะห์ความรู้สึก ทัศนคติและข้อมูลเชิงลึกทางธุรกิจจากสื่อสังคมออนไลน์ทวิตเตอร์ของอาหรับ โดยศึกษาความรู้สึกที่แตกต่างกัน ทำให้เข้าใจความรู้สึกของนิติบุคคลบางอย่างที่เป็นสิ่งสำคัญสำหรับใช้ในการตัดสินใจของผู้ผลิต ในการวางแผนที่จะขยายงานในอนาคต Gitto และ Mancuso [3] นำเสนอการวิเคราะห์ทัศนคติ ความพึง

พอใจของผู้รับบริการของสนามบินโดยการวิเคราะห์ความรู้สึกของลูกค้า ความคิดด้านบวก (Positive) ด้านลบ (Negative) และ ความคิดเห็นเป็นกลาง (Neutral) เพื่อนำผลการวิเคราะห์ไปปรับปรุงการให้บริการของสนามบิน เป็นต้น

ในการวิเคราะห์ความรู้สึกเพื่อทราบถึงทัศนคติในทางบวกหรือทางลบ นั้นบางครั้ง ไม่เพียงพอที่จะทำให้ทราบถึงความคิดเห็นที่แท้จริง ทำให้การวิเคราะห์ข้อมูลนั้นอาจ ผิดพลาดได้เพราะบางครั้งข้อความที่สื่อสารออกมานั้นไม่ตรงกับความหมายที่แท้จริง ข้อความที่สื่อออกมาในลักษณะนี้เรียกว่าเป็นข้อความประเภท “การประชดประชัน” คือ ข้อความนั้นมีความหมายตรงกันข้ามกับความหมายที่แท้จริง เช่น “วันนี้อากาศดีจริงเลยยยยยยยยยยยยย” “โรงแรมนี้บริการดีมากจนอยากกลับมาพักอีกซะเมื่อไหร่” “นายกรัฐมนตรีบริหารงานดีมากจนประชาชนอดอยาก ขอให้ท่านบริหารงานต่อเลยนะประชาชนทนได้” “I love being ignored all the time” เป็นต้น จากประโยคตัวอย่างจะเห็นได้ว่าข้อความที่เขียนขึ้นเขียนโดยใช้คำเชิงบวกแต่ต้องการจะสื่อถึงความรู้สึกเชิงลบ หรือบางข้อความเขียนขึ้นโดยใช้คำในเชิงลบแต่ต้องการสื่อถึงความรู้สึกเชิงบวก ดังนั้นการพิจารณาความหมายของคำจากข้อความเหล่านี้ว่าเป็นบวก ลบ นั้นยังไม่เพียงพอที่จะทำให้ทราบถึงความคิดเห็นที่มีความหมายจริงๆ ของข้อความ การจำแนกข้อความประชดประชันนั้นจึงเป็นเรื่องที่ยากและท้าทายในงานด้านการวิเคราะห์ความรู้สึก

ปัจจุบันได้มีการนำเสนอเทคนิควิธีการจำแนกข้อความประชดประชัน ด้วยวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) เช่น Jasso และ Meza [4] จำแนกข้อความประชดประชันในภาษาสเปน ที่อยู่บนเครือข่ายสังคมออนไลน์ทวิตเตอร์ โดยใช้ขั้นตอนวิธี Support Vector Machine และ Random Forest, Bouazizi และ Otsuki [5] ใช้รูปแบบของข้อความในการตรวจจับข้อความประชดประชัน Razali และคณะ [6] ใช้เทคนิคการเรียนรู้เชิงลึกกับคุณลักษณะบริบทของข้อความในการตรวจจับข้อความประชดประชัน และยังมีนักวิจัยให้ความสนใจในการศึกษาการจำแนกข้อความคิดเห็นประชดประชันในหลากหลายภาษายังคงเป็นงานที่ท้าทายเนื่องจากแต่ละภาษามีเอกลักษณ์ทางภาษาที่แตกต่างกัน ทำให้กระบวนการวิธีในการจำแนกแตกต่างกันไป เช่น ภาษาไทยเป็นภาษาที่มีความซับซ้อนกว่าภาษาอังกฤษ การจำแนกข้อความประชดประชันภาษาไทยจึงถือว่าเป็นงานที่ท้าทายและยังถือเป็นเรื่องใหม่เนื่องจากยังไม่ได้มีการศึกษาการจำแนกข้อความประชดประชันภาษาไทย

ดังนั้นงานวิจัยนี้จึงนำเสนอกระบวนการในการจำแนกข้อความคิดเห็นประชดประชันภาษาไทยบนเครือข่ายสังคมออนไลน์และศึกษาคุณลักษณะที่สามารถจำแนกข้อความคิดเห็นประชดประชันภาษาไทยได้อย่างมีประสิทธิภาพ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนากระบวนการวิธีการจำแนกข้อความประชดประชันภาษาไทยบนเครือข่ายสังคมออนไลน์

2. เพื่อศึกษาคุณลักษณะที่ใช้ในการจำแนกข้อความประชดประชัน

1.3 ความสำคัญของการวิจัย

1. งานวิจัยนี้นำเสนอกระบวนในการจำแนกข้อความประชดประชันภาษาไทยเพื่อให้ทราบอารมณ์ความรู้สึกที่แท้จริงและซึ่งจะช่วยเพิ่มประสิทธิภาพในงานด้านการวิเคราะห์อารมณ์ความรู้สึกได้ถูกต้องมากยิ่งขึ้น

2. งานวิจัยนี้ทำการศึกษาคุณลักษณะข้อความของภาษาไทยที่สามารถใช้ในการจำแนกข้อความประชดชัน ทำให้ทราบถึงคุณลักษณะที่เหมาะสมในการจำแนกข้อความประชดประชันภาษาไทยและพบว่าการรวมกันระหว่างคุณลักษณะจากบริบทของข้อความ (Context-based feature) และ คุณลักษณะจากเนื้อหาในข้อความ (Content-based feature) ช่วยให้การจำแนกข้อความประชดชันมีประสิทธิภาพยิ่งขึ้น

1.4 ขอบเขตของการวิจัย

1. จำแนกข้อความประชดประชันภาษาไทยที่อยู่นเครือข่ายสังคมออนไลน์ Facebook จำนวน 10,800 ข้อความ แบ่งเป็นข้อความประชดประชันจำนวน 5,400 ข้อความ และข้อความไม่ประชดประชันจำนวน 5,400 ข้อความ

2. คุณลักษณะที่ใช้ในการจำแนกข้อความประชดประชันประกอบไปด้วยคุณลักษณะจากบริบทของข้อความและคุณลักษณะจากเนื้อหาในข้อความซึ่งสกัดจากข้อความภาษาไทย เครื่องหมายวรรคตอนและไอคอนแสดงอารมณ์ (Emoticons)

3. ใช้แฮชแท็ก #ประชด #ประชดประชัน สำหรับเก็บรวบรวมข้อความประชดประชัน

4. การจำแนกข้อความประชดประชันแบ่งออกเป็น 2 แบบ คือ ข้อความประชดประชัน และข้อความไม่ประชดประชัน

5. ประเมินประสิทธิภาพในการจำแนกข้อความประชดประชัน โดยใช้ ค่าความถูกต้อง ค่าระลึก ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวม

1.5 นิยามศัพท์เฉพาะ

1. การจำแนกความคิดเห็นประชดประชัน คือ การจำแนกข้อความประชดประชันที่สามารถบ่งบอกว่าข้อความนั้นเป็นข้อความประชดประชันหรือเป็นข้อความไม่ประชดประชัน

2. คุณลักษณะจากบริบทของข้อความ คือ การใช้คำในชุดข้อมูลมาสร้างเป็นคุณลักษณะเพื่อใช้เป็นตัวแทนคุณลักษณะทั้งหมดของชุดข้อความ

3. คุณลักษณะจากเนื้อหาในข้อความ คือ การใช้เนื้อหาในชุดข้อมูลมาสร้างเป็นคุณลักษณะ เช่น การมีไอคอนแสดงอารมณ์ การมีเครื่องหมายคำถาม การมีข้อความแสดงการหัวเราะ เป็นต้น

4. การวิเคราะห์ความรู้สึกบนเครือข่ายสังคมออนไลน์ คือ การวิเคราะห์ความรู้สึก ทัศนคติ ผ่านข้อความที่ผู้ใช้เครือข่ายสังคมออนไลน์ได้ โพสต์ การแสดงความคิดเห็น การรีวิวสินค้า

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้ศึกษา ทบทวนทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อใช้ในการดำเนินงานวิจัย ซึ่งมีหัวข้อดังต่อไปนี้ การวิเคราะห์ข้อคิดเห็นประชดประชัน การเตรียมข้อมูล (Data Preparation) วิธีการเรียนรู้ของเครื่อง (Machine Learning approach) วิธีการเรียนรู้เชิงลึก (Deep Learning) การวัดประสิทธิภาพการจำแนก และงานวิจัยที่เกี่ยวข้อง เพื่อใช้เป็นทฤษฎีในการศึกษาการพัฒนาวิธีการจำแนกข้อความประชดประชันสำหรับการวิเคราะห์ความรู้สึกข้อความภาษาไทยบนเครือข่ายสังคมออนไลน์ โดยรายละเอียดในแต่ละหัวข้อมีดังนี้

2.1 การวิเคราะห์ข้อคิดเห็นประชดประชัน

การประชดประชัน เป็นลักษณะที่สื่อถึงการแสดงออกทางข้อความที่มีความหมายที่อาจจะตรงข้ามกับความหมายที่แท้จริงของผู้สื่อสาร ซึ่งแสดงออกถึงอารมณ์หลากหลาย [5] เช่น การแสดงการประชดประชันที่แสดงออกถึงการตลกขบขัน การแสดงถึงการแสดงออกที่มีแนวโน้มการพูดเกินจริง การใช้น้ำเสียงที่แตกต่าง การแสดงออกถึงการเสียดสีที่แสดงออกถึงความรำคาญหรือมีความโกรธ การใช้คำที่ใช้การสื่อสารในแง่บวกอย่างมากเพื่ออธิบายสถานการณ์เชิงลบ หรือการใช้คำเสียดสีประชดประชันเพื่อแสดงออกถึงสถานการณ์ที่ต้องการหลีกเลี่ยงการให้คำตอบที่ชัดเจน บุคคลนั้นจะใช้ประโยคที่ซับซ้อน คำที่ไม่ได้ใช้ในปกติทั่วไป และใช้นิพจน์บางอย่างที่ไม่ปกติ การแสดงออกถึงการประชดประชันนั้นบุคคลสามารถรับรู้ได้ง่ายหากการแสดงออกนั้นเกิดขึ้นซึ่งๆ หน้า การแสดงออกต่อหน้า เห็นสีหน้าท่าทาง ได้ยินเสียงจะสามารถวิเคราะห์ได้ว่าบุคคลแสดงออกถึงการประชดประชันหรือไม่ แต่การแสดงถึงการประชดประชันด้วยข้อความบนเครือข่ายสังคมออนไลน์นั้น เป็นเรื่องยากที่จะทราบได้ว่าบุคคลนั้นได้แสดงออกออกถึงการประชดประชันหรือไม่ นอกจากผู้ที่เป็นเจ้าของข้อความนั้น ดังนั้นงานด้านการวิเคราะห์ข้อความประชดประชันบนเครือข่ายสังคมออนไลน์นั้นจึงเป็นงานที่มีความสำคัญที่จะช่วยส่งเสริมการพัฒนางานด้านการวิเคราะห์อารมณ์ความรู้สึกให้มีประสิทธิภาพมากขึ้น

การวิเคราะห์ความรู้สึกเป็นเทคนิคในการระบุความคิดเห็น ทัศนคติ และอารมณ์ของบุคคลที่มีต่อสิ่งต่างๆ [1, 2, 7, 8] โดยมีวัตถุประสงค์ในการจำแนกความคิดเห็นโดยการนำหลักการทำเหมืองความคิดเห็น (Opinion Mining) ในการวิเคราะห์ความรู้สึกนั้นจะช่วยให้เห็นว่าผู้บริโภคชอบและไม่ชอบ พึงพอใจต่อสินค้าและบริการหรือไม่ ความคิดเห็นของผู้ใช้งานจากสื่อสังคมออนไลน์เป็นแหล่งที่มีประโยชน์อย่างมาก เนื่องจากเป็นเครือข่ายที่ใหญ่ มีผู้คนแสดงความคิดเห็น หรือโพสต์ข้อความอยู่

ตลอดเวลา งานวิจัยด้านการวิเคราะห์ความรู้สึกนั้นมีการนำมาประยุกต์ใช้กับข้อมูลการแสดงความคิดเห็นต่อสินค้าและบริการ จุดประสงค์เพื่อวิเคราะห์ความรู้สึกของผู้บริโภค ว่ามีความรู้สึกอย่างไรต่อสินค้าและบริการ สามารถวิเคราะห์ความคิดเห็นออกเป็น ความคิดเห็นเป็นบวก ลบ และเป็นกลางต่อสินค้าและบริการนั้น

ในบางครั้งข้อความที่สื่อออกมาอาจจะเป็นข้อความที่มีความหมายไม่ตรงกับความหมายที่แท้จริง สาเหตุเนื่องจากเหตุผลบางอย่างที่ทำให้ไม่สามารถแสดงความคิดเห็นได้ เช่น ความคิดเห็นที่แสดงออกไปอาจผิดกฎหมาย จึงจำเป็นต้องหลีกเลี่ยงการใช้คำที่มีความตรงๆ เปลี่ยนไปใช้คำที่ต่างจากปกติ หรือ ใช้สัญลักษณ์อื่นๆ ในการแสดงข้อความประเภทนี้ เรียกว่า “ข้อความคิดเห็นประชดประชัน” ซึ่งในการแสดงความคิดเห็นประชดประชันนั้นเกิดจากเหตุผลหลายอย่าง เช่น มีจุดประสงค์เพื่อต้องการสื่อถึงเรื่องตลกขบขัน แสดงออกเมื่อมีอารมณ์โกรธ และแสดงอาการในการหลีกเลี่ยงการตอบคำถาม เป็นต้น ในการจำแนกหรือการวิเคราะห์ความคิดเห็นนั้นบางครั้งหากข้อความเหล่านั้นถูกวิเคราะห์ความเห็นออกมาเป็นความหมายเชิงบวก ซึ่งความหมายที่แท้จริงหากข้อความนั้นเป็นข้อความประเภทประชดประชัน นั่นอาจจะให้ความหมายเชิงลบ ดั้งนั้น ในงานทางด้านการตรวจจับข้อความคิดเห็นประชดประชันนั้นถือว่าเป็นงานที่มีความสำคัญที่จะช่วยส่งเสริมให้งานด้านการวิเคราะห์อารมณ์ความรู้สึกสามารถวิเคราะห์ข้อมูลได้ถูกต้องตามความหมายที่แท้จริงมากยิ่งขึ้น

2.2 การเตรียมข้อมูล

ขั้นตอนในการเตรียมข้อมูลนั้นเป็นขั้นตอนที่สำคัญในการที่จะนำข้อมูลไปทำการวิเคราะห์ เนื่องจากข้อความคิดเห็นเป็นข้อมูลที่อยู่ในรูปแบบไม่มีโครงสร้าง (Unstructured Data) จึงจำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบที่มีโครงสร้างก่อน โดยวิธีการเตรียมข้อมูล [9-11] เพื่อทำการกำจัดข้อมูลที่มีสิ่งรบกวน (Noise) และข้อมูลที่ไม่สำคัญกับการนำไปใช้งาน เพื่อให้ได้ข้อมูลให้อยู่ในรูปแบบเดียวกัน (Consistency) ที่สามารถนำไปประมวลผลและทำให้การวิเคราะห์ข้อมูลไม่ผิดพลาด ทำให้มีความถูกต้อง (Accuracy) มากยิ่งขึ้น โดยกระบวนการเตรียมข้อมูลประกอบไปด้วยดังหัวข้อต่อไปนี้

2.2.1 การตัดคำ (Tokenization)

การตัดคำถือได้ว่าเป็นความจำเป็นพื้นฐานที่สำคัญขั้นตอนหนึ่งที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ซึ่งข้อความเกิดจากการรวมกันของคำ เป็นประโยค ดังนั้นกระบวนการตัดคำคือกระบวนการที่การนำเอาเอกสารหรือข้อความมาแบ่งเป็น

ประโยค (Sentence) หรือ คำ (Word) [12, 13] การตัดคำในข้อความภาษาอังกฤษนิยมใช้ ช่องว่าง (White Space) คอมมา (Comma: ,) จุดทศนิยม (Point: .) เครื่องหมายอัฒภาค (Semicolon: ;) เครื่องหมายคำถาม (Question Mark: ?) เครื่องหมายวรรคตอนหรือสัญลักษณ์ต่างๆ ในขั้นตอนการตัดคำจะเริ่มจากการค้นหาตรวจสอบข้อความทั้งหมดเพื่อหาขอบเขตของคำและประโยค ซึ่งในภาษาอังกฤษนั้นจะใช้ช่องวางในการแบ่งคำออกจากกัน และใช้จุดเป็นตัวบอกว่าจบประโยค ในส่วนของการตัดคำในภาษาไทยนั้น สามารถทำได้หลายวิธี เช่น การตัดคำโดยการเทียบคำที่ยาวที่สุด (Longest Matching) วิธีนี้จะค้นหาคำ โดยเริ่มจากการพิจารณาข้อความทั้งข้อความแล้วเปรียบเทียบกับคำในพจนานุกรมว่ามีคำนั้นหรือไม่ หากไม่มี จะพิจารณาตัดตัวอักษรตัวสุดท้ายออกและนำข้อความที่เหลือไปเปรียบเทียบกับคำในพจนานุกรมอีกครั้ง และจะทำซ้ำไปเรื่อยจนกว่าจะได้ข้อความที่มีในพจนานุกรม ตัวอย่างเช่น การแบ่งคำในประโยค "ฉันนั่งตากลมที่ริมหาด" จะเริ่มพิจารณาทั้งประโยคแล้วนำไปเปรียบเทียบกับพจนานุกรม หากพบว่าทั้งข้อความมีอยู่ในพจนานุกรมก็จะได้ข้อความนั้น หากไม่พบก็จะเริ่มตัดตัวอักษรออกทีละตัวจนพบข้อความ หลังจากนั้น ก็จะทำการเปรียบเทียบต่อไป ซึ่งจะได้ผลลัพธ์ออกมาคือ "ฉัน นั่ง ตาก ลม ที่ ริม หาด" ตัวอย่างการตัดคำดังแสดงในตารางที่ 1

วิธีการตัดคำแบบเหมือนมากที่สุดหรือให้มีจำนวนคำน้อยที่สุด (Maximal Matching) โดยจะใช้วิธีในการตัดคำที่สามารถจะเป็นไปได้ทั้งหมด จากนั้นจะเลือกจำนวนคำที่น้อยที่สุดเป็นคำตอบ เช่น การแบ่งคำในประโยค "ฉันชอบไปโรงเรียน" เมื่อตรวจสอบตามพจนานุกรมจะได้ผลลัพธ์การตัดคำออกมาเป็นสอบแบบคือ “ฉัน ชอบ ไป โรง เรียน” และ “ฉัน ชอบ ไป โรงเรียน” แบบที่ 1 ได้จำนวนคำ 5 คำ แบบที่ 2 ได้จำนวนคำ 4 คำ ดังนั้นจริงเลือกแบบที่ 2 ซึ่งมีจำนวนคำน้อยกว่า ดังแสดงตัวอย่างในตารางที่ 2

ตารางที่ 1 ตัวอย่างการตัดคำโดยการเทียบคำที่ยาวที่สุด (Longest Matching)

ประโยค	Longest Matching	ผลของการตัด
"ฉันนั่งตากลมที่ริมหาด"	ฉันนั่งตากลมที่ริมหาด	
	ฉันนั่งตากลมที่ริมหา~~ด~~	
	. .	
	~~ฉันนั่งตากลมที่ริมหาด~~	ฉัน
	นั่ง~~ตากลมที่ริมหาด~~	ฉัน | นั่ง

ตารางที่ 1 ตัวอย่างการตัดคำโดยการเทียบคำที่ยาวที่สุด (Longest Matching) (ต่อ)

	ตาก~~ลมที่ริมหาด~~	ฉัน | นั่ง | ตาก
	ลม~~ที่ริมหาด~~	ฉัน | นั่ง | ตาก | ลม
	~~ที่ริมหาด~~	ฉัน | นั่ง | ตาก | ลม | ที่
	ริม~~หาด~~	ฉัน | นั่ง | ตาก | ลม | ที่ | ริม
	หาด	ฉัน | นั่ง | ตาก | ลม | ที่ | ริม | หาด

ตารางที่ 2 ตัวอย่างการตัดคำภาษาไทยจำนวนคำน้อยที่สุด (Maximal Matching)

ประโยค	Maximal Matching	จำนวนคำ
ฉันชอบไปโรงเรียน	ฉัน | ชอบ | ไป | โรง | เรียน	5
	ฉัน | ชอบ | ไป | โรงเรียน	4

2.2.2 การสกัดคุณลักษณะ

การสกัดคุณลักษณะ [10] หมายถึง การนำเอาคุณลักษณะของเอกสารออกมา ซึ่งคำที่นำมาเป็นคุณลักษณะนั้นจะต้องสามารถเป็นตัวแทนของเอกสารนั้นได้ ซึ่งในการเลือกคุณลักษณะนั้นสามารถทำได้หลายวิธี เช่น วิธีแบบชุดลำดับคำ (N-Grams) [4, 5] การใช้การแทนข้อความด้วยถุงคำ (Bag-of-Word) [14] การใช้หน้าที่ของคำ (Part of Speech) [15-17] หลังจากคำคุณลักษณะที่เป็นตัวแทนแล้วจะแทนค่าคุณลักษณะเหล่านั้นให้อยู่ในรูปแบบเวกเตอร์ (Vector) โดยมีวิธีการคำนวณค่าน้ำหนักได้หลายวิธีและที่นิยมใช้เพื่อแทนค่าคุณลักษณะ มีดังนี้

1) Boolean Weighting เป็นการคำนวณค่าน้ำหนักจากการปรากฏคำในเอกสาร ซึ่งได้รวบรวมไว้แล้วในถุงคำ หากมีคำปรากฏอยู่ในเอกสารที่ตรงกับถุงคำ จะให้ค่าน้ำหนักเป็น 1 ถ้าไม่ปรากฏอยู่ในเอกสาร จะให้ค่าน้ำหนักเป็น 0 ค่าน้ำหนักในลักษณะนี้เรียกอีกอย่างว่า ค่าคุณลักษณะความจริง (Boolean Feature) ซึ่งมีค่าเป็นไบนารี ดังแสดงในสมการที่ (1)

$$B_{td} = \begin{cases} 1, for \text{ term present in document} \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

โดยที่ B_{td} คือ ค่าการเกิดคุณลักษณะ t ในเอกสาร d

ตัวอย่างการคำนวณค่า Boolean Weighting ของข้อความ ต่อไปนี้ “เป็นไปไม่ได้”, “ความพยายามอยู่ที่ไหน”, “ทำดีย่อมได้ดี” เมื่อตัดคำแล้วแค่ค่าในตารางเวกเตอร์สเปส แล้วแทนค่าน้ำหนักในถุงคำ ซึ่งหากมีคำดังนี้ “เป็น, ไป, ได้, ไม่, รัก, ความ, พยายาม, เสมอ, ดี, ทำ” ตัวอย่างดังแสดงในตารางที่ 3

ตารางที่ 3 ตัวอย่างการคำนวณค่า Boolean Weighting

Doc	เป็น	ไป	ได้	ไม่	รัก	ความ	พยายาม	เสมอ	ดี	ทำ
1	1	1	1	1	0	0	0	0	0	0
2	0	0	0	0	0	1	1	0	0	0
3	0	0	1	0	0	0	0	0	1	1

2) Term Frequency: TF [12] เป็นการแทนข้อความด้วยค่าน้ำหนักความถี่ในการเกิดคำในเอกสาร หากคุณลักษณะใดปรากฏบ่อยในเอกสารค่าน้ำหนักความถี่ย่อมมีค่าสูงและหากคำใดไม่ปรากฏในเอกสารเลยจะให้ค่าน้ำหนักความถี่เป็น 0 สามารถคำนวณได้ดังสมการที่ (2) ในกรณีที่เอกสารแต่ละเอกสารมีความยาวไม่เท่ากัน หรือแตกต่างกันมาก เช่น เอกสารที่ 1 มี 5 คำ แต่เอกสารที่ 2 มีจำนวน 100 คำ เพื่อป้องกันไม่ให้ความยาวของเอกสารมีผลต่อการคำนวณและช่วยป้องกันไม่ให้เกิดความแตกต่างกันในการคำนวณของแต่ละเอกสารนิยมนำวิธี normalization ซึ่งเหมาะสำหรับข้อมูลที่มีความยาวของเอกสารไม่เท่ากัน ดังสมการที่ (3)

$$tf_{td} = freq(t,d) \tag{2}$$

เมื่อ $freq(t,d)$ คือ ค่าความถี่ของการปรากฏคุณลักษณะ t ในเอกสาร d

$$Normalization(tf_{td}, K) = K + (1-K)\frac{freq(t,d)}{\max(freq(d))} \tag{3}$$

เมื่อ K คือ มีค่าอยู่ระหว่าง 0 ถึง 1

เมื่อ $\max(freq(d))$ คือ ค่าความถี่สูงสุดของการปรากฏคุณลักษณะในเอกสาร d

3) Term Frequency – Inverse Document Frequency: TF-IDF เป็นการคำนวณค่าน้ำหนักจากความถี่และความถี่ผกผันของการปรากฏคำ ในเอกสาร และจะพิจารณาความถี่ของคำที่ปรากฏในเอกสาร ร่วมด้วย การคำนวณด้วยวิธีนี้เมื่อการแทนค่าความถี่การปรากฏของคุณลักษณะอย่างเดียวไม่เพียงพอ เนื่องจากจะไม่สามารถจำแนกได้หากคุณลักษณะนั้นปรากฏขึ้นเป็นจำนวนมากในทุกเอกสาร แสดงว่าคุณลักษณะดังกล่าวไม่สามารถใช้เป็นตัวแทนของเอกสารนั้นได้ การหาน้ำหนักแบบ TF-IDF สามารถคำนวณได้ดังสมการที่ (4)

$$idf_{td} = \log(\frac{N}{D_t}) \tag{4}$$

เมื่อ N คือ จำนวนเอกสารทั้งหมด

D_t คือ จำนวนเอกสารทั้งหมดที่มีคุณลักษณะ t ปรากฏอยู่

การหาคำนวนหาค่าความถี่และความถี่ผกผัน จะคำนึงถึงความถี่ของการปรากฏคุณลักษณะในเอกสาร และค่าความถี่ผกผัน สามารถคำนวณได้ดังสมการที่ (5)

$$tfidf_{td} = tf_{td} \times idf_{td} \quad (5)$$

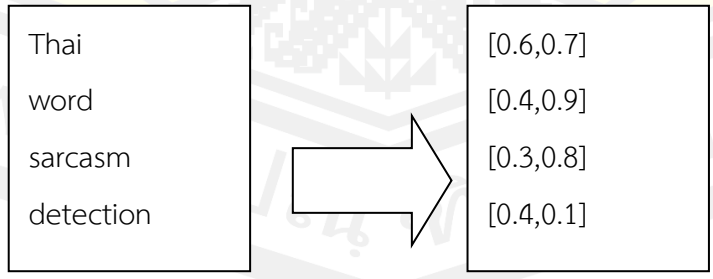
2.2.3 รูปแบบการแทนข้อความ (Document representation)

2.2.3.1 โมเดลเชิงพื้นที่แบบเวกเตอร์ (Vector Space Model)

โมเดลเชิงพื้นที่แบบเวกเตอร์ [18] เป็นรูปแบบในการแทนข้อความ ซึ่งมีวัตถุประสงค์เพื่อใช้ตัวเลขเป็นตัวแทนของเอกสารเป็นการให้ค่าน้ำหนักคำเอกสารที่ไม่มีโครงสร้างเพื่อให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้ ซึ่งในเอกสารสามารถระบุเอกสาร $D = d_x, x = 1,2,3,...n$ ซึ่งแต่ละ d_x แทนจำนวนคำที่ไม่ซ้ำกันที่มาจากชุดของเอกสาร D

2.2.3.2 Word Embedding

Word Embedding [19-21] เป็นวิธีการสร้างคุณลักษณะของประโยคหรือเอกสารขึ้นมาให้อยู่ในรูปแบบเวกเตอร์ หรือเรียกว่าการสร้างคุณลักษระเวกเตอร์ขึ้นมาจากประโยคหรือเอกสารที่มีอยู่ในข้อมูลเพื่อสร้างคุณลักษณะที่อยู่ในรูปแบบตัวเลข โดยการเข้ารหัสคำแต่ละคำให้อยู่ในรูปแบบที่สามารถนำไปคำนวณที่คอมพิวเตอร์สามารถทำความเข้าใจได้ ซึ่งข้อดีของการทำ Word Embedding นั้นจะสามารถนำไปใช้ในการคำนวณความคล้ายคลิงกับคำอื่น ๆ ในบริบทของคำที่แตกต่างกันได้ โดยลักษณะการสร้างเวกเตอร์คุณลักษณะจะเริ่มจากการเข้ารหัสคำแต่ละคำด้วยวิธี One-Hot Encoding ซึ่งจะทำงานโดยการนำจำนวนคำที่ปรากฏขึ้นมาในชุดข้อมูล และนำประโยคในชุดข้อมูลหรือเอกสารที่ได้ทำการกำหนดไว้ในชุดข้อมูลมาเข้ารหัส ทำให้ได้เวกเตอร์ตามจำนวนคำในประโยคที่กำหนดให้อยู่ในรูปของบิต จากนั้นทำการรวมเวกเตอร์ที่ได้จาก One-Hot Encoding ซึ่งสามารถกำหนดจำนวนมิติหรือคุณลักษณะของเวกเตอร์ได้ ตัวอย่างการใช้งาน Word Embedding ในการแปลงข้อความเป็นเวกเตอร์ ดังแสดงในรูปที่ 1



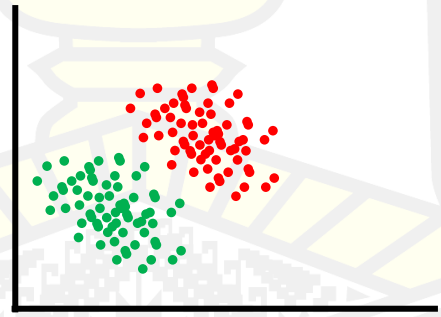
รูปที่ 1 ตัวอย่างการใช้งาน Word Embedding ในการแปลงข้อความให้อยู่ในรูปแบบเวกเตอร์

2.3 การจำแนกข้อมูลด้วยวิธีการเรียนรู้ด้วยเครื่อง

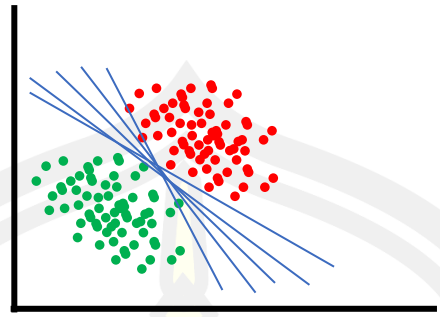
การจำแนกข้อมูลด้วยวิธีการเรียนรู้ด้วยเครื่อง (Machine Learning approach) เป็นความพยายามทำให้เครื่องคอมพิวเตอร์หรือเครื่องอิเล็กทรอนิกส์ให้สามารถทำงานที่มีความฉลาดได้ โดยการสร้างเครื่องจักรที่สามารถเรียนรู้ได้ ซึ่งเป็นเทคนิคหนึ่งที่ใช้ในการจำแนกข้อความหรือเอกสาร ซึ่งนิยมนำมาใช้ในงานด้านการวิเคราะห์ความรู้สึก [3, 22-24] โดยจะใช้เทคนิควิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งเป็นเทคนิคในการเรียนรู้ด้วยเครื่องซึ่งเป็นการนำเข้าข้อมูลที่มีอยู่เข้าสู่ระบบเพื่อใช้สร้างข้อมูลชุดสอน (Training dataset) เพื่อใช้สำหรับการหาคำตอบให้กับชุดข้อมูลใหม่ที่ยังไม่รู้คำตอบ (Testing dataset) และวิธีการที่นิยมในการจำแนกความคิดเห็นประกอบไปด้วย ซัพพอร์ทเวกเตอร์แมคชีน นาอีฟเบย์ เพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ

1) ซัพพอร์ทเวกเตอร์แมคชีน

ซัพพอร์ทเวกเตอร์แมคชีนนิยมนำมาใช้ในการจัดกลุ่มข้อมูล [4, 17, 25] ซึ่งเป็นการแบ่งข้อมูลในลักษณะการใช้เส้นตรงเรียกว่าเส้นระนาบแบ่งข้อมูล (Hyper plane) ในการแบ่งข้อมูลออกจากกันเพื่อหาเส้นแบ่งที่เหมาะสมที่สุด ข้อดีของซัพพอร์ทเวกเตอร์แมคชีนคือสามารถแบ่งกลุ่มข้อมูลได้ทั้งรูปแบบข้อมูลที่แบ่งกลุ่มที่เป็นเส้นตรงและไม่เป็นเส้นตรง นอกจากนี้ยังรองรับคุณลักษณ์จำนวนมากได้ เนื่องจากเป็นการใช้การแทนข้อมูลด้วยเวกเตอร์แลจะพิจารณาเส้นแบ่งข้อมูลจากเวกเตอร์ซัพพอร์ท (Support Vector) แต่ข้อเสียคือต้องทดลองเพื่อปรับค่าพารามิเตอร์ให้เหมาะสมสำหรับแต่ละเคอร์เนล (Kernel) ที่เลือกใช้ จากชุดข้อมูลดังรูปที่ 2 และ รูปที่ 3



รูปที่ 2 ตัวอย่างกลุ่มข้อมูล



รูปที่ 3 เส้นไฮเปอร์เพลนแบ่งกลุ่มข้อมูล

จากรูปที่ 3 เส้นไฮเปอร์เพลนที่ใช้ในการแบ่งข้อมูลที่เกิดขึ้นมากกว่า 2 เส้น จึงต้องหาเส้นแบ่งที่เหมาะสมที่สุด คือเส้นที่ทำให้ชุดข้อมูลทั้งสองกลุ่มมีระยะห่างกันมากที่สุด (Maximum Margin Hyper plane: MMH) กำหนดให้ชุดข้อมูลเรียนรู้ $D = \{\vec{x}_i, y_i\}$ โดยที่ $\vec{x}_i(w_{i1}, w_{i2}, w_{i3}, ..w_{im})$ เป็นข้อมูลเวกเตอร์ตัวแทนของข้อความ และแต่ละ $\vec{x}_i$ ถูกกำหนดคลาสไว้ด้วยคลาส เมื่อ เป็นค่าจำนวนจริงตั้งแต่ -1 ถึง +1 ดังสมการที่ (6)

$$y = \begin{Bmatrix} +1, \vec{w} * \vec{x} + b > 0 \\ -1, \vec{w} * \vec{x} + b > 0 \end{Bmatrix} \quad (6)$$

โดย $\vec{w}$ คือ เวกเตอร์ที่ตั้งฉากกับเส้นไฮเปอร์เพลน

$\vec{x}_i$ คือ ค่าเวกเตอร์ข้อมูล

b คือ ค่าโน้มเอียง (Bias)

เมื่อมีข้อมูล $\vec{x}$ เข้ามาใหม่ จะทำนายหาคลาส y จากชุดข้อมูลเรียนรู้ $(\vec{x}_i, y_i)$ ที่มีค่าใกล้เคียงที่สุด

2) นาอีฟเบย์ (Naïve Bayes)

เป็นวิธีการเรียนรู้แบบมีผู้สอนวิธีการหนึ่งที่จัดว่าง่ายและได้รับความนิยมนำมาใช้ในการจำแนกความคิดเห็นที่มีพื้นฐานมาจากทฤษฎีความน่าจะเป็นของเบย์ (Bayes theorem) [26] หรือกฎของเบย์ ตั้งชื่อตามโทมัส เบย์ (Tomas Bayes) นักสถิติและนักปราชญ์ชาวอังกฤษ ซึ่งกล่าวถึงความสัมพันธ์ระหว่างเหตุการณ์ในปัจจุบันและสิ่งที่เกิดก่อนหน้า โดยมีความน่าจะเป็นจะเป็นแบบมีเงื่อนไขเป็นประเด็นสำคัญในทฤษฎีนี้ ความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) หมายถึงความน่าจะเป็นของการเกิดเหตุการณ์ A เมื่อกำหนดว่าเหตุการณ์ B เกิดขึ้นแล้ว แทนด้วยสัญลักษณ์ $P(A \mid B)$ ซึ่งสามารถคำนวณได้จากความน่าจะเป็นร่วม (Joint Probability) แทนด้วยสัญลักษณ์ $P(A \wedge B)$ หรือความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B เกิดขึ้นร่วมกันดังสมการที่ (7)

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad (7)$$

การจำแนกประเภทด้วยทฤษฎีของเบย์ [26] เพื่อใช้ในการจำแนกประเภท (Classification) จากกฎของเบย์สามารถกำหนดได้ดังสมการที่ (8)

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \quad (8)$$

โดยที่ $P(D)$ คือ ความน่าจะเป็นก่อนของเซตตัวอย่างฝึกฝน D หรือเรียกว่า “Evidence”

$P(h)$ คือ ความน่าจะเป็นก่อน ของสมมติฐาน $h \in H$ หรือเรียกว่า “Prior”

$P(h \mid D)$ คือ ความน่าจะเป็นภายหลังของสมมติฐาน h เมื่อกำหนดเซตตัวอย่างฝึกฝน D หรือเรียกว่า “Posterior”

$P(D \mid h)$ คือ ความน่าจะเป็นภายหลังของเซตตัวอย่างฝึกฝน D เมื่อกำหนดสมมติฐาน h หรือเรียกว่า “Likelihood”

3) เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor: K-NN)

วิธีการจำแนกข้อมูลด้วย K-Nearest Neighbor หรือ เพื่อนบ้านที่ใกล้ที่สุด [27, 28] ถูกนำมาใช้ในการจัดกลุ่มข้อมูลที่อยู่ใกล้กันหรือเป็นกลุ่มข้อมูลเดียวกัน ซึ่งจะมีการกำหนดค่า K เป็นจำนวนกลุ่ม ซึ่งสามารถคำนวณหาระยะห่างระหว่างข้อมูลในแต่ละกลุ่มด้วยระยะห่างจากสมการที่ (9)

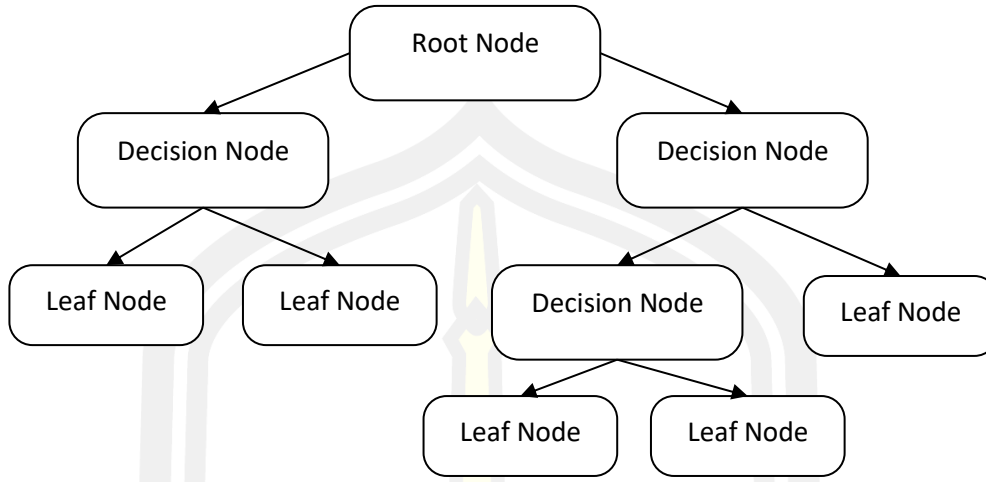
$$D_{Euclidian}(X_i, Y_i) = \sqrt{\sum_{k=1}^{n}(X_{i,k} - X_{j,k})^2} \quad (9)$$

โดยที่ $D_{Euclidian}(X_i, Y_i)$ คือ ระยะห่างระหว่างข้อมูล X_i และ Y_i

k คือ คุณลักษณะทั้งหมดของตัวอย่าง

4) ต้นไม้ตัดสินใจ (Decision tree)

วิธีการจำแนกข้อมูลด้วย Decision tree หรือ ต้นไม้ตัดสินใจ [29] เป็นเทคนิคเรียนรู้ในการจำแนกข้อมูลในลักษณะโครงสร้างแบบต้นไม้ เพื่อแสดงเส้นทางในการตัดสินใจที่เป็นไปได้และผลลัพธ์ของแต่ละเส้นทาง โครงสร้างของต้นไม้ตัดสินใจดังแสดงในรูปที่ 4



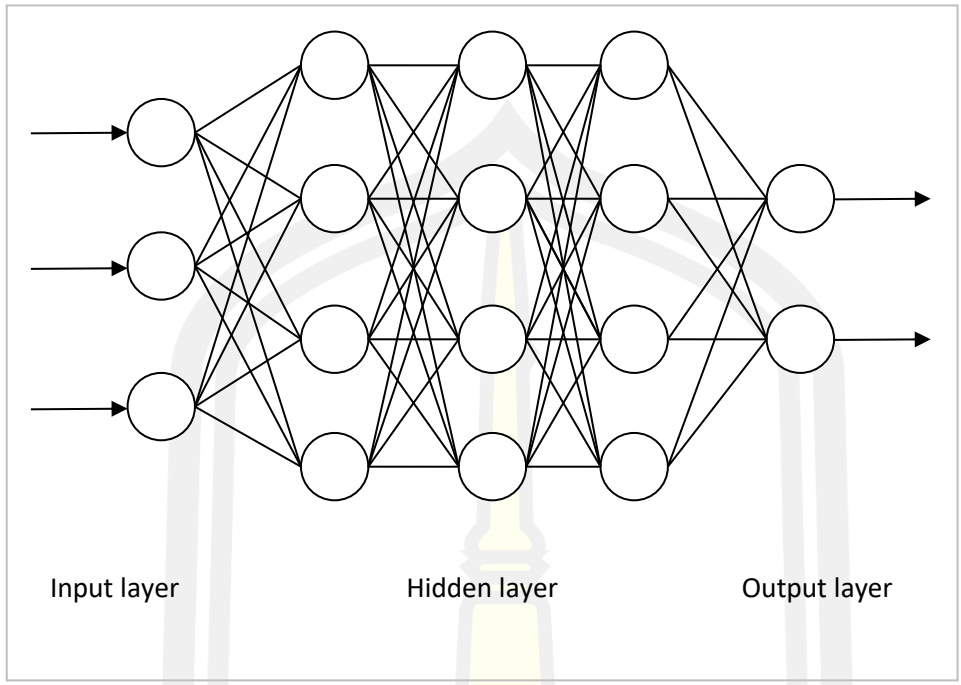
รูปที่ 4 โครงสร้างสำหรับการตัดสินใจของต้นไม้ตัดสินใจ

2.4 เทคนิคการเรียนรู้เชิงลึก

เทคนิคการเรียนรู้เชิงลึก (Deep learning) [6, 20, 30] เป็นวิธีการหนึ่งของการเรียนรู้ของเครื่อง (Machine Learning) [31] เป็นการพัฒนาเพื่อทำให้เครื่องจักรหรือคอมพิวเตอร์สามารถเลียนแบบวิธีการทำงานของโครงข่ายประสาท (Neurons) ที่เปรียบเสมือนสมองของมนุษย์ที่เรียกว่าโครงข่ายประสาทเทียม (Neural Network: NN) โดย Deep Learning ถูกสร้างขึ้นจากการนำเอา NN หลายชั้น (Layer) มาใช้ในการวิเคราะห์มากกว่า 2 ชั้น

2.4.1 Deep Neural Network

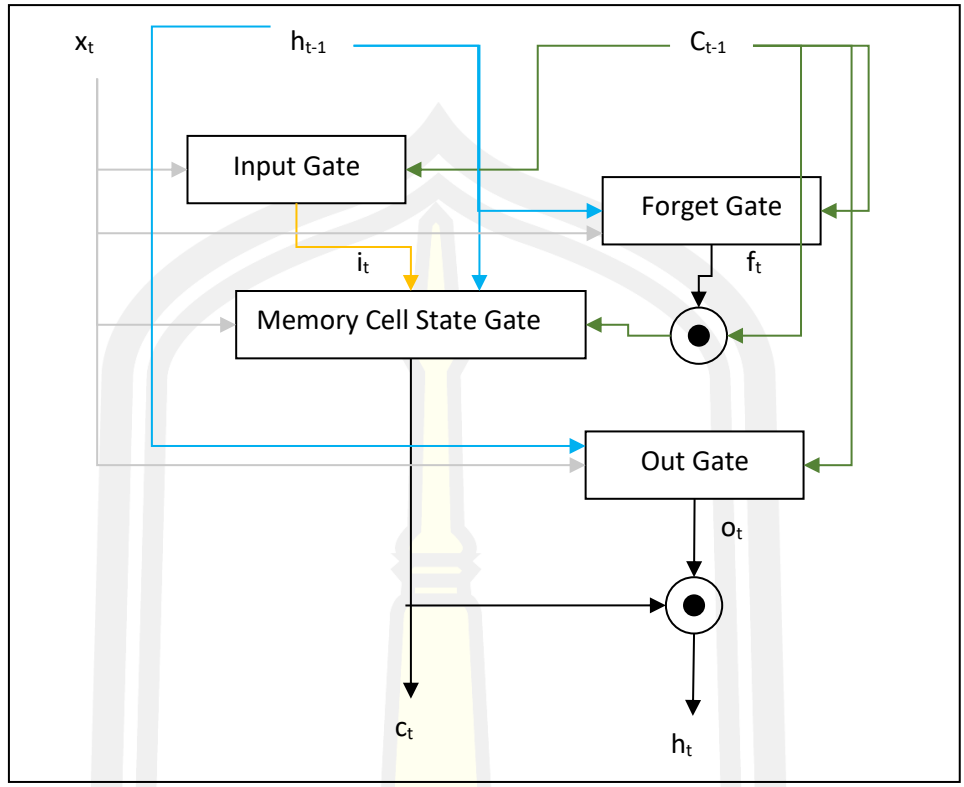
เทคนิคการเรียนรู้โครงข่ายประสาทเชิงลึก (Deep Neural Network: DNN) [19, 32] เป็นวิธีการนำอัลกอริทึมการเรียนรู้เชิงลึกไปใช้งานในรูปแบบที่ง่ายที่สุด โดยเป็นการต่อยอดมาจากโครงข่ายประสาทเทียมด้วยการเพิ่มจำนวนชั้นเข้าไป ทำให้มีตัวแปรที่ใช้ในการแยกคุณลักษณะที่มากขึ้น เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์และการจำแนก ซึ่งการทำงานของโครงสร้างนั้นมีลักษณะเช่นเดียวกันกับโครงข่ายประสาทเทียมดังแสดงในรูปที่ 5



รูปที่ 5 ตัวอย่างโครงสร้างอัลกอริทึม Deep Neural Network

2.4.2 Long Short-Term Memory

เทคนิคการเรียนรู้หน่วยความจำระยะยาว-ระยะสั้น (Long Short-Term Memory: LSTM) [19, 33, 34] เป็นอัลกอริทึมที่ถูกพัฒนาต่อยอดมาจาก Recurrent Neural Network (RNN) [35] ซึ่งสามารถใช้งานได้ดีกับข้อมูลที่มีความต่อเนื่องกัน เช่น ข้อความ ข้อมูลเสียง รูปภาพ หรือ รูปแบบอนุกรมเวลา การปรับปรุงจาก RNN ผู้สร้าง LSTM ได้นำ RNN มาปรับปรุงโดยใช้วิธีการเพิ่มฟังก์ชันจาก tanh และประกอบกับประตูที่มี 3 ประตู ได้แก่ ประตูทางเข้า (Input Gate) ประตูลืม (Forgot Gate) และ ประตูทางออก (Output Gate) พร้อมกับสถานะหน่วยอับเดต (Update Cell State) ดังแสดงในรูปที่ 6



รูปที่ 6 ตัวอย่างโครงสร้างโครงข่ายประสาทแบบ LSTM

ที่มา [19]

Input Gate เป็นหน่วยที่ใช้ในการกำหนดว่าข้อมูลที่นำเข้ามาวิเคราะห์ในเซลล์โดยพิจารณาสถานะข้อมูลที่ซ่อนอยู่จากขั้นตอนก่อนหน้าประตูเข้าเพื่อกำหนดว่าข้อมูลที่ป้อนมีค่าควรเก็บไว้หรือไม่ ดังแสดงในสมการที่ (10) [19, 36]

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{10}$$

เมื่อ i_t แทนผลลัพธ์ที่ได้จาก Input Gate

σ แทนฟังก์ชัน Sigmoid

W_{xi} แทนค่าน้ำหนักสำหรับคำนวณ Input ใน Input Gate

X_t แทนค่า Input ที่นำเข้ามาคำนวณ

W_{hi} แทนค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Input Gate

h_{t-1} แทนค่า Hidden State ที่ได้มาจากการคำนวณในหน่วยเวลาก่อนหน้า

W_{ci} แทนค่าน้ำหนักสำหรับคำนวณ Memory Cell State ใน Input Gate

c_{t-1} แทนค่า Memory Cell State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

b_i แทนค่า Bias ที่ใช้ในการคำนวณใน Input Gate

Forget Gate เป็นหน่วยที่นำมาใช้ในการกำหนดข้อมูลที่จะนำมาทำการวิเคราะห์ใน Cell โดยทำการกำหนดว่าข้อมูลนั้นควรจะถูกบันทึกหรือถูกลืม โดยสามารถกำหนดได้จากสมการที่ (11) [19, 36]

$$f_t = \sigma\ W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f \tag{11}$$

เมื่อ f_t แทนผลลัพธ์ที่ได้จาก Forget Gate

σ แทนฟังก์ชัน Sigmoid

W_{xf} แทนค่าน้ำหนักสำหรับคำนวณ Input ใน Forget Gate

X_t แทนค่า Input ที่นำเข้ามาคำนวณ

W_{hf} แทนค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Forget Gate

h_{t-1} แทนค่า Hidden State ที่ได้มาจากการคำนวณในหน่วยเวลาก่อนหน้า

W_{cf} แทนค่าน้ำหนักสำหรับคำนวณ Memory Cell State ใน Input Gate

c_{t-1} แทนค่า Memory Cell State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

b_f แทนค่า Bias ที่ใช้ในการคำนวณใน Forget Gate

Memory Cell Gate เป็นหน่วยที่นำมาใช้ในการกำหนดข้อมูลที่จะนำมาทำการวิเคราะห์ใน Cell และทำการคำนวณค่าสถานะ เพื่อใช้ในการคำนวณค่าในครั้งถัดไป โดยสามารถกำหนดได้จากสมการที่ (12) [19, 36]

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh\ W_{xc}x_t + W_{hc}h_{t-1} + b_c \tag{12}$$

เมื่อ c_t แทนผลลัพธ์ที่ได้จาก Memory Cell Gate

f_t แทนผลลัพธ์ที่ได้จาก Forget Gate

c_{t-1} แทนค่า Memory Cell Gate จากหน่วยเวลาก่อนหน้า

i_t แทนค่าผลลัพธ์ที่ได้จาก Input Gate

$\tanh$ แทนฟังก์ชัน Hyperbolic tangent

W_{xc} แทนค่าน้ำหนักสำหรับคำนวณค่า Input จาก Memory Cell State Gate

x_t แทนค่า Input ที่นำเข้ามาคำนวณ

W_{hc} แทนค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Memory Cell State Gate

h_{t-1} แทนค่า Hidden State ที่ได้มาจากการคำนวณในหน่วยเวลาก่อนหน้า

b_c แทนค่า Bias ที่ได้มาจากการคำนวณใน Forget Gate

Output Gate เป็นหน่วยที่นำมาใช้ในการคำนวณ Output ของ Cell ซึ่งผลลัพธ์ที่ได้จาก Cell นี้จะมีอยู่ 2 อย่าง ได้แก่ Output และ Hidden State สำหรับใช้ในการคำนวณครั้งถัดไป โดยสามารถกำหนดได้จากสมการที่ (13) และสมการที่ (14) ตามลำดับ [19, 36]

$$o_t = \sigma\ W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o \tag{13}$$

$$h_t = o_t \cdot \tanh\ c_t \tag{14}$$

เมื่อ o_t แทนผลลัพธ์ที่ได้จาก Output Gate

σ แทนฟังก์ชัน Sigmoid

W_{xo} แทนค่าน้ำหนักสำหรับคำนวณ Input ใน Output Gate

x_t แทนค่า Input ที่นำเข้ามาคำนวณ

W_{ho} แทนค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Output Gate

h_{t-1} แทนค่า Hidden State ที่ได้มาจากการคำนวณในหน่วยเวลาก่อนหน้า

W_{co} แทนค่าน้ำหนักสำหรับคำนวณ Memory Cell State ใน Output Gate

c_{t-1} แทนค่า Memory Cell State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

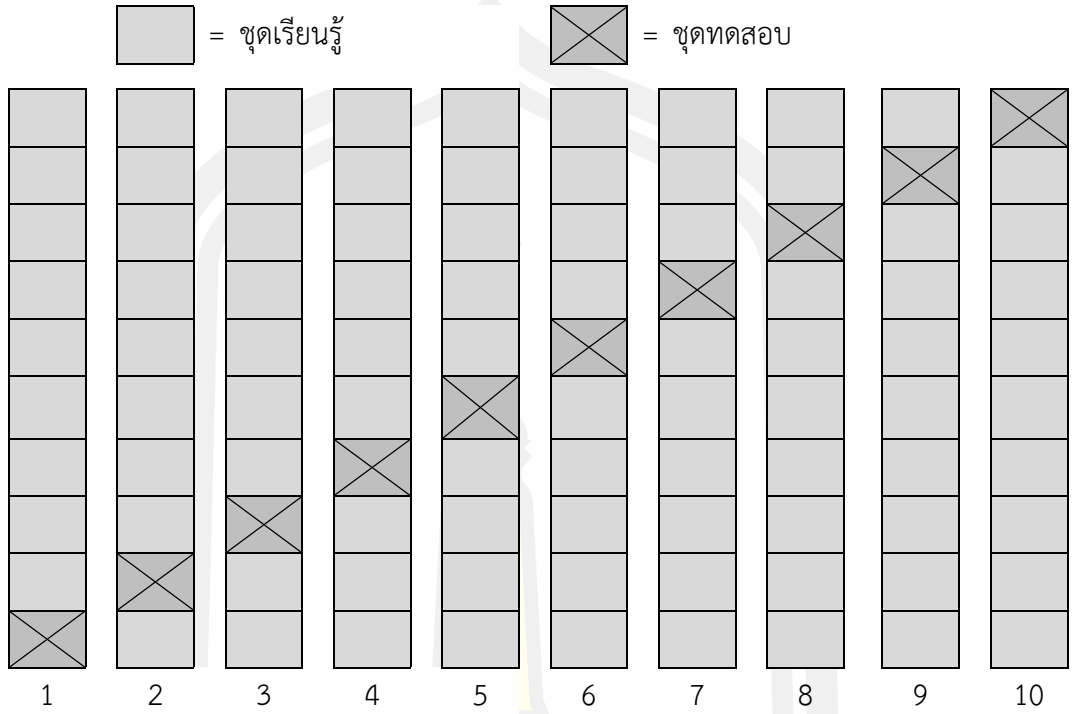
b_o แทนค่า Bias ที่ใช้ในการคำนวณใน Output Gate

h_t แทนค่า Hidden State จากการคำนวณ

2.5 การวัดประสิทธิภาพการจำแนก

ในการวัดประสิทธิภาพการจำแนกนั้นโดยทั่วไปจะใช้วิธีการวัดประสิทธิภาพ ได้แก่ [7, 17] การวัดค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure) และค่าความถูกต้อง (Accuracy) ของผลลัพธ์ที่ได้จากแต่ละอัลกอริทึม ใช้วิธีการ k-fold cross validation ในการแบ่งชุดข้อมูลเรียนรู้ (training set) และชุดข้อมูลทดสอบ (testing set) โดยแบ่งข้อมูลออกเป็น 10 ชุดข้อมูลเท่า ๆ กัน (k = 10) และทำการประเมินประสิทธิภาพของตัวจำแนกในแต่ละรอบ โดยรอบแรกชุดข้อมูลที่ 2 ถึง 10 ถูกนำไปเป็นชุดข้อมูลเรียนรู้เพื่อสร้างตัวจำแนก ชุดข้อมูลที่ 1 เป็นชุดข้อมูลทดสอบเพื่อวัดประสิทธิภาพของตัวจำแนก รอบที่สองชุดข้อมูลที่ 2 เป็นชุดข้อมูลทดสอบ การแบ่งข้อมูลด้วยวิธีการ Cross-validation Test ดังแสดงในรูปที่ 7 และชุดข้อมูลที่เหลือเป็นชุดข้อมูลเรียนรู้ ในแต่ละรอบจะทำการวัดประสิทธิภาพด้านต่าง ๆ ซึ่งสามารถอธิบายได้โดยใช้ตาราง Confusion Matrix ซึ่งเป็นตารางในรูปแบบจัตุรัส ที่มีจำนวนคอลัมน์เท่ากับจำนวนแถวและเท่ากับจำนวนคลาส เช่น มีคลาสของชุดสอน 2 คลาส คือ C1 และ C2 โดยคลาสที่

เป็นการทำนาย (Predicted) อยู่ด้านคอลัมน์ และคลาสที่เป็นชุดสอน (Actual) อยู่ด้านแถวดังตารางที่ 4



รูปที่ 7 ตัวอย่างขั้นตอนการทำงานของ 10-Fold Cross Validation

ตารางที่ 4 ตารางแสดง Confusion Matrix

		Predicted	
		C1	C2
Actual	C1	*a*	*b*
	C2	*c*	*d*

เมื่อ *a* คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส C1

b คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส C2 แต่คำตอบคือ C1

c คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส C1 แต่คำตอบคือ C2

d คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส C2

1) การวัดค่าความถูกต้อง (Accuracy)

การวัดค่าความถูกต้องในการจำแนกโดยรวมเป็นการคำนวณจากผลรวมของค่าที่ทำนายถูกต้องว่าเป็นคลาสที่ต้องการพิจารณา หารด้วยผลรวมของจำนวนทั้งหมด ซึ่งสามารถคำนวณได้ดังสมการที่ (15)

$$Accuracy = \frac{(a+d)}{(a+b+c+d)} \tag{15}$$

2) การวัดค่าความแม่นยำ (Precision)

ในการวัดค่าความแม่นยำเป็นการคำนวณจากค่าที่ทำนายถูกต้องว่าเป็นคลาสที่ต้องการพิจารณา หารด้วยผลรวมของค่าที่ทำนายถูกต้องว่าเป็นคลาสที่กำลังพิจารณาและค่าที่ทำนายว่าเป็นคลาสอื่นแต่ความจริงแล้วเป็นคลาสที่กำลังพิจารณา ดังแสดงในสมการที่ (16) (17)

$$\Pr ecision_{c1} = \frac{a}{a+c} \tag{16}$$

$$\Pr ecision_{c2} = \frac{d}{b+d} \tag{17}$$

3) การวัดค่าความระลึก (Recall)

เป็นการวัดค่าความระลึก โดยจะคำนวณจากค่าที่สามารถทำนายถูกต้องว่าเป็นคลาสที่กำลังพิจารณา หารด้วยผลรวมของค่าที่ทำนายถูกต้องว่าเป็นคลาสที่กำลังพิจารณาและค่าที่ทำนายว่าเป็นคลาสที่กำลังพิจารณาแต่คำตอบเป็นคลาสอื่น ดังแสดงในสมการที่ (18) และ (19)

$$\mathrm{Re}\, call_{c1} = \frac{a}{a+b} \tag{18}$$

$$\mathrm{Re}\, call_{C2} = \frac{d}{c+d} \tag{19}$$

4) การวัดค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure)

เป็นการพิจารณานำเอาค่าความระลึกและค่าความแม่นยำมาพิจารณารวมกัน ระบบที่มีประสิทธิภาพจะต้องมีค่าความแม่นยำและค่าความระลึกสูงใกล้เคียงกัน ดังสมการที่ (20) (21)

$$F - measure_{c1} = 2 \times \frac{\mathrm{Precision}_{c1} \times \mathrm{Re}\, call_{c1}}{\mathrm{Precision}_{c1} + \mathrm{Re}\, call_{c1}} \tag{20}$$

$$F - measure_{c2} = 2 \times \frac{\mathrm{Precision}_{c2} \times \mathrm{Re}\, call_{c2}}{\mathrm{Precision}_{c2} + \mathrm{Re}\, call_{c2}} \tag{21}$$

2.6 งานวิจัยที่เกี่ยวข้อง

Razali และคณะ [6] นำเสนอวิธีการตรวจจับข้อความประชดประชันข้อความจากทวิตเตอร์ โดยใช้คุณลักษณะที่ได้จากเทคนิคการเรียนรู้เชิงลึกรวมกับคุณลักษณะจากบริบทข้อความ โดยใช้ขั้นตอนวิธีจากสถาปัตยกรรม Neural Network Convolutional (CNN) ในการสกัดคุณลักษณะเพื่อค้นหาคุณลักษณะที่ดีที่สุด และการรวมกันของคุณลักษณะที่รวมรวมจากเนื้อหาในข้อความ ผลการทดลองพิจารณาจากการวัดค่าประสิทธิภาพโดยรวมเท่ากับ 0.94 ซึ่งได้ผลดีกว่าผลการทดลองที่นำมาเปรียบเทียบคือเทคนิคการเรียนรู้ของเครื่องด้วยขั้นตอนวิธี Logistic regression

Pasupa และ Seneewong [30] นำเสนอโมเดลสำหรับการวิเคราะห์อารมณ์ความรู้สึกบนภาษาไทยด้วยวิธีการเรียนรู้เชิงลึกแบบไฮบริด โดยใช้คลังข้อมูล Thai-SenticNet5 โดยการใช้คุณลักษณะการฝังคำ การใช้บางส่วนของคำพูด คุณลักษณะที่ละเอียดอ่อน และการผสมผสานคุณลักษณะเหล่านี้เข้าด้วยกัน อีกทั้งยังใช้เทคนิคอัลกอริทึมการเรียนรู้เชิงลึก (Convolutional neural Network) และอัลกอริทึมหน่วยความจำสั้นระยะยาวแบบสองทิศทาง (Bidirectional long short term memory) และเปรียบเทียบกับชุดข้อมูลภาษาไทย 3 ชุด ได้แก่ ThaiTales, ThaiEconTwitter และ Wisesight ผลการทดลองด้วยเทคนิคการเรียนรู้เชิงแบบไฮบริดที่มีประสิทธิภาพดีที่สุดคือ BLSTM-CNN ได้ค่าคะแนนประสิทธิภาพโดยรวม (F-measure) เท่ากับ 0.7436, 0.7707 และ 0.5521 บนชุดข้อมูล ThaiTales, ThaiEconTwitter และ Wisesight ตามลำดับ จากผลการทดลองสรุปได้ว่าการผสมผสานคุณลักษณะและอัลกอริธึมการเรียนรู้เชิงลึกแบบไฮบริดสามารถปรับปรุงประสิทธิภาพโดยรวมได้

A. Onan และ M. A. Tocoglu [33] นำเสนอการวิจัยโดยมีวัตถุประสงค์เพื่อการนำเสนอกรอบการประชดประชันที่มีประสิทธิภาพบนข้อมูลโซเชียลมีเดียโดยดำเนินตามกระบวนทัศน์ของแบบจำลองภาษาธรรมชาติและโครงข่ายประสาทเทียมเชิงลึก พิจารณาคุณลักษณะตัวแทนของข้อความด้วยการฝังคำแบบถ่วงน้ำหนัง และนำเสนอการใช้เทคนิคหน่วยความจำระยะสั้นระยะยาว 3 ชั้นเลเยอร์ โมเดลที่นำเสนอให้ผลลัพธ์ด้วยความแม่นยำในการจำแนกประเภทที่ 95.30%

C. I. Eke และคณะ [37] นำเสนอเทคนิคคุณลักษณะตามบริบทสำหรับการระบุการประชดประชันโดยใช้เทคนิคการเรียนรู้เชิงลึกด้วยโมเดล BERT โดยใช้ชุดข้อมูล Twitter และ Internet Argument Corpus เวอร์ชันที่สอง (IAC-v2) สองชุด ใช้สำหรับการจัดประเภทโดยใช้โมเดลการเรียนรู้สามรูปแบบ โมเดลแรกใช้การแทนแบบฝังผ่านโมเดลการเรียนรู้เชิงลึกที่มีหน่วยความจำระยะสั้นแบบสองทิศทาง (Bi-LSTM) ซึ่งเป็นตัวแปรของ Recurrent Neural Network (RNN) โดยใช้การแสดงเวกเตอร์ GloVe เพื่อสร้างการฝังคำ (word embedding) และบริบทการเรียนรู้ โมเดลที่สองใช้การแสดงข้อมูล Encoder แบบสองทิศทางล่วงหน้าและ Transformer (BERT) ในทางตรงกันข้าม โมเดลที่สามอิงจากการผสมผสานคุณลักษณะที่ประกอบด้วยคุณลักษณะ BERT ที่เกี่ยวข้องกับความรู้สึก เกี่ยวกับการสร้างประโยค และคุณลักษณะการฝัง GloVe ด้วยการเรียนรู้ของเครื่องแบบเดิม ประสิทธิภาพของเทคนิคนี้ได้รับการทดสอบด้วยการทดลองประเมินผลต่างๆ อย่างไรก็ตาม การประเมินเทคนิคบนชุดข้อมูลเปรียบเทียบ Twitter สองชุดมีความแม่นยำสูงสุด 98.5% และ 98.0% ตามลำดับ ในทางกลับกัน ชุดข้อมูล IAC-v2 มีความแม่นยำสูงสุด 81.2% ซึ่งแสดงให้เห็นถึงความสำคัญของเทคนิคที่เสนอเหนือแนวทางพื้นฐานสำหรับการวิเคราะห์การประชดประชัน

Thititorn Seneewong และคณะ [34] ศึกษาเกี่ยวกับการวิเคราะห์อารมณ์ความรู้สึกภาษาไทยด้วยเทคนิค LSTM-CNN แบบสองทิศทางด้วยเวกเตอร์แบบฝังตัวและใช้คุณลักษณะด้วย

Sentic ซึ่งพยายามรวมคุณลักษณะเพิ่มเติมอีก 2 อย่าง ได้แก่ ส่วนหนึ่งของคำพูดและคุณลักษณะที่ละเอียดอ่อน เพื่อให้การวิเคราะห์มีความแม่นยำมากขึ้น คุณลักษณะส่วนหนึ่งของคำพูดจะระบุประเภทของคำที่สื่อถึงความรู้สึกต่างๆ ได้ดีกว่า ในขณะที่คุณลักษณะเกี่ยวกับความรู้สึกจะระบุอารมณ์ที่อยู่ภายใต้คำบางคำ การผสมผสานหน่วยความจำระยะสั้นแบบสองทิศทางและโมเดล Convolutional Neural Networks เข้ากับคุณลักษณะต่างๆ ที่กล่าวถึง เราทำการวิเคราะห์ความคิดเห็นเกี่ยวกับเรื่องราวของเด็กไทย และพบว่าการรวมกันของคุณลักษณะทั้งสามให้ผลลัพธ์ที่ดีที่สุดที่ค่า F1-score 78.89%

Rajadesingan และคณะ [38] นำเสนอวิธีการรูปแบบของพฤษติกรรมที่มีผลกระทบต่อการตรวจจับข้อความประชดประชันบนทวิตเตอร์ การเก็บรวบรวมข้อมูลจากเว็บไซต์ทวิตเตอร์โดยใช้แฮ็ชแท็ก (#sarcasm, #not) ในการรวมรวมข้อมูล ข้อมูลที่ได้ผ่านกระบวนการในการเตรียมข้อมูล โดยฟิลเตอร์ข้อมูลที่ไม่สนใจ เช่น กำจัดข้อมูลที่ไม่ใช่ภาษาอังกฤษ ไม่สนใจข้อความที่เป็นรีทวิทต์ ไม่สนใจข้อมูลที่อยู่เว็บไซต์ รูปภาพหรือวิดีโอ และไม่สนใจข้อความที่มีจำนวนคำน้อยกว่า 3 คำ การเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ ได้เลือกใช้ จำนวน 10 คุณลักษณะ ดังนี้ 1) เปอร์เซ็นต์ของไอคอนแสดงอารมณ์ในข้อความทวีต 2) เปอร์เซ็นต์ของคำคุณศัพท์ในทวีต 3) เปอร์เซ็นต์ของคำที่เป็นอดีตกับคะแนนด้านอารมณ์ความรู้สึกเท่ากับ 3 4) จำนวนของคำหลายพยางค์ในทวีต 5) ความหนาแน่นของข้อความทวีต 6) เปอร์เซ็นต์ของคำที่เป็นอดีตกับคะแนนด้านอารมณ์ความรู้สึกเท่ากับ 2 7) เปอร์เซ็นต์ของคำที่เป็นอดีตกับคะแนนด้านอารมณ์ความรู้สึกเท่ากับ –3 8) จำนวนโพสของคำประชดประชันที่ผ่านมา 9) เปอร์เซ็นต์ของการเปลี่ยนแปลงอารมณ์ความรู้สึกจากบวกไปลบที่สร้างโดยผู้ใช้งาน 10) เปอร์เซ็นต์ของแฮชแท็กที่ใช้ตัวอักษรตัวพิมพ์ใหญ่ในทวิต โดยจากคุณลักษณะข้างต้นได้นำมากำหนดกลุ่มของคุณลักษณ์ได้ดังนี้ 1) กลุ่มของคุณลักษณะด้านการแสดงออก ได้แก่ คุณลักษณะข้อ 1,2,5,10 2) กลุ่มของคุณลักษณะด้านอารมณ์ ได้แก่คุณลักษณะข้อ 3,6,7 3) กลุ่มของคุณลักษณะด้านความคุ้นเคยหรือมีความใกล้เคียง ได้แก่คุณลักษณะข้อ 8 4) กลุ่มของคุณลักษณะด้านความแตกต่างหรือความตรงกันข้าม ได้แก่คุณลักษณะข้อ 9 และ 5) กลุ่มของคุณลักษณะด้านความยุ่งยากซับซ้อน ได้แก่คุณลักษณะข้อ 4 ในส่วนของวิธีการที่ใช้ในวัดประสิทธิภาพโดยใช้เทคนิค 10-fold cross-validation ผลของการทดลองจากการวิเคราะห์กลุ่มของคุณลักษณะได้ค่าความถูกต้องโดยรวมที่ 83.46% โดยกลุ่มของคุณลักษณะที่ให้ประสิทธิภาพดีที่สุดคือ กลุ่มที่ 1 กลุ่มของคุณลักษณะด้านการแสดงออกให้ค่าความถูกต้องเท่ากับ 76.72%

Bouazizi M และ Otsuki [5] นำเสนอวิธีการแนวทางตามรูปแบบในการจำแนกคำประชดประชันบนเครือข่ายสังคมออนไลน์ทวิตเตอร์ ใช้ข้อมูลจากทวิตเตอร์จากเดือนธันวาคม 2014 ถึงมีนาคม 2015 วิธีการเก็บข้อมูลโดยดึงข้อมูลโดยใช้เอพีไอ (API) ของทวิตเตอร์ที่เป็นคำประชดประชันที่มี เครื่องหมายแฮชแท็ก #sarcasm โดยเก็บข้อมูลจำนวน 58,609 รายการ โดยหลังจากทำความ

สะอาดข้อมูลแล้วได้แบ่งข้อมูลเป็น 3 ส่วน 1) ข้อมูล 6,000 ทวิต ใช้คนสองคนที่ไม่มีประสบการณ์ด้านการทวิตในการตรวจสอบและจำแนกระดับของคำประชดประชันโดยแบ่งเป็น 6 คลาส จากระดับ 1 มีความประชดประชันน้อย ถึง ระดับ 6 มีระดับความประชดประชันมากในการให้คลาสขึ้นอยู่กับทัศนคติของผู้ที่ติดคลาส เมื่อติดคลาสแล้วจะได้ข้อความที่ประชดประชัน และไม่ประชดประชันเท่ากับ 50/50 ซึ่งใช้ข้อมูลชุดนี้เป็นชุดสอน (Training set) 2) ข้อมูลชุดนี้ประกอบด้วยข้อความประชดประชัน 1,128 ข้อความ และไม่ประชดประชัน 1,128 ข้อความ ซึ่งนำมาใช้เป็นข้อมูลชุดเพิ่มประสิทธิภาพ (Optimization set) 3) ประกอบด้วยข้อความประชดประชัน 500 ข้อความ และไม่ประชดประชัน 500 ข้อความ นำมาใช้เป็นข้อมูลชุดทดสอบ (Test set) และใช้นำมาใช้ในการประเมินผลการทดลอง การสกัดคุณลักษณะใช้วิธีดังนี้ 1) คุณลักษณะที่เกี่ยวข้องกับอารมณ์ (Sentiment-related features) เป็นวิธีการโดยใช้ฐานข้อมูล SentiStrength ในการตรวจสอบความคิดเห็นที่เป็นบวก และลบ โดยกำหนดเป็น 2 คุณลักษณะ คือ pw, nw การใช้หน้าที่ของคำในการวิเคราะห์ค่าความเป็นบวกหรือลบ โดย คำคุณศัพท์ คำกริยา และคำวิเศษณ์ มีค่านำหนักในด้านอารมณ์มากกว่าคำนาม และใช้สัญญลักษณะแสดงอารมณ์ร่วมด้วย เช่น “ :P ” 2) คุณลักษณะที่เกี่ยวข้องกับเครื่องหมายวรรคตอน (Punctuation-related features) โดยพิจารณาจากการจำนวนเครื่องหมาย เช่น จำนวนของเครื่องหมายตกใจ (Exclamation marks) เครื่องหมายคำถาม (Question marks) เครื่องหมายจุด (Dots) การพิมพ์เป็นตัวอักษรพิมพ์ใหญ่ทั้งหมด (All-capital words) เครื่องหมายคำพูด (Quotes) และจำนวนของสระ 3) คุณลักษณะประโยคและความหมาย (Syntactic and Semantic features) พิจารณาจากคำที่ไม่ปกติ (Uncommon words) จำนวนของคำที่ไม่ปกติ (Number of uncommon words) การมีคำประชดประชันในประโยค(Existence of common sarcastic expressions) จำนวนของคำหยาบคาย (rude) การแสดงอารมณ์ขัน (Laughing expressions) 4) คุณลักษณะที่เกี่ยวข้องกับรูปแบบ (Pattern-related features) เป็นการพัฒนารูปแบบที่ใช้หน้าที่ของคำในการกำหนดรูปแบบ ผลการทดลอง ทดสอบโดยใช้ K-fold cross validation ใช้การจำแนกด้วยอัลกอริทึม Random Forest, Support Vector Machine, K-NN และ Maximum Entropy การประเมินประสิทธิภาพของการทดลองในระหว่างการทดสอบให้ค่าความแม่นยำที่ 91.1% ค่าความถูกต้องที่ 83.1% และค่าความระลึกที่ 73.4%

Mukherjee และ Pradip [39] นำเสนอวิธีการจำแนกข้อความประชดประชันบนทวิตเตอร์มีจุดประสงค์ในการเพิ่มประสิทธิภาพการในการจำแนก ด้วยวิธีการเรียนรู้แบบมีผู้สอน สกัดคุณลักษณะโดยวิธี Content words, Function Words, Part of Speech tags, Part of speech n-grams, Content Words+Function Words, Function Words+Part of speech n-grams c และ Content Words+Function Words+Part of speech n-grams และใช้วิธีการจำแนกด้วย Naïve Bayesian classifier และใช้วิธีการจัดกลุ่มด้วยวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised

learning) ด้วย Fuzzy C-mean (FCM) algorithm ผลการทดลองเปรียบเทียบทั้งสองอัลกอริทึม พบว่า วิธีการจัดกลุ่มด้วย Fuzzy ให้ประสิทธิภาพน้อยกว่า Naïve Bayes สำหรับการจำแนกข้อความประชดประชัน ค่าความถูกต้องที่ 65%

Dave และ Desai [25] นำเสนอวิธีโดยการศึกษาความเข้าใจเทคนิคในการจำแนกสำหรับการตรวจจับคำพูดส่อเสียด ประชดประชัน บนข้อความ โดยมีวัตถุประสงค์โดยศึกษาเทคนิคที่แตกต่างที่เป็นไปได้สำหรับการจำแนกคำประชดประชันและเปรียบเทียบประสิทธิภาพแต่ละวิธีการทดลองกับประโยคในภาษาฮินดู การสกัดคุณลักษณะโดยใช้ TF-IDF หน้าที่ของคำ (Part of Speech) ประโยคและคำแสดงความคิดเห็น (Opinion word and Phrase) และการปฏิเสธ (Negation) เช่น not happy มีความหมายเดียวกับ sad การเลือกคุณลักษณะด้วยวิธี ค่าสารสนเทศร่วม (Point wise Mutual Information) และ สถิติไคสแคว์ (Chi-square) การจำแนกด้วยวิธีเรียนรู้ด้วยเครื่องใช้อัลกอริทึมที่ใช้ในการจำแนก Naïve Bayes, Maximum Entropy, Support Vector Machine และ Conditional Random Field การทดลองด้วยข้อมูลที่เป็นบวก 150 ข้อความ และข้อมูลที่เป็นลบ 150 ข้อความ เพื่อใช้เป็นชุดสอน โดยใช้ Support Vector Machine ด้วยใช้การตรวจสอบด้วย 10X Validation ด้วยคุณลักษณะถุงคำ (Bag of words) และวัดคุณลักษณะด้วย TF-IDF ใช้ข้อมูลสำหรับเป็นชุดทดสอบจำนวน 25 ข้อความ เพื่อใช้ทดสอบโมเดล ผลการทดลองได้ค่าความถูกต้องเท่ากับ 50%

Santosh และคณะ [7] นำเสนอวิธีการแยกการรับรู้ความรู้สึกประชดประชันบนทวิตเตอร์ วิธีการที่นำเสนอมีสองวิธีคือ 1) อัลกอริทึมการวิเคราะห์ตามพจนานุกรม (Parsing-Based lexicon generation algorithm: PBLGA) 2) การจำแนกตามการปรากฏขึ้นของคำอุทาน (Occurrence of the interjection word) การรวมกันของทั้งสองวิธีเมื่อเปรียบเทียบกับศาสตร์ของการจำแนกคำประชดประชัน แสดงให้เห็นว่า วิธีแรกให้ค่าความแม่นยำที่ 89% ค่าระลึกที่ 96% และค่าเฉลี่ยประสิทธิภาพโดยรวมที่ 84% วิธีที่สอง ให้ค่าความแม่นยำที่ 85% ค่าระลึกที่ 96% และค่าเฉลี่ยประสิทธิภาพโดยรวมที่ 90% ในข้อความทวิตที่ใช้แฮชแท็ก (#sarcasm)

Archana และ Chitrakala [40] ได้นำเสนอวิธีการในการจัดการข้อความประชดประชันในการคำนวนระดับ ความรู้สึกบนทวิตเตอร์โดยใช้วิธีบิ๊กเดต้า โดยข้อมูลที่นำมาใช้ในการทดลองมาจากทวิตเตอร์การรวบรวมข้อมูลโดยการค้นหาโดยใช้แฮ็ชแท็ก #sarcasm และ #sarcastic เป็นข้อมูลที่ทวิตเกี่ยวกับการเลือกตั้งในสหรัฐในช่วงวันที่ 1 สิงหาคม 2016 ถึง 31 สิงหาคม 2016 ใช้วิธีเก็บข้อมูลด้วย TwitterR API จำนวน 1,150,000 ทวิต และใช้ Hadoop Map-reduce ช่วยในการจัดการข้อมูลขนาดใหญ่ ซึ่งผลการทดลองในการวัดประสิทธิภาพของการใช้ Map-reduce ให้ความเร็วในการประมวลผลที่เร็วว่าการไม่ใช้ โดยวัดจากค่า ความแม่นยำ, ค่าระลึก และการวัด

ประสิทธิภาพโดยรวม จากการใช้ Map-reduce มีค่าเท่ากับ 0.714, 0.51 และ 0.586 ตามลำดับ การใช้วิธีพื้นฐานมีค่าเท่ากับ 0.318, 0.296 และ 0.306 ตามลำดับ

Satoshi และ Kazutaka [41] นำเสนอวิธีการในการสกัดรูปแบบของคำประชดประชันเพื่อประเมินผลการแสดงอารมณ์ ข้อมูลที่นำมาใช้ในการทดลองเป็นข้อมูลเกี่ยวกับการรีวิวสินค้าจาก Rakuten Ichba ประเทศญี่ปุ่น จำนวน 10,000 รีวิว ซึ่งมีจำนวนประโยคทั้งหมด 34,917 ประโยค วิธีการที่นำเสนอ คือการจำแนกข้อความประชดประชันออกเป็น 8 คลาส และทำการตรวจสอบประสิทธิภาพในการจำแนกตามกฎของ 8 คลาส ผลการทดลองวิธีที่นำเสนอให้ประสิทธิภาพดีกว่าวิธีพื้นฐานโดยว่าจากค่าความแม่นยำ, ค่าระลึก และค่าประสิทธิภาพโดยรวม ดังนี้ วิธีพื้นฐานเท่ากับ 0.006, 0.414 และ 0.012 วิธีที่นำเสนอ 0.028, 0.543 และ 0.053 ตามลำดับ และผลการประเมินประสิทธิภาพจากค่าความแม่นยำของการทดลองจากข้อมูลชุดทดสอบ วิธีที่นำเสนอให้ประสิทธิภาพดีกว่าวิธีพื้นฐาน เท่ากับ 0.036 และ 0.009

Salas และคณะ [42] นำเสนอวิธีการตรวจจับข้อความประชดประชันอัตโนมัติบนทวิตเตอร์ด้วยวิธีการทางจิตวิทยา การรวบรวมข้อมูลจากทวิตเตอร์โดยใช้ Twitter4J โดยรวมรวมสองส่วนคือจากสเปนและแม็กซิโก แยกเป็นข้อความประชดประชันอย่างละ 5,000 ทวิต และ ข้อความทั่วไปที่ไม่ใช่ข้อความประชดประชันอย่างละ 5,000 ข้อความ เมื่อนำมาผ่านกระบวนการในการเตรียมข้อมูลแล้วจะได้ข้อมูล ที่เป็นคำประชดประกับจำนวน 5,000 ข้อความ (สเปน 2,500, แมกซิโก 2,500) และไม่ประชดประชัน 5,000 ข้อความ (สเปน 2,500, แมกซิโก 2,500) อัลกอริทึมที่ใช้คือ Sequential Minimal Optimization (SMO), J48 Decision tree และ Bayes Net ผลการการทดลองโดยการวัดค่าความถูกต้อง ความความแม่นยำ ค่าความระลึก ค่าความถูกต้องโดยรวม เท่ากับ SMO คือ 85.50%, J48 คือ 75.20% และ BayesNet คือ 75.60%, 75.70%, 75.60%, 75.60% ตามลำดับ

Bouazizi และ Ohtsuki [43] นำเสนอการทำเหมืองความคิดเห็นบนทวิตเตอร์เพื่อการเพิ่มประสิทธิภาพการวิเคราะห์อารมณ์ความรู้สึกในข้อความประชดประชัน หัวข้อที่เก็บข้อมูล คือ การเมือง รีวิวโทรศัพท์ กีฬา รีวิวหนัง และ สินค้าอิเล็กทรอนิค โดยข้อมูลที่ใช้เป็นชุดเทรนจำนวน 20,000 ทวิตและ 1,000 ทวิตใช้เป็นชุดทดสอบ อัลกอริทึมที่ใช้ คือ Support Vector Machine Maximum Entropy และ Naïve Bayes ผลการทดลองสรุปได้ว่า Support Vector Machine ให้ประสิทธิภาพดีกว่า Maximum Entropy และ Naïve Bayes ผลจากการทดลองโดยการวัดค่า Recall ของการจำแนกซึ่งเปรียบเทียบกับวิธีการพื้นฐาน ได้ดังนี้ ผลของค่า Recall วิธีพื้นฐาน Naïve Bayes 83.9%, SVM 85.7% และ Maximum Entropy 82.3% ผลของวิธีการที่นำเสนอ Naïve Bayes 85.9%, SVM 92.0% และ Maximum Entropy 83.8% ซึ่งให้ผลดีกว่าวิธีแบบพื้นฐาน

Miljana และคณะ [44] นำเสนอวิธีการในการใช้คลังคำในการจำแนกข้อความเสียดสีและประชดประชัน ข้อมูลที่ใช้เป็นภาษาเซอเบียร์ จากทวิตเตอร์ โดยรวบรวมข้อมูลจากคำค้นโดยใช้แฮช

แท็ก #irony, #sarcasm, #not, #yeahright ในกระบวนการนำข้อมูลมาวิเคราะห์โดยการนำแฮชแท็กออก คลังคำที่นำมาใช้ในการจำแนกจาก Serbian WordNet ซึ่งเป็นพจนานุกรมที่บรรจุคำจำนวน 4,593 ข้อความ ซึ่งเป็นข้อความที่มีขั้วความรู้สึก และ 62 วลี ข้อมูลที่ใช้ในการทำมาทดลองจำนวน 2,127 รายการ ประสิทธิภาพได้ค่าความแม่นยำเท่ากับ 68.6% และค่าความถูกต้องเท่ากับ 86.1%

Thu และ New [45] นำเสนอวิธีการใช้คุณลักษณะทางอารมณ์ในการตรวจจับข้อความเสียดสีประชดประชัน ประเภทของชุดข้อมูลที่ใช้มีสอบประเภทคือข้อความประชดและไม่ประชด เกี่ยวกับบทความข่าวจำนวน 226 ข้อความ และ 674 ข้อความ, รีวิวสินค้าอเมซอน จำนวน 1,000 ข้อความ และข้อมูลจากทวิตเตอร์ จำนวน 22,126 ข้อความ และ 32,681 ข้อความ วิธีการที่ใช้ในการสะกัดคุณลักษณะ 1) word-based ใช้โมเดล N-grams โดยเลือกคำที่พบบ่อยที่สุด 1,000 คำ 2) Emotion ใช้การแสดงอารมณ์ 8 ตัว คือ ความโกรธ ความคาดหมาย รังเกียจ ความกลัว ความสุข ความเศร้า แปลกใจ และความไว้วางใจ 3) Sentiment ใช้ขั้วแสดงอารมณ์ บวก ลบ เป็นกลาง และ ค่าคะแนนความเชื่อมั่น 4) Bog-of Sorted Emotion (BOSE) เป็นรวบรวมระหว่างการแสดงอารมณ์ 8 ตัวและขั้วความรู้สึก ในการเรียงลำดับ 5) BOSE-TFIDF การทำวิธีการที่ 4 มาให้ค่าน้ำหนักด้วย TFIDF 6) BOSE-TFRF เป็นการคำนวณแบบจำลองโดยใช้ความถี่และแล้วความเกี่ยวข้องกันของความถี่ 7) SenticNet คือฐานข้อมูล จำนวน 13,000 คำที่มี 4 มิติทางอารมณ์ คือ ความรู้สึกไว, ไหวพริบ, ความสนใจ และ ความร่าเริง ในการทดสอบประสิทธิภาพของแบบจำลองด้วย 10-Fold cross-validation อัลกอริทึมที่ใช้ Support Vector Machine และ Ensemble โดยสรุปแล้วลักษณะอารมณ์ที่คลุมเคลือการจำแนกด้วย SVM ให้ประสิทธิภาพได้ไม่ได้ในวิธีการคุณลักษณะพื้นฐานอารมณ์ (Emotion-based features) ในขณะเดียวกันการจำแนกด้วยตัวจำแนก Ensemble Classifier ให้ประสิทธิภาพ ค่าความถูกต้องที่ดีขึ้นในข้อความสั้นและยาว

Reganti และคณะ [46] นำเสนอแบบจำลองเสียดสีในข้อความภาษาอังกฤษสำหรับการตรวจจับอัตโนมัติ ด้วยวิธีการคุณลักษณะคลังคำเพื่อเปรียบเทียบกับวิธีพื้นฐาน (n-grams) โดยวิธีที่ใช้ในการเลือกคุณลักษณะ คือ Lexical Feature ซึ่งเป็นคลังคำทางอารมณ์ที่บรรจุคำจำนวน 14,000 คำ ซึ่งแต่ละข้อความได้ระบุความรู้สึกที่เป็น Positive และ Negative และ 8 สัญลักษณ์ทางอารมณ์ คือ ความโกรธ ความคาดหมาย รังเกียจ ความกลัว ความสุข ความเศร้า แปลกใจ และความไว้วางใจ และใช้ SenticNet ซึ่งเป็นคลังคำความรู้สึกที่ใหญ่แต่ละแนวคิดของอารมณ์ถูงระบุขั้วจำนวนของขั้วบวก ลบ และเป็นกลาง โดยจะถูกนำมาใช้เป็นคุณลักษณะ ชุดข้อมูลที่ใช้คือ 1) รีวิวสินค้าอเมซอน จำนวน 1,254 ข้อความ แบ่งเป็นข้อมูลประชดจำนวน 437 ข้อความ และไม่ประชดจำนวน 817 ข้อความ 2) เอกสารข่าว จำนวน 4,000 ข้อความ และจำนวน 233 ข้อความ ที่เป็นบทความคำประชด 3) ทวิตเตอร์โพสต์ จำนวน 3,000 ข้อความประชด จากการใช้คำค้น #satire, #irony และ

#sarcasm อัลกอริทึมที่ใช้ในการจำแนกคือ Logistic Regression(LR), Random Forest(RF), Support Vector Machine(SVM), Decision Tree(DT) และ Ensemble ผลการจำแนกประสิทธิภาพโดยรวมแต่ละชุดข้อมูลทุกคุณลักษณะ 1) รีวิวสินค้าอเมซอน 75.30%, 68.93%, 66.63%, 67.22% และ 77.96% 2) เอกสารข่าว 75.88%, 63.89%, 69.34%, 63.22% และ 79.02% 3) ทวิตเตอร์โพสต์ 76.89%, 71.06%, 74.03%, 68.11% และ 78.16% สรุปได้ว่าอัลกอริทึมที่ให้ประสิทธิภาพในการจำแนกดีที่สุดคือ Ensemble เท่ากับ 77.96%, 79.02% และ 78.16% จากชุดข้อมูลทั้งสามชุดตามลำดับ

Suhaimin และคณะ [47] นำเสนอวิธีการตรวจจับข้อความประชดประชันด้วยคุณลักษณะการประมวลผลภาษาธรรมชาติในข้อความสองภาษาบนเครือข่ายสังคมออนไลน์ ชุดข้อมูลที่นำมาใช้ในการวิจัยเป็นหัวข้อความคิดเห็นเกี่ยวกับข่าวด้านเศรษฐกิจจากเฟสบุ๊คแฟนเพ็จสาธารณะ จำนวน 3,000 รายการ แยกเป็นข้อความประชด 969 ความคิดเห็น และ 1,001 ข้อความคิดเห็นที่ไม่ใช่ความคิดเห็นประชดประชัน วิธีการที่ใช้ในการสกัดคุณลักษณะในรูปแบบการแปลสองภาษาตาม 5 กลุ่มของการประมวลผลภาษาธรรมชาติ คือ Lexical, Pragmatic, Prosodic, Syntactic และ Idiosyncratic จำแนกโดยใช้อังกอริทึม Non-linear Support Vector Machine จากการทดลองให้ประสิทธิภาพโดยรวมที่ดีกว่าวิธีพื้นฐานโดยค่าประสิทธิภาพโดยรวมที่ 85.2%

Jasso และ Meza [4] นำเสนอวิธีการตรวจจับคำเสียดสีรูปแบบข้อความสั้นในภาษาสเปนด้วยวิธี 1) รูปแบบคำ (Word Based) ใช้วิธี word N-grams และ Word2Vec 2) รูปแบบตัวอักษร (Character Based) ใช้วิธี Character N-grams โดยพิจารณาลักษณะเครื่องหมายวรรคตอน และไอคอนแสดงอารมณ์ร่วมด้วย ชุดข้อมูลที่ใช้เป็นคำเสียดสีและไม่ใช่คำเสียดสี แบ่งเป็น 3 กลุ่ม 1) รายการชุดเรียนรู้ที่มีความสมดุล (14,511, 14,511) ชุดทดสอบ (1,483, 1,483) 2) รายการชุดเรียนรู้ที่ไม่มีความสมดุล (14,511, 33,859) ชุดทดสอบ (1,483, 3,458) และ 3) รายการชุดเรียนรู้ที่นำเสนอ (14,511, 130,599) ชุดทดสอบ (1,483, 13,347) ผลการจำแนกด้วยอังกอริทึม Support vector machine (SVM) และ Random Forest (RF) โดยการวัดประสิทธิภาโดยรวม (F-score) ของทั้งสองวิธีกับข้อมูลทั้ง 3 ชุด ข้อมูลชุดที่ 1 ให้ผลดังนี้ 1) รูปแบบคำ word-gram ผลคือ (RF = 68%, SVM = 67%) word2vec ผลคือ (RF = 76%, SVM = 78%) 2) รูปแบบตัวอักษร char-gram ผลคือ (RF = 87%, SVM = 86%) ข้อมูลชุดที่ 2 ให้ผลดังนี้ 1) รูปแบบคำ word-gram ผลคือ (RF = 48%, SVM = 37%) word2vec ผลคือ (RF = 38%, SVM = 61%) 2) รูปแบบตัวอักษร char-gram ผลคือ (RF = 80%, SVM = 80%) ข้อมูลชุดที่ 3 ใช้วิธีรูปแบบตัวอักษร char-gram ผลคือ (RF = 87%, SVM = 86%)

Justo และคณะ [48] นำเสนอวิธีการในการสกัดความรู้ที่เกี่ยวข้องสำหรับการตรวจจับข้อความประชดประชันและข้อความหยาบคายบนเว็บเครือข่ายสังคมนำเสนอชุดของการเรียนรู้แบบมี

ผู้สอนในการทดลองตรวจจับข้อความประชดประชันและความหยาบคายในกล่องโต้ตอบออนไลน์ (Onling Dialog) เพื่อเปรียบเทียบช่วงของชุดคุณลักษณะที่พัฒนาโดยใช้เกณฑ์ที่แตกต่างกัน โดยนำเสนอวิธีการดังนี้ วิธีการค้นหาการตั้งค่าที่เหมาะสมของคุณลักษณะสำหรับประเภทที่แตกต่างกันของภาษาในสังคม วิธีการจัดหมวดหมู่อย่างชัดเจนว่าพิจารณาความเป็นไปในรูปแบบที่แตกต่างกันของการเสียดสี เช่น พูดและประชด วิธีการใช้คุณลักษณะที่ชัดเจนในการพิจารณาความหมายและปัญหาที่เกิดจากความยาวของคำพูดที่รวมทั้งเป้าหมายประเภทของประโยค เช่น ประโยคประชดประชัน รวมทั้งไม่ได้อยู่ในหมวดหมู่ของเป้าหมาย เช่น ไม่ได้ประชด วิธีการเปรียบเทียวความสะดวกในการตรวจสอบข้อความทั้งสองประเภท ประชดและหยาบคาย ในบริบทที่เหมือนกัน วิธีการที่เลือกใช้ในการสกัดคุณลักษณะ 1) Machanical Turk Cues เป็นการพิจารณาคุณสมบัติที่ประกอบด้วยตัวชี้นำที่ระบุว่าเป็นการวัดการประชดประชันหรือข้อความหยาบคาย 2) Statistical Cues เนื่องจากการระบุตัวชี้นำที่เกี่ยวข้องกับอารมณ์การใช้วิธีที่ 1 อาจไม่เพียงพอในการระบุอารมณ์ที่ต้องการได้อย่างถูกต้อง ดังนั้นชุดคุณลักษณะที่สองจึงถูกนำออกมาจากชุดเรียนรู้โดยอัติโนมัติ ซึ่งชุดนี้ประกบด้วย unigrams, bigrams และ trigrams ที่สกัดจากคำพูดแต่ละคำในชุดทดสอบ จากนั้นจะมีการใช้ขั้นตอนการเลือกคุณลักษณะเพื่อลดจำนวนคุณลักษณะที่จะลบข้อมูลที่ไม่เกี่ยวข้องออกไป 3) Linguistic information ข้อมูลทางภาษาศาสตร์ เพื่อพิจารณาทางสถิติ จำเป็นต้องมีชุดเรียนรู้ขนาดใหญ่เพื่อดึงชุดตัวชี้นำให้เพียงพอเพื่อใช้กับชุดทดสอบต่างๆ จึงได้นำคุณลักษณะทางด้านภาษา POS-tagging ที่เกี่ยวข้องกับแต่ละคำ ในชุดนี้เป็นการใช้งานชุด POS n-grams ซึ่ง n = 1,2,3 4) Semantic information ข้อมูลเกี่ยวกับความหมาย: แม้ว่ารูปแบบบางอย่างอาจเป็นตัวชี้นำที่ดีต่ออารมณ์ที่แตกต่างกัน แต่ก็มีข้อมูลเพิ่มเติมในการสนทนาออนไลน์ซึ่งเกี่ยวข้องกับประเภทความหมายมากกว่าคำเฉพาะหรือประเภทของ POS 5) Length information ข้อมูลความยาว เนื่องจากคลังข้อมูลที่เฉพาะเจาะจงที่กำลังรับมือซึ่งในโพสต์ความยาวแตกต่างกันมากจึงสังเกตเห็นว่าข้อมูลความยาวสามารถเป็นประโยชน์อย่างมากสำหรับการตีความคุณลักษณะชุดก่อนหน้านี้อย่างถูกต้อง ดังนั้นในงานนี้จะมีเวกเตอร์ที่มีข้อมูลต่อไปนี้สำหรับคำพูดแต่ละคำเช่นจำนวนคำจำนวนอักขระจำนวนประโยคคำเฉลี่ยต่อประโยคตัวอักษรเฉลี่ยต่อประโยค 6) Concept and Polarity Information แนวคิดและข้อมูลเกี่ยวกับขั้ว พิจารณาว่าการประชดประชันและความหยาบคายทั้งสองอย่างเชื่อมโยงกันอย่างใกล้ชิดกับสถานะทางอารมณ์ที่เฉพาะเจาะจงชุดคุณลักษณะใหม่ ๆ รวมทั้งข้อมูลทางอารมณ์ได้รับการพิจารณาด้วย โดยเฉพาะ SenticNet-3.0 ถูกนำมาใช้ วิธีการจำแนกประเภท คือ rule-based classifier และ Naïve Bayes classifier

Kunneman และคณะ [49] ได้นำเสนอระบบการตรวจจับข้อความประชดประชันสำหรับทวิต ข้อความบนไมโครบล็อคจากเว็บไซต์ทวิตเตอร์ ในการเก็บรวบรวมข้อมูลสำหรับการทำการวิจัยใช้ข้อมูลจากแฮชแท็กจากผู้ใช้ที่นำเสนอข้อความเป็นข้อความประชดประชันอย่างชัดเจนผ่านการใช้

แฮชแท็ก เช่น #sarcasm or #not โดยข้อมูลที่เก็บรวบรวมในภาษาดัชมาจำนวน 2.25 ล้านทวิต จำนวนชุดเรียนรู้ทั้งหมด 406000 ทวิตที่ไม่มีแท็กดังกล่าว อัลกอริทึมที่ใช้ในการจำแนกคือ Winnow classification การจำแนกด้วยค่าคะแนน AUC (Area Under the Curve) มีค่าเท่ากับ 0.84 และสามารถระบุจุดที่ต้องการได้อย่างถูกต้อง 309 จาก 353 ทวีตในจำนวน 2.25 ล้านทวิตที่ทำเครื่องหมายแฮชแท็กที่เอาเครื่องหมายแฮชแท็กออก จากการทดสอบโดยตัวแบ่งประเภทใน 250 อันดับแรกของทวีตที่ได้รับการจัดอันดับมากที่สุดน่าจะเป็นข้อความประชดที่ไม่มีแฮชแท็กประชด แต่มีความแม่นยำเพียง 35%เท่านั้น

Schifanella และคณะ [50] นำเสนอการศึกษาความสัมพันธ์ระหว่างแง่มุมของข้อความและภาพในบทความแบบหลายรูปแบบจากแพลตฟอร์มสื่อสังคมออนไลน์ Instagram, Tumblr และ Twitter สิ่งที่ศึกษามีดังนี้ 1) การศึกษาการมีอิทธิพลซึ่งกันและกันระหว่างข้อความและรูปภาพในโพสของสื่อสังคมทั้งสาม 2) ศึกษาคุณภาพของการตรวจจับข้อความประชดด้วยภาพเปรียบเทียบกับการติดคลาสด้วยมนุษย์ 3) แสดงให้เห็นถึงประสิทธิภาพที่ดีขึ้นเมื่อเทียบกับแพลตฟอร์มพื้นฐานและวิธีการต่างๆ ที่เป็นข้อความ ชุดของข้อมูลเก็บรวบรวมข้อมูลที่เป็นข้อความและมีรูปภาพประกอบในข้อความ จาก Instagram, Tumblr และ Twitter การใช้ตัวกรองข้อมูลโดยไม่สนใจโพสที่ไม่มีรูปภาพ , ตัดข้อความที่เป็น @username, URL, ไม่สนใจข้อความที่มีคำ sarcasm อยู่โดยไม่มีแฮชแท็ก และข้อความที่มีจำนวนน้อยกว่า 4 คำ ในแต่ละแพลตฟอร์มได้จำนวนชุดข้อมูลที่ได้มาจากการสุ่มจำนวนแพลตฟอร์มละ 10000 โพส วิธีการที่นำมาใช้ในการจำแนก คือ 1) SVM Approach ในส่วนของ NLP Features (Lexical, Subjectivity and sentiment score, n-grams, word2vec และ combination) 2) Deep learning ในส่วนของ Visual features (Visual Semantics Features: VSF) โดยสรุปการใช้วิธีการรวมกันระหว่าง NLP Features และ VSF ให้ประสิทธิภาพที่สูง คือ Instagram เท่ากับ 82.3% Tumblr เท่ากับ 81.0% และ Twitter เท่ากับ 80.0%

Tasliogl u และ Karagoz [51] นำเสนอเทคนิคในการตรวจจับข้อความประชดประชันบนไมโครบล็อกไซต์ทวิตเตอร์ในภาษาตุรกี ด้วยวิธีการเรียนรู้ด้วยเครื่อง ชุดข้อมูลที่ใช้ศึกษาถูกเก็บรวบรวมจากทวิตเตอร์ด้วยทวิตเตอร์เอพีไอ (Twitter API) จำนวน 54362 ทวิต เป็นทวิตภาษาตูรกีในช่วงเวลาตุลาคม 2556 – กันยายน 2557 วิธีการที่เลือกใช้ในการสกัดคุณลักษณะ ซึ่งคุณลักษณะบางอย่างเหล่านี้มีเฉพาะสำหรับตุรกีเท่านั้น การใช้เครื่องหมายวรรคตอน “!”, “?”, “ ” ”, “.”, :), :(, คำอุทาน, ตัวอักษรพิมพ์ใหญ่ และ ช่องว่างระหว่างค่าคะแนนความรู้สึก ใช้วิธีการเรียนรู้ด้วยเครื่องแบบมีผู้สอนด้วยอัลกอริทึม Naïve Bayes, Support Vector Machine, K-Nearest Neighbors และ Random Forest ผลการเปรียบเทียบแต่ละอัลกอริทึม K-Nearest Neighbors ให้ประสิทธิภาพสูงที่สุด precision = 94.4% Recall = 93.8% และ F1 = 93.7%

Freitas และคณะ [45] ได้ศึกษาการจรวจสอบข้อความประชดบนทวิต นำเสนอรูปแบบภาษาสำหรับการประชด โดยเก็บรวบรวมข้อมูลทวิตภายใต้หัวข้อ “End of the world” ประกอบด้วย 2779 ทวีต (55663 คำ) การจำแนกประเภทที่ถูกนำมาใช้มีจำนวน 6 ประเภท ดังนี้ รายชื่อ (lists), การแสดงออกที่แน่นอน (exact expressions), หน้าที่ของคำ (Part of speech) + exact expressions, หน้าที่ของคำ (Part of speech) + lists, Part of speech + named entities และ สัญลักษณ์ (Symbols) สามารถอธิบายรูปแบบของคุณลักษณะได้ดังนี้คือ 1) รายการของการแสดงอาการหัวเราะ (List of Laughter Expression) 2) รายการของสัญลักษณ์แสดงอารมณ์ (List of Emoticons) 3) só que ในภาษาบลาซิล เป็นวิธีที่มีอยู่แล้วในการแสดงให้เห็นว่าคำพูดไม่ได้หมายความว่าสิ่งที่เห็นได้ชัด 4) sim แสดงให้เห็นถึงความสำคัญของสิ่งที่จะกล่าวและทำให้อาจหมายถึงการประชด 5) na boa นิพจน์นี้มีมากกว่าหนึ่งความหมาย ใช้เมื่อพูดถึงหัวข้อที่อาจทำให้เกิดความเห็นแย้งและตลก 6) การใช้ #ironia #sarcasmo #joking #kiding แท็กนี้เป็นที่ชัดเจนว่าผู้ใช้ต้องการเน้นความเห็นขัน 7) tão + Adjective หรือ tão + Adverb ซึ่งเป็นคำวิเศษณ์ในภาษาโปรตุเกส อาจใช้เพื่อทำเครื่องหมายระดับความเท่าเทียมกันของคำวิเศษณ์อื่น ๆ ของคำคุณศัพท์ แต่อาจให้ความรุนแรงกับคำเหล่านี้ด้วย 8) Adjective + list of Emoticons ถ้าทวีตนำคำคุณศัพท์เชิงบวก แต่อารมณ์เป็นลบจะไปสำหรับรายการของผลนอกจากนี้คำคุณศัพท์เชิงลบและอีโมติคอนในเชิงบวกแสดงให้เห็นถึงผลดี ถ้าคำคุณศัพท์และอีโมติคอนมีขั้วเดียวกันทวีตจะถูกปฏิเสธ 9) DET + ADJ + (PREP+DET) + NE รูปแบบนี้ต้องการการวิเคราะห์ Part of speech ในลำดับที่แนะนำ 10) Demonstrative Pronouns + NE การเกิดขึ้นของคำสรรพนามชี้เฉพาะ (This, that) ให้ประสิทธิผลดี 11) <EXPR>! การบรรยายข้อความประชดประชันผู้เขียนมักจะเขียนต้องการเพิ่มความคิดเห็นซึ่งเป็นการแสดงความคิดเห็นเกี่ยวกับสิ่งที่ได้กล่าวมาก่อนหน้า 12) !*|?*|!*?*|?*!* การใช้เครื่องหมายวรรคตอนส่งเนื้อหาเพิ่มเติมไปยังคำพูด โดยเฉพาะการใช้เครื่องหมายอัศเจรีย์และเครื่องหมายคำถามทำให้อาจได้อารมณ์จากการสื่อสาร 13) Quotation Marks นอกเหนือจากการถ่ายทอดสิ่งที่คนพูดไว้ในข้อความสัญลักษณ์เหล่านี้ใช้เพื่อเน้นคำบางคำหรือสำนวนและให้ความสำคัญกับความหมายที่เป็นรูปเป็นร่าง เช่น เรื่องประชดประชัน

Dana และ คณะ [52] ได้นำเสนอการตรวจจับข้อความประชดประชันในภาษาอาราบิคบนเครือข่ายสังคมออนไลน์ทวิตเตอร์ โดยใช้เทคนิคและอัลกอริทึมด้านดาต้าไมนิ่ง โดยการเก็บรวบรวมข้อมูลภาษาอาราบิค จำนวน 350 ที่เป็นคำประชดและไม่ใช่คำประชดจาก 11 แฮชแท็กที่แตกต่างกันโดยไม่สนใจรูปภาพและวิดีโอ กระบวนการในการเตรียมข้อมูลโดยการนำแฮชแท็ก ไอคอนแสดงอารมณ์ และ ที่อยู่เว็บไซต์ออกจากข้อความ จากนั้นจะได้ข้อมูลจำนวน 344 ทวีต แบ่งเป็น 238 เป็นข้อความประชดประชัน และ 106 ที่ไม่ใช่ข้อความประชดประชัน การเลือกคุณลักษณะโดยการใช้เครื่องหมาย เช่น !, ?, ..., (), “” , คำที่เป็น เชิงบวก และ เชิงลบ ซึ่งหากพบคำที่เป็นเชิงบวกและเชิง

ลบในข้อความเดียวกันจะระบุว่าข้อความนั้นเป็นข้อความประชดประชัน การทดลองใช้อัลกอริทึมนาอีฟเบย์ ในการจำแนก วัดประสิทธิภาพโดยใช้ค่า ความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวม ผลการวัดประสิทธิภาพ ค่าความแม่นยำเท่ากับ 65.9% ค่าความระลึก 71% และค่าประสิทธิภาพโดยรวมเท่ากับ 67.6%

Hiai และ Shimada [41] นำเสนอวิธีในการสกัดรูปแบบในการประเมินผลข้อความประชดประชัน ซึ่งได้ให้ข้อคิดเห็นว่าการแสดงด้วยข้อความประชดประชันนั้นเป็นการแสดงทางความหมายที่เป็นเชิงลบด้วยข้อความเชิงบวก ซึ่งงานทางด้านการตรวจสอบหรือตรวจหาข้อความประชดประชันจะช่วยเพิ่มประสิทธิภาพในงานทางด้านการวิเคราะห์อารมณ์ความรู้สึก ดังนั้นจึงได้เสนอวิธีการสกัดประโยคประชดประชันจากข้อความรีวิวสินค้า โดยการจำแนกประโยคออกเป็น 8 คลาส ชุดข้อมูลที่ใช้เป็นชุดข้อมูลความคิดเห็นจากรีวิวสินค้าญี่ปุ่น จากบริษัท รากุเทน อิชิบา (Rakuten Ichiba) โดยข้อมูลประกอบด้วยชื่อสินค้า การให้คะแนนสินค้า หัวข้อรีวิว และประโยคที่รีวิว จำนวนทั้งหมด 10,000 รีวิว ที่มีจำนวนประโยคทั้งหมด 34,917 ประโยค ในการวิเคราะห์ประโยคเพื่อจำแนกคลาสทั้ง 8 คลาส ได้ดังนี้ 1) การแสดงข้อความที่เป็นเชิงบวก และ เชิงลบที่ให้คะแนนในระดับน้อยในประโยคเดียวกัน 2) การแสดงข้อความคิดเห็นในเชิงบวกในหัวข้อการรีวิวและแสดงความคิดเห็นเชิงลบในข้อความเนื้อหารีวิว ซึ่งให้ความสำคัญถึงความหมายเชิงลบ 3) การแสดงข้อความ “ii” ซึ่งหมายถึง “Good” ในภาษาอังกฤษ ซึ่งการรวมกันของข้อความที่มีความหมายซับซ้อน เช่น “ii (Good)” และ “nedan (price)” ทำให้เปลี่ยนความหมายเป็น แพงมาก (too expensive) เป็นการแสดงออกเชิงบวกของคลาสนี้ 4) คลาสนี้แสดงเป้าหมายของการประเมินผลในด้านการส่งของ แต่ไม่ใช่ผลิตภัณฑ์ที่ต้องส่ง ดังนั้นจึงเป็นการให้ความหมายเชิงลบกับผลิตภัฑณ์ซึ่งแสดงออกผ่านข้อความเชิงบวก 5) การแสดงออกในเชิงบวกในสถานการณ์ที่ไม่ดี เช่น มีข้อความ “ii (better)” การสูญเสียเงินอีกครั้ง ซึ่งแสดงถึงความหมายเชิงลบ 6) การแสดงออกในเชิงบวกกับผลิตภัณฑ์อื่น ซึ่งเป็นการบงบอกว่าผลิตภัณฑ์นี้ไม่ดี 7) การแสดงออกเชิงบวกโดยใช้ข้อความที่ไม่ใช่ข้อความปกติ ซึ่งเป็นการบงบอกถึงความหมายเชิงลบกับผลิตภัณฑ์ ด้วยการแสดงข้อความเชิงบวก 8) เป็นการแสดงออกเชิงบวกแต่ไม่ได้แสดงถึงจุดบวกที่ชัดเจนเกี่ยวกับผลิตภัณฑ์ กล่าวอีกนัยหนึ่งผลิตภัณฑ์ไม่มีจุดเด่นชัดแม้ว่าประโยคนั้นจะมีนิพจน์เชิงบวก

Jain และคณะ [53] นำเสนอวิธีการในการตรวจจับข้อความประชดประชันบนทวิตเตอร์ โดยให้ความเห็นว่า การตรวจจับข้อความประชดประชันนั้นถือเป็นงานที่ท้าทายเนื่องจากข้อความประชดประชันที่ผู้คนสื่อสารออกมาผ่านเครือข่ายสังคมออนไลน์นั้นมีลักษณะคลุมเคลือประกอบด้วยข้อความที่เป็นคำแสลง การแสดงออกถึงประโยคที่เป็นเชิงบวกหรือเชิงลบ แต่ใช้ไอคอนแสดงอารมณ์ที่เป็นลักษณะตรงกันข้ามกับข้อความที่สื่อสารออกมา และการแสดงสื่อสารอารมณ์ความรู้สึกเชิงบวกในสถานการณ์ที่เป็นเชิงลบ วิธีการที่นำเสนอสำหรับการจำแนกข้อความประชดประชันนั้นใช้ชุดข้อมูล

จากทวิตเตอร์โดยเก็บข้อมูลจำนวน 35,000 เพื่อใช้เป็นชุกการเรียนรู้และจำนวน 15,000 เพื่อใช้เป็นข้อมูลชุดทดสอบรูปแบบการเก็บข้อมูลโดยใช้แฮชแท็ก #sarcasm และ #sarcastic ด้วยทวิตเตอร์เอพีไอ ในการทำความสะอาดข้อมูลสิ่งที่กำจัดออกไปเนื่องจากไม่สนใจในงานนี้คือ เครื่องหมายวรรคตอน ข้อมูลตัวเลข และตรวจสอบและแก้ไขการสะกดคำ ขั้นตอนในการสะกัดคุณลักษณะใช้ N-grams และใช้หน้าที่ของคำในการระบุความหมาย เช่น ใช้ 1-Gram โดยใช้ VERB ในการระบุขั้วในเชิงบวกหรือเชิงลบ การให้ค่าน้ำหนักได้เลือกใช้ TF-IDF การใช้เทคนิคด้านการจำแนกโดยใช้การเรียนรู้ของเครื่องและเหมืองข้อมูล คือ นาอีฟเบย์ ลิเนียรีเกรสชั่น แรนดอมฟอเรส และ ใช้การให้ค่าน้ำหนักในการเลือกการวัดการจำแนกด้วยเอ็นเซมเบลอที่เป็นการรวมกันด้วยอัลกอริทึม นาอีฟเบย์ ลิเนียรีเกรสชั่น แรนดอมฟอเรส ผลการทดลอง พบว่าค่าความถูกต้องด้วยวิธีการโหวตให้ประสิทธิภาพที่ดีและอัลกอริทึมแรนดอมฟอเรส ด้วยค่าความถูกต้อง 84 และ 85%

Razali และคณะ [54] นำเสนอการศึกษาความสำคัญในการจำแนกข้อความประชดประชันสำหรับงานด้านการวิเคราะห์อารมณ์ความรู้สึก ซึ่งได้ศึกษาวิธีการต่างๆ ดังนี้ 1) การใช้วิธีการจำแนกด้วยกฎแฮชแท็ก โดยกฎระบุว่า หากพบแฮชแท็ก (#sarcasm) ก็จะไม่สนใจข้อความอื่นๆ นั้นหมายถึงเป็นข้อความประชดประชันและหากมีการใช้กริยาที่เป็นเชิงบวกในสถานการณ์ที่เป็นเชิงลบก็จำแนกได้ว่าเป็นข้อความประชดประชัน 2) วิธีที่สองใช้วิธีทางสถิติโดยใช้รูปแบบเป็นฐานของคุณลักณะโดยพิจารณาจากสถานการณ์ที่ตรงกัน สถานการณ์ที่แตกต่างกัน และ สถานการณ์ที่ไม่เกี่ยวข้องกัน ฐานข้อมูลด้านอารมณ์ความรู้สึก และไอคอนแสดงอารมณ์ได้ถูกนำมาใช้ในการพิจารณาด้วย 3) ใช้วิธีการการเรียนรู้เชิงลึก (Deep learning) ในการพิจารณาเนื้อหาจากรูปภาพ เนื่องจากการแสดงการประชดประชันบนเครือข่ายสังคมออนไลน์มีการใช้รูปภาพประกอบเป็นจำนวนมาก

Gidhe และ Ragha [55] นำเสนอวิธีการจำแนกข้อความคิดเห็นประชดประชันบนข้อความที่ไม่มีแฮชแท็กด้วยวิธี Multilayer perceptron-Backpropagation (MLP-BP) โดยให้เหตุผลของการศึกษาว่า ข้อความประชดประชันนั้นเป็นข้อความที่แสดงถึงอารมณ์ความรู้สึกที่ถูกซ่อนเอาไว้ซึ่งสื่อสารออกมาผ่านข้อความที่มีความหมายถึงกันข้าม ชุดข้อมูลที่นำมาใช้จากฐานข้อมูล Raddit การสกัดคุณลักษณะมีดังนี้ 1) คุณลักษณะทางโครงสร้าง โดยการนำการเกิดขึ้นของเครื่องหมายวรรคตอน เครื่องหมายทางการสนทนา และไอคอนแสดงอารมณ์ 2) คุณลักษณะทางอารมณ์โดยใช้คลังคำแสดงอารมณ์จาก Whissell Dictionary 3) คุณลักษณะเชิงความหมายของความคล้ายคลึกกันระหว่างคำประชดประชันกับความรู้สึกที่แท้จริงของคำเป้าหมาย

Chaudhari และ Chandankhede [56] นำเสนอการศึกษาในการสำรวจเกี่ยวกับการทดลองการจำแนกข้อความประชดประชัน จากการศึกษาพบว่าสามารถแบ่งชนิดของการจำแนกคำประชดประชดประชันได้ดังนี้ 1) คำประชดประชันที่การบ่งบอกความแตกต่างของอารมณ์ความรู้สึก เช่น ความแตกต่างระหว่างอารมณ์ความรู้สึกทางบวกและสถานการณ์ทางลบ ความแตกต่างระหว่าง

อารมณ์ความรู้สึกทางลบและสถานการณ์ทางบวก ความแตกต่างทางความหมาย ความแตกต่างระหว่าปัจจุบันกับอดีต การปฏิเสธความจริง และการกล่าวถึงข้อเท็จจริงชั่วคราว 2) คำประชดประชันที่เป็นสื่อในการถ่ายทอดอารมณ์ความรู้สึก เช่น การแสดงความรอบรู้และความเฉลียวฉลาดหรือมีปฏิภาณ การแสดงถึงอารมณ์การครำ่ครวญ การพูดเลอะเทอะ การแสดงการพูดหลบหลีก การพูดเลี่ยงหรือการพูดโกหก การแสดงออกถึงความรุนแรง การอาละวาด 2) คำประชดประชันที่เป็นรูปแบบของการแสดงออกเป็นลายลักษณ์อักษร เช่น การใช้เครื่องหมายวรรคตอน การใช้ตัวอักษรพิมพ์ใหญ่ทั้งหมด เพื่อเป็นตัวบ่งชี้ถึงการประชดประชัน การแสดงออกในรูปแบบโครงสร้างของประโยคซึ่งมีสองส่วนในประโยคเดียวที่มีขั้วแตกต่างกัน การแสดงออกในรูปแบบการวิเคราะห์คำศัพท์ซึ่งส่วนนี้เป็นการใช้แฮชแท็ก เช่น #sarcasm #irony 3) คำประชดประชันที่เป็นการแสดงออกถึงการมีความเชี่ยวชาญ เช่น ความเชี่ยวชาญทางด้านภาษา, แสดงถึงทักษะความรู้ทางสภาพแวดล้อม ในส่วนของการเลือกชนิดของคุณลักษณะ 1) คุณลักษณะทางคำศัพท์ ในส่วนนี้เช่น bigram, n-gram, skip-gram, #hashtag 2) คุณลักษณะในการปฏิบัติ เช่น การใช้สัญลักษณ์ต่างตางๆ เช่น รูปภาพไอคอนแสดงอารมณ์ 3) คุณลักษณะของการใจคำพูดเกินจริง เช่น การเน้นย้ำ คำอุทาน เครื่องหมายคำพูด เครื่องหมายวรรคตอนต่างๆ 4) คุณลักษณะตามรูปแบบ เช่น รูปแบบการเกิดคำที่ปรากฏซ้ำๆ 5) คุณลักษณ์ของประโยค เช่น การรวมกันของลักษณะของหน่วยคำ หน้าที่ของคำ 6) คุณลักษณะตามบริบท ที่ใช้ข้อมูลนอกเหนือจากข้อความ เช่น สัญญลักษณ์ต่างๆ 7) คุณลักษณะของคำอุปมาอุปมัย เช่น คำนามที่มีความหมายเชิงบวกหรือเชิงลบ ลักษณะของการเน้นคำคุณศัพท์, สุภาษิตและคำพังเพย ประเด็นปัญหาของการจำแนกข้อความประชดประชัน ได้นำเสนอไว้ดังนี้ 1) ปัญหาทางของชุดข้อมูล ประเด็นด้านข้อมูลที่นำมาใช้ในการวิเคราะห์ทางด้านนี้ยังค่อนข้างคลุมเครือเนื่องจากผู้ที่ศึกษาด้านนี้ส่วนมากใช้การรวบรวมข้อมูลโดยใช้แฮชแท็ก ซึ่งเมื่อนำแฮชแท็กออก ข้อความเหล่าจะมีความที่ไม่ใช่ข้อความประชดประชันทันที 2) ปัญหาของคุณลักษณะ ประโยคที่เป็นข้อความประชดประชัน การใช้ตัวจำแนกด้านอารมณ์ความรู้สึกประสิทธิภาพการจำแนกอาจลดลง ด้านความรู้สึกสามารถใช้เป็นคุณลักษณะสำหรับตัวจำแนกประเภทได้และจำเป็นต้องใชขั้วของประโยค ดังนั้นคุณลักษณะใหม่ควรมีการสำรวจและใช้งานร่วมกับคุณลักษณะอื่นที่มีอยู่เพื่อให้ได้ความถูกต้องที่ดียิ่งขึ้น 3) ปัญหาทางด้านการเลือกใช้ตัวจำแนก ซึ่งบางครั้งนักวิจัยอาจใช้ชุดข้อมูลที่มีขนาดใหญ่ หรือขนาดเล็ก ซึ่งก็ไม่ใช่ประเด็นสำคัญเพราะสามารถทำให้ข้อมูลมีความสมดุลกันได้ ดังนั้นควรใช้เทคนิคในการจำแนกประเภทที่ถูกต้องในชุดข้อมูลเพื่อจำแนกประเภทของประโยคประชดประชันและไม่ประชดประชันที่ถูกต้อง

Bhan และ D'silva [57] นำเสนอรูปแบบหัวข้อข้อความประชดประชันด้วยการวิเคราะห์อารมณ์ความรู้สึก โดยใช้การเรียนรู้ของเครื่อง ชุดของข้อมูลประกอบด้วย 3 ส่วน จากทวิตเตอร์ คือ ชุดเรียนรู้ ชุดการปรับปรุงคุณภาพ และชุดทดสอบ คุณลักษณะที่ใช้ ประกอบด้วย 4 คุณลักษณะคือ

คุณลักษณะที่เกี่ยวข้องกับอารมณ์ความรู้สึก คำที่เป็นเชิงบวก เชิงลบ คุณลักษณะที่เกี่ยวข้องกับเครื่องหมายวรรคตอน ซึ่งใช้การนับจำนวนของเครื่องหมายเหล่านี้ เช่น เครื่องหมายอัศเจรีย์ เครื่องหายคำถาม จุด การใช้ตัวอักษรพิมพ์ใหญ่ทั้งหมด เครื่องหมายคำพูด คุณลักษณะทางศัพท์และประโยค เช่น การนับจำนวนของการใช้คำที่ไม่ปกติ การแสดงข้อความที่เป็นคำประชดทั่วๆ ไป จำนวนของคำอุทาน จำนวนของของการแสดงอารมณ์ขบขัน การใช้คุณลักณะที่เกี่ยวข้องกับรูปแบบของคำและหน้าที่ของคำ อัลกอริทึมที่ใช้ในการจำแนก คือ ซัพพอร์ทเวกร์เตอร์แมคชีน ลอจิสติกรีเกรซชั่น และ นาอีฟเบย์ ซึ่งซัพพอร์ทเวกร์เตอร์แมคชีน ลอจิสติกรีเกรซชั่น ให้ประสิทธิภาพโดยรวมเท่ากับ 86% และ นาอีฟเบย์ได้ค่าประสิทธิภาพโดยรวมเท่ากับ 83%

Miljana และคณะ [44] นำเสนอการจำแนกข้อความประชดประชันด้วยเทคนิคคลังคำ ชุดข้อมูลที่ใช้ในการทดสอบเป็นชุดข้อมูลที่ไม่สมดุล 319 เป็นข้อความประชดประชันและ 1,413 เป็นข้อความไม่ประชดประชัน คุณลักษณะที่ใช้มีดังนี้ การเปรียบเทียบกับ Serbian wordnet, การจับคู่ระหว่างสมาชิกที่มีขั้วอารมณ์ความรู้สึกเชิงบวก, การเรียงลำดับของแท็กอารมณ์ความรู้สึก, หน้าที่ของคำ และการสื่อถึงข้อความประชดประชันโดยตรง ผลของการจำแนกที่ได้ประสิทธิภาพสูงที่สุดคือ การรวมกันของทุกคุณลักษณะ ได้ค่าความถูกต้องเท่ากับ 86.10%

Manohar และ Kulkarni [58] นำเสนอการศึกษาการพัฒนาการวิเคราะห์ข้อความประชดประชันโดยใช้การประมวลผลภาษาธรรมชาติด้วยวิธีการใช้คลังคำ โดยใช้ข้อมูลจำทวิตเตอร์ด้วยทวิตเตอร์เอพีไอ การเตรียมข้อมูลโดยการกำจัดข้อความที่ไม่สนใจ เช่น ชื่อผู้ใช้งาน ที่อยู่เว็บไซต์ และแฮชแท็กออก และให้ความสนใจไอคอนแสดงอารมณ์ การตัดคำใช้โมเดล uni-gram โดยใช้ช่องว่าง มีการใช้คุณลักษณะด้วยการหาหน้าที่ของคำ และวิธีการแบบคลังคำ ในการวัดประสิทธิภาพด้วยค่าความถูกต้องเท่ากับการหาจำนวนของคำที่เป็นคำประชดประชันหารด้วยจำนวนของข้อความผลของการจำแนกแบบเรียลไทม์ให้ผลลัพท์ที่ดี

Jihen และคณะ [59] นำเสนอวิธีการในการจำแนกข้อความประชดประชันในภาษาอาราบิคบนเครือข่ายสังคมออนไลน์ โดยให้เหตุผลของการศึกษาว่าการจำแนกข้อความประชดประชันบนเครือข่ายสังคมออนไลน์นั้นมีความยากเนื่องจากข้อความเหล่านั้นเป็นข้อความที่ยากที่จะเข้าใจความหมายว่าผู้โพสข้อความนั้นต้องการสื่อสารออกมาอย่างไร ในการศึกษานี้ได้ใช้ชุดข้อมูลทั้งหมด 5,479 โดยแบ่งเป็นข้อความประชดประชันจำนวน 1,733 และ ข้อความที่ไม่ใช่ข้อความประชดประชันจำนวน 1,733 ข้อความ วิธีการเลือกคุณลักษณะ 1) Surface features โดยพิจารณาสัญลักษณ์ เช่น เครื่องหมายวรรคตอน, ไอคอนแสดงอารมณ์, เครื่องหมายคำพูด, คำที่เป็นความหมายถึงกันข้าม (but), จำนวนและลำดับของเครื่องอัศเจรีย์และเครื่องหมายคำถาม, การรวมกันของเครื่องอัศเจรีย์และเครื่องหมายคำถาม, คำพูดที่ไม่ได้แสดงถึงความขัดแย้ง, คำอุทาน และจำนวนของไอคอนแสดงอารมณ์ 2) Sentiment features การใช้คลังคำที่รวบรวมคำที่มีความหมายเชิงบวกและ

เชิงลบ เช่น การแปลจากคลังคำ Bing Liu ซึ่งมีคำที่มีความหมายเชิงบวกจำนวน 2,006 และที่มีความหมายเชิงลบจำนวน 4,783 คำ โดยใช้เครื่องมือแปลภาษาของกูเกิล, การแปลคลังคำที่รวมคำเชิงบก เป็นกลาง และเชิงลบ จำนวน 2,718, 570 และ 4,911 ตามลำดับ จากคลังคำ MPQA Subjectivity, คลังคำไอคอนแสดงอารมณ์จากภาษาอาราบิค ซึ่งมีความคิดเห็นอยู่ระหว่าง -7 และ 7 ของคำที่เป็นเชิงลบ และเชิงบวกจำนวน 22,962 และ 20,342 และ คลังคำแฮชแท็กภาษาอาราบิคมีคำเชิงบวกและเชิงลบ จำนวน 11,941 และ 8,179 จากนั้นจะได้คำที่เป็นเชิงบวกและเชิงลบจำนวนทั้งสิ้น 26,777 และ 22,239 ซึ่งได้กำจัดคำที่ซ้ำกันออก การทดลองใช้เครื่องมือ Weka toolkit โดยใช้อัลกอริทึม ซัพพอร์ทเวกร์เตอร์แมคชีน, นาอีฟเบย์, ลอจีสติก รีเกรชชั่น, ลิเนียร์ รีเกรชชั่น, แรนดอมทรี และ แรนดอมฟอเรส โดยใช้ 10-cross validation ในการทดสอบ ผลที่ได้ประสิทธิภาพมากที่สุดคือ แรนดอมฟอเรส จากการวัดด้วยค่าความถูกต้องเท่ากับ 72.36 เปอร์เซนต์ ค่าความแม่นยำเท่ากับ 72.90% ค่าความระลึกเท่ากับ 73.50% และค่าประสิทธิภาพโดยรวมเท่ากับ 72.70%

Liebrecht และคณะ [60] นำเสนอวิธีการในการจำแนกข้อความประชดประชันบนทวิตเตอร์ในภาษาดัช โดยการเก็บข้อมูลทวิตเตอร์จำนวน 78,000 ทวิตด้วยแฮชแท็ก (#sarcasme) ซึ่งหมายถึงคำประชดประชันในภาษาดัช การเตรียมข้อมูลในส่วนของการตัดคำให้ความสนใจเครื่องหมายวรรคตอน และตัวอักษรตัวพิมพ์ใหญ่ กรณีข้อความที่สั้นและแฮชแท็กถูกกำจัดออก ใช้โมเดล Uni, Bi, Tri-grams ในการนำใช้เป็นคุณลักษณะ

Ahmad และคณะ [61] นำเสนอวิธีการตรวจจับข้อความประชดประชันจากเอกสารบนเว็บโดยใช้เทคนิควิธีการเรียนรู้ของเครื่อง ชุดข้อมูลที่ใช้ในการศึกษาจำนวน 2795 แบ่งเป็นข้อมูลไม่ประชดประชันจำนวน 2624 และ ข้อความประชดประชันจำนวน 171 การเตรียมข้อมูลใช้เครื่องมือ NLTK took kit จากนั้นกำจัดคำหยุดซึ่งเป็นคนที่ไม่มีความสำคัญออกไป และ ทำการหารากศัพท์เพื่อลดจำนวนคุณลักษณะ วิธีการที่เลือกใช้ในการสะกัดคุณลักษณะคือ TF-IDF (Term Frequency - Inverse Docment Frequency), TF-BNS (Term Frequency Bi-normal separation), BIN (Binary) และ TF-BNS-IDF (Term Frequency – Bi-normal separation scaling – Inverse Docment Frequency) อัลกอริทึมที่ใช้ในการเรียนรู้คือซัพพอร์ทเวกเตอร์แมคชีน วัดประสิทธิภาพการทดลองด้วย ค่าความถูกต้อง ค่าความแม่นยำ ค่าระลึก และค่าประสิทธิภาพโดยรวม โดยผลการทดลองแยกตามคุณลักษณะดังนี้ 1) TF-IDF ค่าความถูกต้องเท่ากับ 83.41% ค่าความแม่นยำเท่ากับ 72.20% ค่าระลึกเท่ากับ 73.20% และค่าประสิทธิภาพโดยรวมเท่ากับ 72.70% 2) BIN ค่าความถูกต้องเท่ากับ 84.08% ค่าความแม่นยำเท่ากับ 74.60% ค่าระลึกเท่ากับ 78.80% และค่าประสิทธิภาพโดยรวมเท่ากับ 76.70% 3) BNS ค่าความถูกต้องเท่ากับ 89.92% ค่าความแม่นยำเท่ากับ 80.20% ค่าระลึกเท่ากับ 85.90% และค่าประสิทธิภาพโดยรวมเท่ากับ 82.90% และ 4) TF-IDF ค่าความถูก

ต้องเท่ากับ 92.75% ค่าความแม่นยำเท่ากับ 82.60% ค่าระลึกเท่ากับ 87.00% และค่าประสิทธิภาพโดยรวมเท่ากับ 84.00%

Vateekul และคณะ [62] ได้ศึกษาเกี่ยวกับการวิเคราะห์อารมณ์ความรู้สึกข้อความภาษาไทย บนทวิตเตอร์ โดยใช้เทคนิค Long Short-Term Memory (LSTM) และ Dynamic Convolutional Neural Network (DCNN) เพื่อเปรียบเทียบประสิทธิภาพกับการเรียนรู้ของเครื่องด้วยอัลกอริทึม Naïve Bayes, Support Vector Machine และ Maximum Entropy ซึ่ง LSTM และ DCNN ให้ประสิทธิภาพดีกว่า และ DCNN ให้ค่าความถูกต้องมากที่สุด โดยมีผลลัพธ์ค่าความถูกต้องดังนี้ DCNN 75.35% LSTM 75.30% MaxEnt 75.13% SVM 74.71% และ NB 74.05%

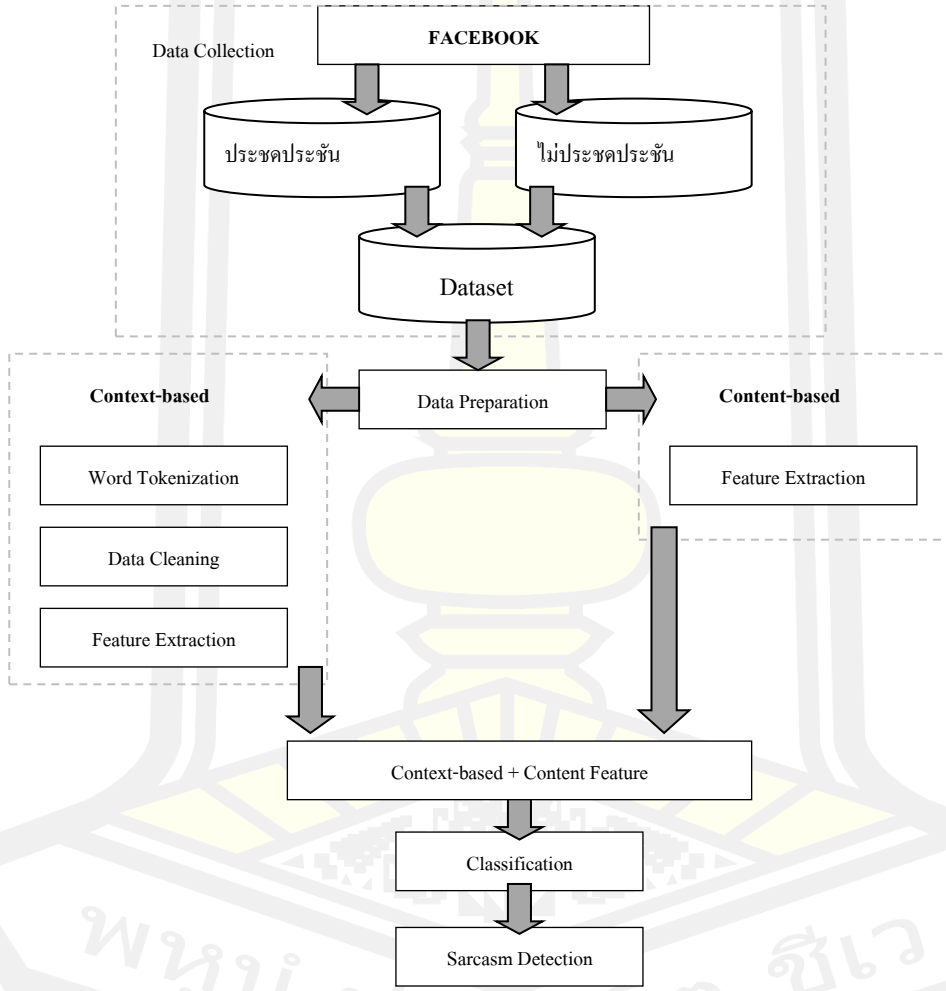
จากการศึกษางานวิจัยที่เกี่ยวข้องจะเห็นได้ว่าการศึกษางานทางด้านการจำแนกข้อความประชดประชันนั้นยังมีความท้าทายและยังมีการศึกษาวิจัยเพื่อเพิ่มประสิทธิภาพในการจำแนกข้อความประชดประชันอย่างต่อเนื่อง ในหลากหลายภาษา ซึ่งโครงสร้างทางด้านภาษาของแต่ละภาษานั้นมีความแตกต่างกัน ทั้งนี้ในภาษาไทยยังไม่มีงานวิจัยทางด้านการตรวจจับข้อความประชดประชันจากข้อความบนเครือข่ายสังคมออนไลน์ ผู้วิจัยจึงให้ความสนใจที่จะศึกษาเพื่อหาระเบียบวิธีในการตรวจจับข้อความประชดประชันภาษาไทยจากข้อความบนเครือข่ายสังคมออนไลน์ คำหรือข้อความประชดประชันเป็นรูปแบบของข้อความที่มีการสื่อสารออกมามีความหมายตรงกันข้ามกับความหมายที่แท้จริง มีความยากในการตรวจจับคำประชดประชันจากข้อความ การใช้ข้อความเชิงบวกในสถานการณ์ทางลบ การใช้ข้อความสั้น จากการศึกษางานวิจัยที่เกี่ยวข้องในงานทางด้านการตรวจจับ/การจำแนกข้อความประชดประชันกับข้อความนั้นยังถือเป็นเรื่องที่ท้าทายและยังถือเป็นเรื่องยากในการที่การเรียนรู้ของเครื่องจะสามารถจำแนกได้ตรงกับการความหมายที่แท้จริง ทั้งนี้ยังมีการศึกษาวิจัยอย่างต่อเนื่อง ซึ่งสามารถสรุปประเด็นต่างๆ ได้ดังนี้ 1) การเลือกชุดข้อมูลส่วนใหญ่เลือกข้อมูลจากเครือข่ายสังคมออนไลน์ เช่น Twitter, Amazon product review, Facebook, Instagram เป็นต้น ภาษาที่มีการศึกษา ภาษาอังกฤษ, ภาษาฮินดู, ภาษาดัช, ภาษาสเปน, ภาษาญี่ปุ่น , ภาษาอินโด, ภาษาเซอเบียร์, ภาษาตุรกี, ภาษาบราซิล เนื่องจากเป็นแหล่งข้อมูลที่สามารถรวบรวมได้ง่ายและมีจำนวนมาก วิธีการในการรวมรวมข้อมูลใช้การค้นหาจากแฮชแท็ก ตัวอย่างเช่น #sarcasm, #sarcastic, #irony, #satire, #not เหตุผลของการใช้แฮชแท็กในการเก็บข้อมูลเพราะจะได้ข้อมูลที่เป็นตัวแทนของคำประชดประชันที่ใช้ในการศึกษางานด้านการจำแนกข้อความประชดประชันได้ดีที่สุดเพราะหากไม่ใช้แฮชแท็กจะได้ชุดข้อมูลที่มีความหลากหลายมากเกินไป ทั้งนี้ขั้นตอนแรกในการเตรียมข้อมูลนั้น ผู้วิจัยได้กำจัดแฮชแท็กออก แล้วให้ผู้เชี่ยวชาญระบุคลาสหรือลาเบลสำหรับคลาสประชดประชัน และไม่ประชดประชัน 2) การเลือกคุณลักษณะเลือกจากคุณลักษณะต่างๆ เหล่านี้ เช่น N-gram, Hashtag, Semantic, Syntactic, POS-tagging, Bag-of-word, Emoticons, Punctuation, Lexical feature, Pattern-based feature, Contextual feature,

Word2Vec 3) การเลือกใช้อัลกอริทึมที่ใช้ในการเรียนรู้ ใช้วิธีจำแนกด้วยการเรียนรู้ของเครื่องแบบมีผู้สอน ด้วยอัลกอริทึม ดังนี้ Logistic Regression/Maximum Entropy, Winnow class, Naïve Bayes, Support Vector Machine, Tree-based, Random Forest, Ensemble, K-nearest Neighbor และการใช้การเรียนรู้เชิงลึก Long Short-Term Memory 4) การวัดประสิทธิภาพในการจำแนก ใช้ 10-fold cross validation เพื่อแบ่งข้อมูลในการวัดประสิทธิภาพ การประเมินผลการจำแนกหรือการทดสอบประเมินผมเพื่อหาความถูกต้องของการจำแนก โดยวัดค่า ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าระลึก, ค่าประสิทธิภาพโดยรวม

นอกจากนี้ยังพบว่า จากการศึกษางานวิจัยที่ผ่านมายังไม่มีการศึกษาการจำแนกข้อความประชดประชันด้วยข้อความภาษาไทย และยังไม่มีการเลือกใช้คุณลักษณะที่รวมกันระหว่างบริบทในข้อความและเนื้อหาในข้อความ ซึ่งการเลือกคุณลักษณะที่ดีนั้นจะช่วยเพิ่มประสิทธิภาพในการจำแนกข้อความประชดประชันได้ดียิ่งขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อการตรวจจับข้อความประชดประชันบนเครือข่ายสังคมออนไลน์ เพื่อให้บรรลุวัตถุประสงค์ผู้วิจัยได้นำเสนอกระบวนการในการดำเนินการวิจัย 5 ขั้นตอน ได้แก่ 1) การรวบรวมข้อมูล (Data Collection) 2) การเตรียมข้อมูล (Data Processing) 3) การสกัดคุณลักษณะ (Feature Extraction) 4) การสร้างตัวจำแนก (Classifier) 5) การวัดประสิทธิภาพการจำแนก (Evaluation) ดังรูปที่ 8 แสดงให้เห็นถึงขั้นตอนของการดำเนินการวิจัยในภาพรวมทั้งหมด



รูปที่ 8 ขั้นตอนในการดำเนินการวิจัย

3.1 การรวบรวมข้อมูล

งานวิจัยนี้ได้เก็บรวบรวมข้อมูลข้อความจากเฟสบุ๊ค ที่มีแฮ็สแท็ก #ประชดประชัน ซึ่งจากการศึกษางานวิจัยที่เกี่ยวข้อง [5, 38, 40, 44, 46, 49] พบว่าการรวมรวมข้อมูลด้านการประชดประชันนั้นใช้วิธีการรวบรวมข้อมูลโดยใช้คำค้นแฮ็สแท็ก ตัวอย่างการวิจัยที่เป็นข้อความภาษาอังกฤษ

ใช้คำค้น #sarcasm, #sarcastic, #irony เป็นต้น งานวิจัยภาษาอื่น เช่น ภาษาดัช [49] ใช้คำค้น #sarcasme ดังนั้นในงานวิจัยนี้จึงใช้วิธีการรวบรวมข้อมูลโดยการใช้แฮ็สแท็ก ซึ่งรวบรวมข้อมูลด้วยแฮ็สแท็กที่มีคำค้นดังนี้ 1) ข้อความที่เป็นข้อความประชดประชันใช้แฮ็สแท็ก #ประชด และ #ประชดประชัน 2) ข้อความไม่ประชดประชันรวมรวมข้อมูลทั่วๆไป ซึ่งใช้หลักการการเก็บข้อมูลรูปแบบเดียวกันกับการเก็บข้อความประชดประชัน โดยประยุกต์ใช้แฮ็สแท็ก #สิ่งดีดี, #คิดดี, มีสุข, ความสุข, #โชคดีจัง ในการระบุคลาสนั้นผู้วิจัยใช้วิธีการระบุคลาสจากแฮ็สแท็ก โดยข้อความที่ได้จากแฮ็สแท็กประชด ถูกระบุคลาสเป็น ประชดประชัน ประกอบด้วยชุดข้อมูลทั้งหมด 5,400 รายการ และข้อความที่ได้จากข้อความทั่วไปที่ไม่ระบุแฮ็สแท็กประชดประชัน ระบุคลาสเป็นไม่ประชดประชัน ประกอบด้วยชุดข้อมูลทั้งหมด 5,400 รายการ ตัวอย่างข้อความดังแสดงในตารางที่ 5

ตารางที่ 5 ตัวอย่างข้อความในชุดข้อมูล

ข้อความ	คลาส
ยุติธรรมจริงๆๆๆๆๆ	Sarcasm
ชื่นชมในความคิดสร้างสรรค์🤣🤣🤣	Sarcasm
เมื่อสั่งให้เพื่อนไปเด็ดกระถินมากินกับส้มตำโถเด็ดมาทีกินได้ทั้งหมู่บ้าน😖😖😖5555	Sarcasm
ย..ยายทองดีเอ้ยนี่ถ้าไปด้วยนะให้ใส่หมวกครอบผมอาบน้ำไปเลย🤣🤣🤣🤣🤣🤣	Sarcasm
อยู่คนเดียวไม่เหงานะจริงๆแล้วการอยู่คนเดียวมันอิสระและโคตรจะมีความสุขเลย😅😅😅	Sarcasm
มีความสุขต้อนรับวันแรกของปีพ่อแม่ลูก	Non-Sarcasm
ยิ้มชิยิ้มมมมมมมีความสุข	Non-Sarcasm
ทำบุญวันปีใหม่มีความสุข	Non-Sarcasm
มีความสุขอย่างบอกไม่ถูกทุกอย่างอยู่ที่ตัวเรา	Non-Sarcasm
มีความสุขทุกครั้งที่ได้อยู่ด้วยขิดื้อน้อย	Non-Sarcasm
ร๊ากน่ะคนดีมีความสุข	Non-Sarcasm
มีความสุขจังถ้าชีวิตมีแต่ความทุกข์ต้องหาความสุขใส่ตัวเรามีความสุขที่ได้อยู่กับครอบครัวลูกหลาน	Non-Sarcasm

3.2 การเตรียมข้อมูล

3.2.1 การสกัดคุณลักษณะจากบริบทในข้อความ

การสกัดคุณลักษณะจากบริบทของข้อความใช้ข้อมูลเฉพาะข้อความที่เป็นข้อความประชดและไม่ประชดจากการเก็บรวบรวมข้อมูล [37] ซึ่งใช้หลักการของการทำเหมืองข้อความ (Text Mining) โดยจะทำการแปลงข้อมูลข้อความให้อยู่ในรูปของเวกเตอร์สเปส (Vector Space Model) ก่อนที่จะเข้าสู่กระบวนการสร้างตัวจำแนก เนื่องจากข้อความเป็นข้อมูลมีคุณลักษณะที่ไม่มีโครงสร้างที่แน่นอน (Unstructured Data) ดังนั้นจะต้องแปลงข้อความให้อยู่ในรูปของข้อมูลที่มีโครงสร้าง (Structural Data) โดยทำการสร้างเวกเตอร์ โดยหนึ่งเวกเตอร์จะแทนหนึ่งเอกสารหรือหนึ่งข้อความ จำนวนมิติของเวกเตอร์ คือ จำนวนคุณลักษณะที่สกัดได้ และค่าที่อยู่ในเวกเตอร์จะคือค่าน้ำหนักของแต่ละคุณลักษณะ โดยขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปของเวกเตอร์มีดังนี้

3.2.1.1 การทำความสะอาดข้อความ

ในกระบวนการทำความสะอาดข้อความจากชุดข้อมูลประชดและไม่ประชด ซึ่งในการวิจัยนี้ศึกษาข้อความภาษาไทย ดังนั้นก่อนเข้าสู่กระบวนการตัด จึงทำการกำจัดภาษาอังกฤษ สัญลักษณ์ต่าง ๆ และตัวเลขเอกออกจากข้อความ สัญลักษณ์ถูกขจัดออกไปเนื่องจากไม่มีนัยสำคัญในการตรวจสอบข้อความ

3.2.1.2 การตัดคำ

งานวิจัยนี้ได้ประยุกต์การตัดคำ 2 วิธีคือ 1) การตัดคำโดยใช้ [63, 64] ซึ่งเป็นการใช้เทคนิคการเรียนรู้เชิงลึกพัฒนามาจากการตัดคำด้วย Deepcut [65] ในการตัดคำภาษาไทยที่เรียนรู้จากคลังคำจากสื่อสังคมออนไลน์ภาษาไทย VISTEC-TP-TH-2021 (VISTEC) ซึ่งประกอบด้วยข้อความจำนวน 49,997 ข้อความ เป็นตัวอย่างข้อความจากทวิตเตอร์ตั้งแต่ปี 2560 – 2562 มีจำนวนประโยค 49,997 ประโยคและมีจำนวนคำทั้งหมด 3.39 ล้านคำ ผลการตัดคำที่ให้ประสิทธิภาพมากที่สุดคือ Deep stacked ensemble (DSE) โมเดล LST20 ให้ประสิทธิภาพการตัดตัวอักษร 99.01% ประสิทธิภาพการตัดคำ 97.33% ดังแสดง

ตารางที่ 6 ผู้วิจัยจึงเห็นว่าเป็นเทคนิคการตัดคำที่เหมาะสมที่สุดที่จะนำมาใช้ในการวิจัยครั้งนี้ ผลการตัดคำดังแสดงได้ดังตารางที่ 7

ตารางที่ 6 ผลการทดลองการตัดคำด้วย OSKut

ที่มา [64]

Method	WS160		TNHC		LST20		TWS21	
	Char	Word	Char	Word	Char	Word	Char	Word
DC	93.47	84.03	89.48	75.40	94.60	84.15	92.77	81.78
AC	93.50	84.04	88.82	73.71	95.24	87.21	91.47	79.31
TL-DC	96.30	90.60	95.43	88.60	98.63	96.30	96.78	90.99
TD-AC	94.10	85.00	90.57	77.54	98.04	94.77	95.47	89.27
SE-DC	95.20	86.90	95.20	84.10	94.96	87.72	94.76	86.33
SE-AC	94.50	85.60	93.70	83.90	96.30	89.87	93.86	84.43
DSE-DC	**96.67**	**91.51**	**95.71**	**89.14**	**99.01**	**97.33**	**97.36**	**92.91**
DSE-AC	94.57	86.24	95.71	88.52	98.46	95.79	97.31	92.78

ตารางที่ 7 ตัวอย่างการตัดภาษาไทย (OSKut Thai Word Segmentation, tl-deepcut-lst20)

ข้อความ	ผลการตัดคำ	คลาส
ยุติธรรมจริงๆๆๆๆๆ	['ยุติธรรม', 'จริง', 'ๆ', 'ๆ', 'ๆ', 'ๆ', 'ๆ']	ประชด
ยิ้มให้กับวันทำงานที่แสนสดใส หลอกตัวเองว่าวันอาทิตย์ได้ม๊ายยย	['ยิ้ม', 'ให้', 'กับ', 'วัน', 'ทำงาน', 'ที่', 'แสน', 'สดใส', ' ', 'หลอก', 'ตัวเอง', 'ว่า', 'วัน', 'อาทิตย์', 'ได้', 'ม๊ายยย']	ประชด
ไม่ต้องการคู่กระทงอันเดียวถือได้	['ไม่', 'ต้องการ', 'คู่', 'กระทง', 'อัน', 'เดียว', 'ถือ', 'ได้']	ประชด
นานแค่ไหนแล้วไม่ได้สนุกขนาดนี้	['นาน', 'แค่', 'ไหน', 'แล้ว', 'ไม่', 'ได้', 'สนุก', 'ขนาด', 'นี้']	ไม่ประชด
ประทับใจในความน่ารักวนไปค่าาาา	['ประทับใจ', 'ใน', 'ความ', 'น่า', 'รัก', 'วน', 'ไป', 'ค่าาาา']	ไม่ประชด

3.2.1.3 การให้ค่าน้ำหนักข้อความ

จำนวนคำที่เหลือจากการขจัดคำหยุดออกจะถูกนำไปใช้เป็นคุณลักษณะ ในงานวิจัยนี้ทำการสกัดคุณลักษณะออกเป็นแบบ unigram ซึ่งคุณลักษณะแต่ละคุณลักษณะคือถุงคำหรือคำที่ตัดได้

จากนั้นทำการสร้างเวกเตอร์จากคุณลักษณะที่สกัดได้ โดย 1 เวกเตอร์ คือ 1 ข้อความ จำนวนคอลัมน์ของแต่ละเวกเตอร์เท่ากับจำนวนคุณลักษณะทั้งหมด ค่าที่อยู่ในแต่ละคุณลักษณะคือค่าน้ำหนักของคุณลักษณะ โดยงานวิจัยนี้ทำการทดลองกำหนดค่าน้ำหนักให้กับคุณลักษณะ 4 วิธี คือ Boolean weighting TF weighting และ TF-IDF weighting เพื่อหาค่าน้ำหนักที่เหมาะสม ตัวอย่างเวกเตอร์ที่ได้จากการใช้คุณลักษณะแบบ unigram และกำหนดค่าน้ำหนักให้คุณลักษณะแบบ Boolean weighting แสดงได้ในตารางที่ 8 โดยค่าน้ำหนักเท่ากับ 1 หมายถึงปรากฏคุณลักษณะในเอกสาร ถ้าค่าน้ำหนัก 0 หมายถึงไม่ปรากฏคุณลักษณะในเอกสาร เช่น ในเอกสาร *d1* ปรากฏคุณลักษณะหรือคำว่า เดิน ดังนั้นค่าน้ำหนักของคุณลักษณะดังกล่าวเท่ากับ 1 แต่ในเอกสาร *d2* ไม่ปรากฏคุณลักษณะ เดิน ดังนั้นค่าน้ำหนักของคุณลักษณะจึงมีค่าเท่ากับ 0 และตัวอย่างเวกเตอร์ที่ได้จากการใช้คุณลักษณะแบบ unigram และกำหนดค่าน้ำหนักให้คุณลักษณะแบบ TF weighting แสดงได้ดังตารางที่ 9 โดยค่าน้ำหนักของแต่ละเวกเตอร์ คือ จำนวนครั้งที่ปรากฏคุณลักษณะ เช่น ในเอกสาร *d1* ปรากฏคุณลักษณะ สุข จำนวน 2 ครั้ง ดังนั้นค่าน้ำหนักของคุณลักษณะดังกล่าวในเอกสาร *d1* มีค่าเท่ากับ 2 และตารางที่ 10 ตัวอย่างเวกเตอร์ที่ใช้คุณลักษณะแบบ unigram และ TF-IDF weighting คือ การคำนวณค่าน้ำหนักของการเกิดคุณลักษณะในเอกสาร *d1* ปรากฏอยู่ในเอกสารอื่น ๆ หรือไม่ เช่น คุณลักษณะ เดิน ปรากฏอยู่ทั้งเอกสาร *d1* และ *d2* ผลการคำนวนดังตารางที่ 10

ตารางที่ 8 ตัวอย่างเวกเตอร์ที่ใช้คุณลักษณะแบบ unigram และ Boolean weighting

Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด
d_1	1	0	1	1	1
d_2	0	1	1	0	0
d_3	1	1	1	1	1

ตารางที่ 9 ตัวอย่างเวกเตอร์ที่ใช้คุณลักษณะแบบ unigram และ TF weighting

Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด
d_1	1	0	1	2	1
d_2	2	2	1	0	0
d_3	0	1	1	1	1

ตารางที่ 10 ตัวอย่างเวกเตอร์ที่ใช้คุณลักษณะแบบ unigram และ TF-IDF weighting

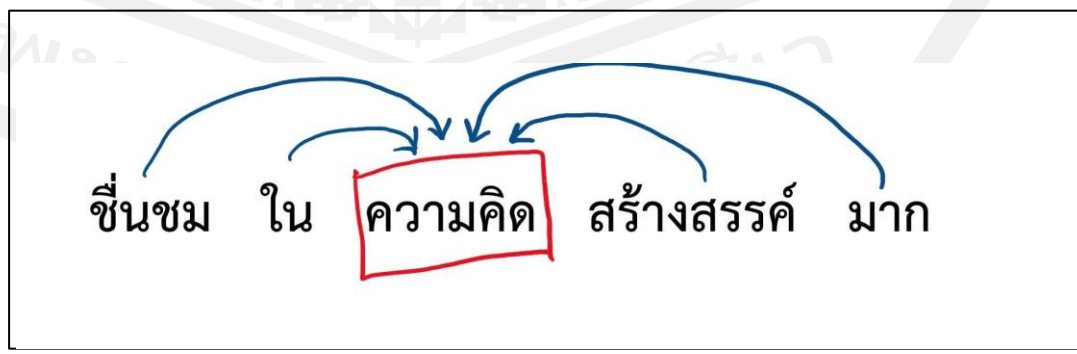
Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด
d_1	0.18	0	0	0.35	0.18
d_2	0.35	0.35	0	0	0

3.2.1.4 การสร้างตัวแทนเชิงความหมายของคำ

การสร้างตัวแทนเชิงความหมายของคำ (Word Embedding) เทคนิคนี้เป็นวิธีการในการแปลงข้อความให้อยู่ในรูปแบบเวกเตอร์ ซึ่งคำหนึ่งคำสามารถแทนค่าคำด้วยเวกเตอร์ที่ทำให้จำนวนมิติลดลง ซึ่งวิธีที่ใช้คือ Word2Vec [66] ซึ่งจะนำมาใช้แก้ปัญหาการแทนความค่าคำด้วยวิธี one-hot encoding จากตารางที่ 11 จะเห็นว่าการ encode ด้วยวิธีนี้ทำให้สิ้นเปลืองหน่วยความจำ และขาดการสื่อความหมาย เช่นคำว่า “ประชด” และ “ประชดประชัน” ต้องถูก encode ด้วยเวกเตอร์สองอันที่ไม่เหมือนกันเลยทั้งที่ความหมายของสองคำนี้มีความหมายเหมือนหรือคล้ายกัน ดังนั้นจึงมีวิธี word embedding เพื่อช่วยแก้ปัญหานี้ด้วยวิธี word2vec ซึ่งจะสร้างเวกเตอร์ของแต่ละคำขึ้นมาและดูความสัมพันธ์ระหว่างคำคำนั้นกับคำที่อยู่รอบข้าง วิธีนี้เป็นการกำหนดมิติของข้อมูลโดยที่ไม่จำเป็นต้องใช้จำนวนมิติของข้อมูลเท่ากับจำนวนคำก็ได้ ซึ่งโดยส่วนมากแล้วจะกำหนดมิติอยู่ที่ระหว่าง 100-300 ซึ่งเพียงพอสำหรับใช้เป็นตัวแทนของคำเพื่อนำไปใช้งานต่อไปได้ สำหรับการวิเคราะห์คำในประโยคโมเดลจะนำเอา word vectors ของคำที่อยู่รอบ ๆ ของคำคำนั้นๆ ภายในระยะของบริบท (context size) ที่กำหนดมาใช้เป็น input สำหรับการจำแนก และใช้คำที่กำลังพิจารณาซึ่งตำแหน่งอยู่ตรงกลางเป็นเป้าหมายในการทำนาย ดังแสดงในรูปที่ 9

ตารางที่ 11 การแทนค่าคำด้วยวิธี one-hot encoding

Index	1	2	...	8999	9000	9001	...	10800
ประชด				1				
ประชดประชัน						1		



รูปที่ 9 การพิจารณาเวกเตอร์ของคำจากบริบทรอบข้าง

3.2.2 การสกัดคุณลักษณะจากเนื้อหาในข้อความ

ในการสกัดคุณลักษณะจากเนื้อหาในข้อความ เป็นการสกัดคุณลักษณะจากข้อความและองค์ประกอบของข้อความ เนื่องจากข้อความประเภทคำประชดประชัน ที่สื่อสารออกมาผ่านข้อความนั้น มีคุณลักษณะที่ผู้สื่อสารแสดงออกมาโดยมีความหมายตรงข้ามกับความหมายที่แท้จริง ดังนั้นในการสกัดคุณลักษณะนี้จึงเป็นการพิจารณาเนื้อหาในข้อความ ซึ่งศึกษาจากงานวิจัย [5] ตัวอย่างข้อความดังแสดงในตารางที่ 17 สำหรับงานวิจัยนี้ใช้คุณลักษณะจำนวน 15 คุณลักษณะ ดังนี้

1. การมีข้อความที่แสดงอารมณ์เชิงบวก (F1) การใช้ความแตกต่างระหว่างขั้วของอารมณ์เชิงบวกอยู่ในข้อความเป็นคุณลักษณะที่สำคัญในการพิจารณาถึงการแสดงออกถึงข้อความประชดประชัน ซึ่งการวิจัยนี้ใช้ข้อความเชิงบวกจาก [67] ดังแสดงในภาคผนวก ก เป็นข้อความในการเปรียบเทียบว่ามีคำที่แสดงอารมณ์เชิงบวก

2. การมีข้อความที่แสดงอารมณ์เชิงลบ (F2) การใช้ความแตกต่างระหว่างขั้วของอารมณ์เชิงลบอยู่ในข้อความเป็นคุณลักษณะที่สำคัญในการพิจารณาถึงการแสดงออกถึงข้อความประชดประชัน ซึ่งการวิจัยนี้ใช้ข้อความเชิงลบจาก [67] เป็นข้อความในการเปรียบเทียบว่ามีคำที่แสดงอารมณ์เชิงบวก ดังแสดงในภาคผนวก ข

3. การใช้คำที่ไม่ปกติ (F3) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่เป็นการใช้คำที่ไม่ปกติหรือไม่ใช้ข้อความที่เขียนอย่างถูกต้องตามพจนานุกรม ตัวอย่างคำที่ไม่ปกติที่นำมาพิจารณา เช่น รักกกกกกกกกกก การใช้ตัวอักษรที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่าเป็นการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป ตัวอย่างคุณลักษณะนี้ดังแสดงในตารางที่ 12

ตารางที่ 12 ตัวอย่างการใช้คำที่ไม่ปกติ

ตัวอย่างข้อความ	ตัวอย่างข้อความ	ตัวอย่างข้อความ
มากกกกก	ยังงงงงง	เลยยยยยย
เหรอออออออ	แล้วววววว	เรียบร้อยยยยย

4. การใช้ปรัศนี (?) (F4) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายปรัศนีที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป ตัวอย่างเช่น สงสัย? หรือ สวยจัง??????? หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

5. การแสดงคำหัวเราะ (F5) จำนวนของการใช้คำแสดงการหัวเราะในประโยคอาจมีส่วนในการบอกถึงการแสดงออกซึ่งการประชดประชัน หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

6. การมีคำที่แสดงถึงการประชดประชันโดยทั่วไป (F6) การมีคำที่เป็นคำเฉพาะที่บ่งบอกได้ถึงการแสดงออกอย่างชัดเจนถึงการประชดประชัน หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0 ตัวอย่างดังแสดงในตารางที่ 13

ตารางที่ 13 ตารางแสดงคำประชดประชันโดยทั่วไป

เล้ย	แหม	แหม่	เอ่อ	เหรออ	แร้ว
เอิ่ม	ป่ะว่ะ	เกิ๊น	เก่งเกิ๊น	จริ้ง	เล๊ย
ถนั๊ด	จริงจริ๊ง	เร้ย	ดี๊ดี	ดีออก	เก่งเกิ้น
เก่งมากจ้า	เปล่าไม่มีไร	ดีมากจ้า	สนุกสุดสุด	เหอะตลก	เหอะๆ
เหอะ	เห้อ	เชอะ	ใช่สิ๊	เถอะ	เอาจ้าา
เอาไปเถอะ	ช่างแม่ง	แล้วแต่จ้า	ไปไหนก็ไป	เราผิดเอง	เรามันไม่ดีเอง
ตามสบายเลย	เอาที่สบายใจ	เอาเล๊ยย	อือจ้า	งี้แหละ	กุไปเอง
ไม่มาปีหน้าอะ	ไม่ต้องรีบกุรออยู่	อร่อยมากกก	สนใจกุดี	ก็งี้อ่ะ	ไม่ต้องรักหรอก
ไม่ต้องสนใจ	ไม่ต้องหรอก	โสดมากจ้า	โสดเว้ยยย	โสดตัวเท่าบ้าน	โสดตัวเท่าควาย
โสดก็ดี๊	สวยดีเนอะ	สวยจ้าสวย	ไม่เห็นจะแคร์	อยู่ได้ไปเถอะ	คิดว่าดีก็ทำไป
มีความสุขก็ทำไป	มีความสุขดี	เค้าก่อน	นานเกิน	ก็ดี๊	ก็งี้
ก็งี้แหละ	ไม่เป็นไรจ่ะพี่รอได้	ยอมโง่	รอได้จริงๆ	ขอให้รักกันนานๆ	เข้าข้างกูชิบหาย
โชคดี	ไม่ต้องมายุ่ง	ผิดตลอดแหละ	แพ้ตลอด	สู้เขาไม่ได้	ตามสบาย

ตารางที่ 13 ตารางแสดงคำประชดประชันโดยทั่วไป (ต่อ)

เชิญเลยจ้า	เชิญเล้ย	ไม่ต้องมาหรอก	คนดีมาก	ดี๊ดีมาก	จะร้องไห้เลยกู
เอาเลย!	เต็มที่จ้า	แดกตามสบาย	ไม่ต้องเกรงใจ	หนีไปไกลๆ	ไม่มีอะไรจะคุย

เราอยู่ได้	เราโอเค	โถ่	โถ่เว้ย	เว้ยเฮ้ย	

7. การใช้เครื่องหมายอัศเจรีย์ (!)(F7) ซึ่งโดยทั่วไปใช้เป็นเครื่องหมายที่ใช้เขียนไว้ข้างหลังคำหรือกลุ่มคำที่แสดงอารมณ์และความรู้สึกต่าง ๆ เช่น เสียใจ! ตกใจ! คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายอัศเจรีย์ที่มากเกินไป ตัวอย่าง เช่น รัก!!!!!!!! การใช้เครื่องหมายอัศเจรีย์ที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป เช่น รัก! หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

8. การใช้มหัพภาค (.) (F8) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายจุลภาคที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

9. คำหยาบ (F9) การใช้คำหยาบในประโยคอาจมีส่วนในการบอกถึงการแสดงออกซึ่งการประชดประชัน ซึ่งการมีคำหยาบคาบในประโยคเมื่อรวมกับคำอื่นอาจมีความหมายที่ตรงข้าม หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0 ซึ่งการวิจัยนี้ใช้ข้อความคำหยาบ [67] ดังแสดงในตารางที่ 14 ทั้งนี้เพื่อความเหมาะสมในการแสดงข้อความจึงขอแสดงตัวอย่างเฉพาะบางคำเท่านั้น

ตารางที่ 14 ตารางแสดงคำหยาบคาย

ขี้	เยี่ยว	หมา
ฟาย	หมาบ้า	มัจจุราช
ส้นตีน	ชาติชั่ว	สถุล
ควาย	แรด	กระทิง
เควี่ย	เชี่ย	แมงดา
ระยำ	สันดาน	ทมิฬ

10. การใช้บุพสัญญา (”) (F10) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายจุลภาคที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

11. การใช้ไม้ยมก (ๆ) (F11) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายจุลภาคที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

12. ไอคอนแสดงอารมณ์เชิงลบ (F12) การใช้ความแตกต่างระหว่างขั้วของอารมณ์ กับไอคอนแสดงอารมณ์ อาจบ่งบอกได้ถึงการแสดงออกถึงการประชดประชัน เช่น เธอช่างน่ารักจริง ๆ 😖 โดยเลือกใช้รูปแบบไอคอนแสดงอารมณ์ที่สามารถใช้ได้บนเฟสบุ๊ค ตัวอย่างดังแสดงในตารางที่ 15 หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

ตารางที่ 15 แสดงไอคอนแสดงอารมณ์เชิงลบ

😶	😐	😑	😒
😥	😞	😟	😠
😠	😔	😕	😣
😖	😭	😩	😭
😮	😱	😦	😢
😯	😦	😧	😰
😢	😪	😓	😱
😵	😲	😂	😴
👎			

13. ไอคอนแสดงอารมณ์เชิงบวก (F13) การใช้ความแตกต่างระหว่างขั้วของอารมณ์ กับไอคอนแสดงอารมณ์ อาจบ่งบอกได้ถึงการแสดงออกถึงการประชดประชัน เช่น เธอช่างน่ารักจริง ๆ 😁โดยเลือกใช้รูปแบบไอคอนแสดงอารมณ์ที่สามารถใช้ได้บนเฟสบุ๊ค ดังแสดงในตารางที่ 16

ตารางที่ 16 แสดงไอคอนแสดงอารมณ์เชิงบวก

😀	😁	😄	😂
😃	😆	😅	😆
😇	😊	☺	🙂
😋	😌	😍	😘
😗	😙	😚	😜
😝	😛	😎	😏
👍			

14. การใช้เครื่องหมายบวก (+) (F14) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายบวกที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

15. การใช้เครื่องหมายลบ (-) (F15) คุณลักษณะดังกล่าวถูกนำมาใช้พิจารณาข้อความประชดประชัน เนื่องจากสมมติฐานที่ว่า ข้อความที่มีการใช้เครื่องหมายลบที่มากเกิน 5 ตัวอักษร ทำให้มีเหตุที่สื่อได้ว่ามีความตั้งใจในการแสดงออกซึ่งข้อความประชดประชันเพราะเป็นการสื่อสารที่เป็นเกิดจากความตั้งใจที่ผิดปกติจากการใช้ข้อความเพื่อการสื่อสารทั่ว ๆ ไป หากมีคุณลักษณะนี้ให้แทนค่าข้อความตามคุณลักษณะนี้เป็น 1 หากไม่มีให้กำหนดเป็น 0

ตารางที่ 17 ตัวอย่างข้อความจากการสกัดคุณลักษณะจากเนื้อหาในข้อความ

Text	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	Class
เราจะกินให้หมดนี้เลยเทอทำให้เราโกด!!!!สำนึกไว้เลยนะะะะเทอทำเราอ้วน	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	1
โอ้ยยยยแบบนี้ไม่ต้องจัดสอบให้เสียเวลาจ้างครูมาสอนต่อเถอะคะ	0	1	1	0	0	1	0	0	0	0	1	0	0	0	0	1
นอกจากเป็นแม่ทูลหัวแล้วยังเป็นแม่บ้านด้วยนะรับสักที่มั้ยคะมีความสุขอยู่กับอาหาร	1	1	0	0	1	0	1	0	0	1	0	0	0	0	0	2
ภาพบอกความรู้สึกก็จะประมาณนี้แหล่ะค้าการเริ่มต้นใหม่ที่ดีมีความสุข	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

จากการสกัดคุณลักษณะจากเนื้อหาในข้อความ สามารถสรุปการเลือกคุณลักษณะในส่วนนี้ได้ว่า คุณลักษณะที่เลือกใช้เป็นคุณลักษณะที่มีนัยสำคัญที่บ่งบอกได้ว่าข้อความประชดประชันส่วนมาก มีการสื่อออกมาเป็นข้อความในรูปแบบที่ไม่ปกติ หรือเกิดจากความตั้งใจสื่อสารออกมาให้ไม่ปกติจากผู้สื่อสาร เพื่อบ่งบอกว่าข้อความเหล่านี้มีนัยเป็นการประชดประชดประชัน ซึ่งอาจบอกได้ว่าข้อความประชดประชันนั้นมีรูปแบบ [5] การสื่อสารที่แตกต่างจากข้อความปกติ ซึ่งการเลือกกำหนดจำนวนคุณลักษณะที่นำมาใช้เป็นตัวแทนของเอกสารที่เป็นข้อความประชดเพียงแค่ 15 คุณลักษณะ

ดังกล่าวนี้ อาจจะเป็นข้อจำกัดในการจำแนกข้อความไม่ประชดประชันเพราะคุณลักษณะเหล่านี้ถูกกำหนดใช้เป็นตัวแทนของข้อความประชดประชันในชุดข้อมูลการทดลองในการวิจัยนี้เท่านั้น

3.2.3 คุณลักษณะการรวมบริบทในข้อความและเนื้อหาในข้อความ

การใช้วิธีการรวมคุณลักษณะระหว่างคุณลักษณะที่สกัดจากบริบทในข้อความรวมกับคุณลักษณะที่สกัดจากเนื้อหาในข้อความ สามารถรวมกันได้ดังนี้ 1) การรวมกันระหว่าง Boolean Weighting และ 16 คุณลักษณะจากเนื้อหาในข้อความ ดังแสดงในตารางที่ 18 2) การรวมกันระหว่าง TF Weighting และ 16 คุณลักษณะจากเนื้อหาในข้อความ 3) การรวมกันระหว่าง TF-IDF Weighting และ 16 คุณลักษณะจากเนื้อหาในข้อความ ซึ่งข้อจำกัดของการรวมกันทั้งสามข้อคืออาจทำให้มีความโน้มเอียงในการประมวลผล เนื่องจากเป็นการรวมกันของข้อมูลที่มีรูปแบบไม่เหมือนกันสำหรับการรวมกันของรูปแบบที่ 2 และ 3

ตารางที่ 18 ตัวอย่างข้อมูลคุณลักษณะจากบริบทในข้อความด้วย Boolean Weighting รวมกับคุณลักษณะที่สกัดจากเนื้อหาในข้อความ

Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด	Positive Sentiment	Negative Sentiment	Positive Emoticon	Negative Emoticon
d_1	1	0	1	1	1	1	0	0	1
d_2	0	1	1	0	0	1	1	1	1
d_3	1	1	1	1	1	0	1	1	0

ตารางที่ 19 ตัวอย่างข้อมูลคุณลักษณะจากบริบทในข้อความด้วย TF Weighting รวมกับคุณลักษณะที่สกัดจากเนื้อหาในข้อความ

Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด	Positive Sentiment	Negative Sentiment	Positive Emoticon	Negative Emoticon
d_1	1	0	1	2	1	1	0	0	1
d_2	2	2	1	0	0	1	1	1	1
d_3	0	1	1	1	1	0	1	1	0

ตารางที่ 20 ตัวอย่างข้อมูลคุณลักษณะจากบริบทในข้อความด้วย TF-IDF Weighting รวมกับคุณลักษณะที่สกัดจากเนื้อหาในข้อคววาม

Documents	เดิน	ทาง	เหนื่อย	สุข	ที่สุด	Positive Sentiment	Negative Sentiment	Positive Emoticon	Negative Emoticon
d_1	0.18	0	0	0.35	0.18	1	0	0	1
d_2	0.35	0.35	0	0	0	1	1	1	1

3.3 การสร้างตัวจำแนก

3.3.1 การสร้างตัวจำแนกด้วยวิธีการเรียนรู้ด้วยเครื่อง

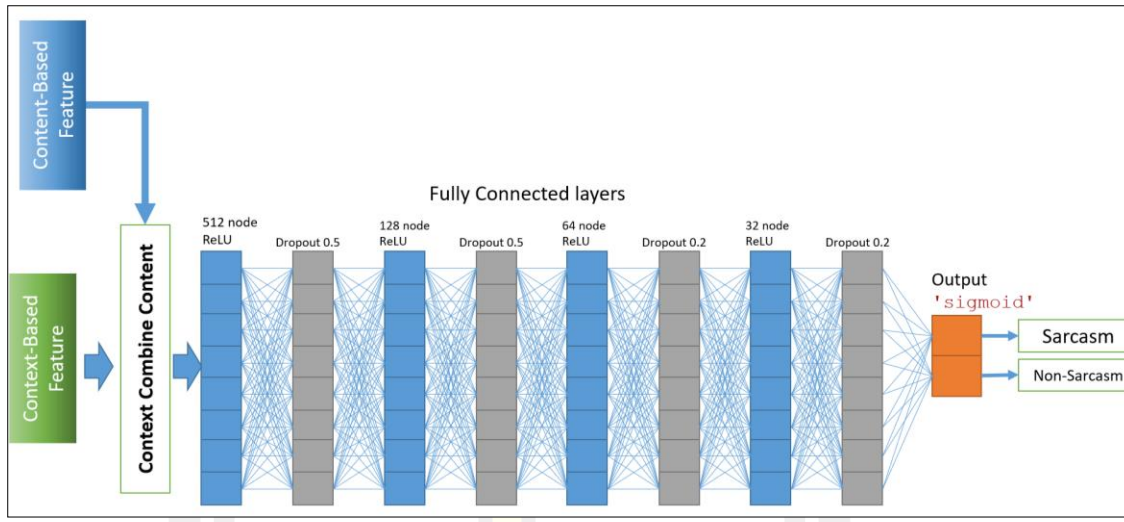
จากการศึกษางานวิจัยที่เกี่ยวข้องทำให้ทราบถึงความนิยมในการเลือกใช้ตัวจำแนกที่มีความนิยมและมีประสิทธิภาพในการจำแนก ทั้งนี้ผู้วิจัยได้เลือกใช้ตัวจำแนกโดยเลือกมา 4 ตัวจำแนก ดังต่อไปนี้ 1) นาอีฟเบย์ 2) ซัพพอร์ทเวกเตอร์แมคชีน 3) เพื่อนบ้านใกล้ที่สุด 4) ต้นไม้ตัดสินใจ เนื่องจากเป็นอัลกอริทึมที่นักวิจัยด้านการตรวจจับข้อความประชดประชันส่วนมากนำมาใช้ในการจำแนก ซึ่งจะนำมาเปรียบเทียบกับประสิทธิภาพ การใช้เฉพาะคุณลักษณะจากบริบทในข้อความและการรวมคุณลักษณะจากบริบทในข้อความกับคุณลักษณะจากเนื้อหาในข้อความ ซึ่งในอัลกอริทึมแต่ละตัวเลือกใช้พารามิเตอร์ในการจำแนกของแต่ละวิธีดังนี้

1. ขั้นตอนวิธีซัพพอร์ทเวกเตอร์แมคชีน เลือกใช้ linear kernel ในการเรียนรู้สำหรับการสร้างตัวจำแนก
2. ขั้นตอนวิธีนาอีฟเบย์ ไม่มีการเลือกใช้พารามิเตอร์ในวิธีนี้ เนื่องจากไม่มีการปรับพารามิเตอร์
3. ขั้นตอนวิธีต้นไม้ตัดสินใจ ใช้อัลกอริทึม ID3 เลือกใช้ Gini เป็นพารามิเตอร์สำหรับเกณฑ์ในการสร้างโครงสร้างการตัดสินใจ ในการเรียนรู้สำหรับการสร้างตัวจำแนก
4. ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด เลือกใช้ Euclidean distance และกำหนดค่า K แตกต่างกันในการเรียนรู้สำหรับการสร้างตัวจำแนก

3.3.2 การสร้างตัวจำแนกด้วยวิธีการเรียนรู้เชิงลึก

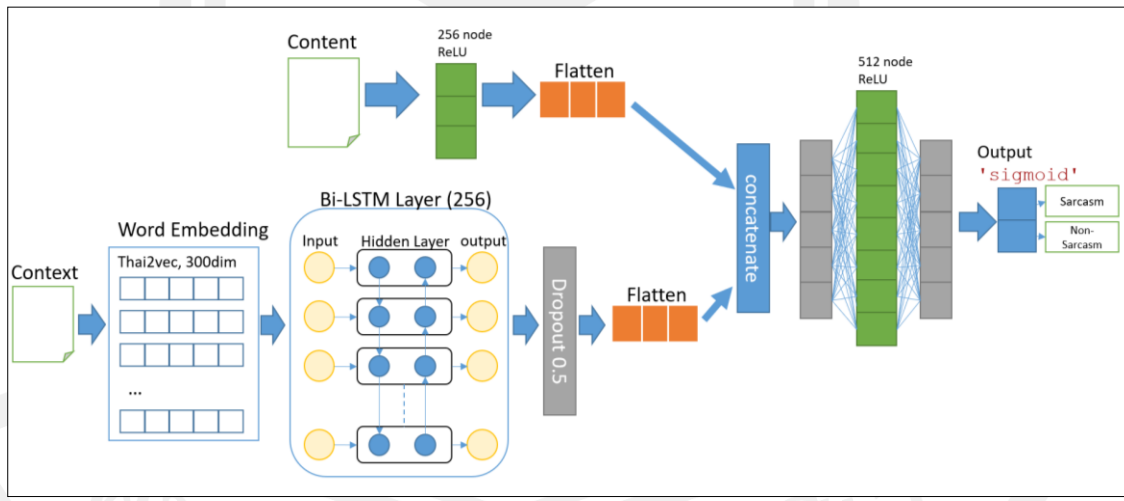
จากการศึกษางานวิจัยที่เกี่ยวข้อง [32, 33, 68] ทำให้ทราบถึงความนิยมในการเลือกใช้ตัวจำแนกด้วยเทคนิควิธีการเรียนรู้เชิงลึก ในการวิจัยนี้ผู้วิจัยเลือกใช้อัลกอริทึม 2 อย่างคือ Deep Neural Network และ Long Short-Term Memory ซึ่งในอัลกอริทึมแต่ละตัวเลือกใช้พารามิเตอร์ในการจำแนกของแต่ละวิธีดังนี้

1. ขั้นตอนวิธี Deep Neural Network (DNN) ปรับค่าพารามิเตอร์ในการเรียนรู้ดังนี้ Dense = 512, Activation Function = ReLU, Dropout = 0.2, Output Activation = Sigmoid, loss Function = binary_crossentropy และกำหนด Optimizer = adam



รูปที่ 10 แสดงแผนภาพโมเดล DNN

2. ขั้นตอนวิธี Long Short-Term Memory (BiLSTM) ปรับค่าพารามิเตอร์ในการเรียนรู้ ดังนี้ เลือกใช้การเรียนรู้ Bidirectional LSTM จำนวน 256 node, Activation Function = ReLU, Dropout = 0.2, Output Activation = Sigmoid, loss Function = binary_crossentropy และกำหนด Optimizer = adam



รูปที่ 11 แสดงแผนภาพโมเดล Bi-LSTM

3.4 การวัดประสิทธิภาพ

ขั้นตอนในการวัดประสิทธิภาพการจำแนก เป็นขั้นตอนในการประเมินความสามารถในการจำแนกข้อมูลของตัวจำแนก งานวิจัยนี้ พิจารณาใช้ค่า ความถูกต้อง ความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวม ใช้ 10-fold cross validation ในการแบ่งชุดข้อมูลการเรียนรู้ (Training set) และชุดข้อมูลการทดสอบ (Testing set) ซึ่งจะแบ่งข้อมูลออกเป็น 10 ชุดข้อมูลในจำนวนเท่าๆ

กัน และจะทำการประเมินประสิทธิภาพแต่ละรอบ ในรอบที่ 1 ข้อมูลชุดที่ 1 จะถูกนำมาใช้เป็นข้อมูลชุดทดสอบเพื่อวัดประสิทธิภาพของตัวจำแนก ข้อมูลชุดที่ 2 ถึง 10 จะถูกนำมาเป็นชุดข้อมูลการเรียนรู้ ในรอบที่ 2 ข้อมูลชุดที่ 2 จะถูกนำมาใช้เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่ 6 และ ข้อมูลชุดที่ 3 ถึง 10 จะถูกนำมาเป็นข้อมูลชุดการเรียนรู้ และในแต่ละรอบจะทำการวัดประสิทธิภาพ ดังนี้ โดยพิจารณาจากตาราง Confusion Matrix ดังตารางที่ 21

ตารางที่ 21 Confusion Matrix

		Predicted	
		Sarcasm	Non-Sarcasm
Actual	Sarcasm	*a*	*b*
	Non-Sarcasm	*c*	*d*

เมื่อ *a* คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นข้อความประชดประชัน

b คือ จำนวนข้อมูลที่ทำนายว่าไม่เป็นข้อความประชดประชัน แต่คำตอบคือ เป็นข้อความประชดประชัน

c คือ จำนวนข้อมูลที่ทำนายว่าเป็นข้อความประชดประชัน แต่คำตอบคือ ไม่เป็นข้อความประชดประชัน

d คือ จำนวนข้อมูลที่ทำนายถูกว่าไม่เป็นข้อความประชดประชัน

1. การวัดค่าความถูกต้องในการจำแนกโดยรวม (Accuracy) ในการจำแนกข้อความประชดประชัน และไม่ประชดประชัน ซึ่งสามารถคำนวณได้ดังสมการที่ (22)

$$Accuracy = \frac{(a+d)}{(a+b+c+d)} \tag{22}$$

2. การวัดค่าความแม่นยำ (Precision) ในการจำแนกข้อความประชดประชัน ซึ่งสามารถคำนวณได้ดังสมการที่ (23)

$$\Pr ecision_{sarcasm} = \frac{a}{(a+c)} \tag{23}$$

3. การวัดค่าความแม่นยำ ในการจำแนกข้อความไม่ประชดประชัน ซึ่งสามารถคำนวณได้ดังสมการที่ (24)

$$\Pr ecision_{non-sarcasm} = \frac{d}{(b+d)} \tag{24}$$

4. การวัดค่าความระลึก (Recall) ในการจำแนกข้อความประชดประชัน ซึ่งสามารถคำนวณได้ดังสมการที่ (25)

$$\mathrm{Re}\,call_{sarcasm} = \frac{a}{(a+b)} \tag{25}$$

5. การวัดค่าความระลึก (Recall) ในการจำแนกข้อความไม่ประชดประชัน ซึ่งสามารถคำนวณได้ดังสมการที่ (26)

$$\mathrm{Re}\,call_{non-sarcasm} = \frac{d}{(c+d)} \tag{26}$$

6. การวัดค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure) เป็นการพิจารณานำเอาค่าความระลึกและค่าความแม่นยำของคลาสประชดประชันและไม่ประชดประชัน มาพิจารณารวมกัน ระบบที่มีประสิทธิภาพจะต้องมีค่าความแม่นยำและค่าความระลึกสูงใกล้เคียงกัน ดังสมการที่ (27) และ (28) ตามลำดับ

$$F-measure_{sarcasm} = 2 \times \frac{\mathrm{Pr}\,ecision_{sarcasm} \times \mathrm{Re}\,call_{sarcasm}}{\mathrm{Pr}\,ecision_{sarcasm} + \mathrm{Re}\,call_{sarcasm}} \tag{27}$$

$$F-measure_{non-sarcasm} = 2 \times \frac{\mathrm{Pr}\,ecision_{non-sarcasm} \times \mathrm{Re}\,call_{non-sarcasm}}{\mathrm{Pr}\,ecision_{non-sarcasm} + \mathrm{Re}\,call_{non-sarcasm}} \tag{28}$$

บทที่ 4
ผลการวิจัยและการอภิปรายผล

การทดลองการวิจัยนี้ เป็นการทดลองการใช้การสกัดคุณลักษณะจากบริบทในข้อความ และคุณลักษณะจากเนื้อหาในข้อความ การให้ค่าน้ำหนักคำคุณลักษณะจากบริบทในข้อความด้วย Boolean weighting TF Weighting TF-IDF Weighting และ Word Embedding ซึ่งทำการทดลอง 3 รูปแบบ คือ 1) ทดลองด้วยคุณลักษณะจากบริบทในข้อความอย่างเดียว 2) ทดลองด้วยคุณลักษณะจากเนื้อหาในข้อความอย่างเดียว และ 3) ทดลองด้วยการการรวมคุณลักษณะระหว่างบริบทในข้อความและเนื้อหาในข้อความ และเปรียบเทียบด้วยเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน 4 เทคนิคคือ 1) อัลกอริทึมซัพพอร์ทเวกร์เตอร์แมคชีน 2) อัลกอริทึมนาอีฟเบย์ 3) อัลกอริทึมเพื่อนบ้านใกล้ที่สุด 4) อัลกอริทึมต้นไม้ตัดสินใจ และเทคนิคการเรียนรู้เชิงลึกใช้การเรียนรู้ 2 อัลกอริทึมคือ 1) DNN และ 2) BiLSTM ซึ่งแสดงผลการทดลอง ได้ดังนี้

4.1 เครื่องมือและข้อมูลที่ใช้ในการทดลอง

4.1.1 เครื่องมือที่ใช้ในการทดลอง

เครื่องมือที่ใช้ในการทดลองในการวิจัย ได้แก่ ด้านฮาร์ดแวร์ ประกอบด้วย เครื่องคอมพิวเตอร์แบบพกพา Apple MacBook Pro 16-inch, 2019 ซีพียู 2.3 GHz 8-Core Intel Core i9 แรม 16 GB ด้านซอฟต์แวร์และภาที่ใช้ในการเขียนการทดลอง ได้แก่ ระบบปฏิบัติการ macOS Moterey 12.1.1 64-bit พัฒนาซอฟต์แวร์โดยใช้ภาษาไพธอน (Python) บนบริการระบบคลาวด์ของบริษัท Google (Google Colaboratory Pro) และใช้ Library ด้าน Machine Learning ได้แก่ Tensorflow เวอร์ชั่น 2.8.0

4.1.2 ผลการรวบรวมข้อมูลในการทดลอง

การรวบรวมข้อมูลผู้วิจัยรวบรวมข้อมูลที่ใช้ในการวิจัยนี้จากเครือข่ายสังคมออนไลน์เฟสบุ๊ค จำนวน 1 ชุดข้อมูล ข้อมูลทั้งหมดถูกรวบรวมด้วยการใช้คำค้นด้วยแฮ็สแท็กจำนวนทั้งหมด 10,800 ข้อความ โดยแบ่งเป็นข้อความ 1) ข้อความประชดประชัน ใช้แฮ็สแท็ก #ประชด และ #ประชดประชัน จำนวนข้อมูล 5,400 ข้อความ และ 2) ข้อความไม่ประชดประชัน ใช้แฮ็สแท็ก #สิ่งดีดี, #คิดดี, มีสุข, ความสุข, #โชคดีจัง จำนวนข้อมูล 5,400 ข้อความ จำนวนข้อมูลทั้งหมดที่ใช้ในการทดลองแสดงตารางที่ 22

ตารางที่ 22 ลักษณะข้อมูลที่ในในงานวิจัย

ข้อมูล	**จำนวนข้อความ**
ข้อความประชดประชัน	5,400

ตารางที่ 22 ลักษณะข้อมูลที่ในในงานวิจัย (ต่อ)

ข้อความไม่ประชดประชัน	5,400
ความยาวต่ำสุดของข้อความ (คำ)	2
ความยาวสูงสุดของข้อความ (คำ)	252
ความยาวเฉลี่ยของข้อความ (คำ)	16

4.2 ผลการทดลองจำแนกข้อมูลจากบริบทในข้อความ

งานวิจัยนี้ได้ทำการคัดเลือกจำนวนข้อความที่เป็นข้อความประชดประชันทั้งหมด 5,400 รายการและไม่ประชดประชันทั้งหมด 5,400 ข้อความ จากข้อความบนเฟสบุ๊คโดยใช้แฮชแท็ก # ประชด และ #ประชดประชัน และข้อความไม่ประชดประชันจากข้อความบนเฟสบุ๊คโดยใช้แฮชแท็ก #สิ่งดีดี #คิดดี #มีสุข #ความสุข #โชคดีจัง เมื่อได้ชุดข้อมูลแล้วทำความสะอาดข้อมูลและทำการตัดข้อมูลด้วยวิธีการตัดคำ และแปลงข้อมูลให้อยู่ในรูปแบบเวกเตอร์จะได้เวกเตอร์ที่มีคุณลักษณะจำนวน 12,973 คุณลักษณะ โดยในชุดข้อมูลมีจำนวนคำที่น้อยที่สุดคือ 2 คำ และจำนวนคำที่มากที่สุดคือ 252 คำ จากนั้นกำหนดค่าคุณลักษณะแต่ละคุณลักษณะโดยทำการคำนวณค่าน้ำหนักของคุณลักษณะด้วยวิธีการ 4 วิธี คือ Boolean weighting, TF weighting TF-IDF weighting และ Word Embedding จากนั้นการทดลองจำแนกข้อมูลกับชุดข้อมูลกับตัวจำแนกที่สร้างขึ้นด้วย 1) เทคนิคการเรียนรู้ของเครื่องโดยใช้อัลกอริทึม ซัพพอร์ทเวกเตอร์แมคชีน, นาอีฟเบย์, ต้นไม้ตัดสินใจ และ เพื่อนบ้านใกล้เคียงที่สุด 2) เทคนิคการเรียนรู้เชิงลึกด้วย DNN และ BiLSTM และทำการวัดประสิทธิภาพโดยใช้ 10-fold cross validation ในการแบ่งข้อมูลเรียนรู้และทดสอบ และวัดค่าความถูกต้องเฉลี่ย ค่าความแม่นยำเฉลี่ย ค่าระลึกเฉลี่ย ค่าอัตราการเรียนรู้เฉลี่ยสำหรับแต่ละคลาสและเวลาในการประมวลผล ซึ่งผลการทดลองแสดงได้ดังนี้

4.2.1 การทดลองเปรียบเทียบ Remove Stop Words และ ไม่ Remove Stop Words

การทดลองเปรียบเทียบการลบคำที่ไม่สำคัญออกและการไม่ลบคำไม่สำคัญออก ซึ่งโดยปกติแล้วการลบคำที่ไม่สำคัญออกนั้นทำให้คุณลักษณะของข้อมูลลดลง ช่วยทำให้การประมวลผลเร็วขึ้นและให้ประสิทธิภาพในการทดลองไม่แตกต่างจากเดิม จากการทดลองนี้แสดงผลการทดลองด้วยเทคนิคการเรียนรู้ของเครื่อง โดยการใช้คุณลักษณะจากบริบทในข้อความ ด้วยการวัดค่าน้ำหนักด้วยวิธี Boolean Weighting เพื่อแสดงการทดลองการลบคำที่สำคัญและการไม่ลบคำที่ไม่สำคัญ จากการทดลองแสดงให้เห็นว่าการไม่ลบคำสำคัญออกทำให้ได้ค่าความถูกต้องในการทดลองมาขึ้นเกือบทุกเทคนิค มีเพียงเทคนิค KNN เท่านั้นที่การลบคำสำคัญให้ค่าความถูกต้องมากกว่าการไม่ลบคำสำคัญ

ผลการทดลองดังแสดงในตารางที่ 23 ดังนั้นในการทดลองในการวิจัยนี้จึงเลือกใช้การไม่ลบคำสำคัญออก

ตารางที่ 23 การทดลองเปรียบเทียบการ Remove Stop Words และ ไม่ Remove Stop Words

	Accuracy	Precision		Recall		F-measure	
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm
With Stop words							
KNN	71.72	97.86	45.59	64.26	95.54	77.56	61.68
SVM	89.31	92.95	85.68	86.66	92.39	89.69	88.90
Decision-Tree	84.94	89.92	79.96	81.78	88.82	85.65	84.15
Naïve Bayes	86.38	85.58	87.19	87.02	85.77	86.28	86.46
DNN	89.38	89.27	89.49	89.27	89.49	89.27	89.49
LSTM	92.96	90.98	87.83	88.18	90.75	89.55	89.26
Remove Stop words							
KNN	75.78	86.70	64.83	71.16	83.03	78.14	72.77
SVM	86.64	90.60	82.71	83.95	89.76	87.14	86.08
Decision-Tree	82.44	87.19	77.71	79.61	85.86	83.21	81.56
Naïve Bayes	85.91	84.43	87.47	87.04	84.86	85.69	86.11
DNN	87.81	88.90	86.74	86.78	88.86	87.83	87.79
LSTM	91.39	90.45	89.58	89.70	90.38	90.03	89.94

4.2.2 การทดลองการกำหนดจำนวน K ที่ดีที่สุดสำหรับ KNN

จากการเลือกใช้เทคนิค KNN การสร้างแบบจำลองด้วยการเรียนรู้ของเครื่อง เพื่อหาจำนวน K ที่เหมาะสมที่สุด จึงทดลองเปรียบเทียบจำนวนการเลือกใช้ K เท่ากับ 3, 5, 7 และ 9 โดยการใช้คุณลักษณะจากบริบทในข้อความ ด้วยการวัดค่าน้ำหนักด้วยวิธี Boolean Weighting ดังตารางที่ 24 จากผลการทดลองแสดงให้เห็นว่าจำนวน K ที่ให้ค่าความถูกต้องมากที่สุดคือ K เท่ากับ 5 ดังนั้นในการวิจัยนี้จึงเลือกใช้จำนวน K = 5 กับการทดลองสำหรับ KNN

ตารางที่ 24 ผลการทดลองเปรียบเทียบการเลือกใช้จำนวน K สำหรับเทคนิค KNN

	Accuracy	Precision		Recall		F-measure		times
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	
KNN=3	71.58	98.60	44.58	64.02	96.96	77.61	61.02	44s
KNN=5	71.72	97.86	45.59	64.26	95.54	77.56	61.68	61s
KNN=7	69.40	98.70	40.11	62.23	96.87	76.31	56.70	46s
KNN=9	67.47	99.02	35.92	60.71	97.36	75.26	52.43	47s

4.2.3 ผลการทดลองด้วยคุณลักษณะ Boolean Weighting

จากผลการทดลองในตารางที่ 25 จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้อง 91.39% ด้วยค่า std = 0.03 นอกจากนี้ยังให้ค่าระลึกดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าระลึกเท่ากับ 89.70% ส่วน KNN ให้ค่าความแม่นยำดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าความแม่นยำในการทำนายข้อความประชดประชันถึง 97.67% แต่เมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุด คือ 89.94% ส่วนขั้นตอนวิธี Naïve bayes ใช้เวลาในการประมวลผลน้อยที่สุด กล่าวโดยสรุปขั้นตอนวิธี LSTM ให้ประสิทธิภาพโดยรวมในการทำนายดีที่สุด เมื่อสกัดคุณลักษณะแบบ Boolean weighting แต่อย่างไรก็ LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นอันดับที่ 4 ซึ่งโดยคุณลักษณะการทำงานของโมเดล Bi-LSTM ที่มีการทำงานที่เหมาะสมในการคำนวณค่าแบบต่อเนื่อง ซึ่งเหมาะกับการทำงานกับข้อมูลรูปแบบข้อความ อีกทั้งยังมีการคำนวณค่าแบบสองทิศทางทำให้เรียนรู้รูปแบบข้อความได้ดี จึงทำให้ประสิทธิภาพโดยรวมในการทำงานให้ผลที่ดีที่สุด

ตารางที่ 25 ผลการทดลองด้วยคุณลักษณะ Boolean Weighting

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	69.07	97.67	44.26	63.67	94.97	77.08	60.36	49.99	0.01
SVM	84.44	85.90	82.78	83.33	85.42	84.59	84.07	3145.48	0.01
Decision-Tree	77.50	79.99	77.23	77.83	79.43	78.88	78.30	537.39	0.01
Naïve Bayes	84.07	82.83	86.61	86.14	83.44	84.44	84.98	22.33	0.01
DNN	85.65	84.20	87.09	86.82	84.57	85.47	85.79	653.88	0.01
LSTM	91.39	90.45	89.58	89.70	90.38	90.03	89.94	132.60	0.03

4.2.4 ผลการทดลองด้วยคุณลักษณะ TF Weighting

จากผลการทดลองในตารางที่ 26 จากการสกัดคุณลักษณะแบบ TF Weighting จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้องเท่ากับ 93.61% ด้วยค่า SD เท่ากับ 0.05 นอกจากนี้ยังคงให้ค่าความระลึกมากที่สุดในการทำนายข้อความประชดประชัน ซึ่งให้ค่าระลึกเท่ากับ 91.31% และเมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุดคือ 90.43% ทั้งนี้ขั้นตอนวิธี Naive Bayes ยังคงใช้เวลาในการประมวลผลน้อยที่สุด จากการใช้คุณลักษณะที่สกัดแบบ TF Weighting แต่อย่างไรก็ตาม LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นอันดับที่ 3

ตารางที่ 26 ผลการทดลองด้วยคุณลักษณะ TF Weighting

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	71.85	96.51	49.36	65.59	93.43	78.09	64.55	50.52	0.01
SVM	83.24	85.53	83.17	83.60	85.15	84.54	84.14	3206.56	0.01
Decision-Tree	77.04	80.76	77.41	78.13	80.10	79.41	78.72	488.22	0.02
Naïve Bayes	83.52	83.26	85.39	85.10	83.59	84.16	84.47	22.03	0.01
DNN	86.11	85.75	86.89	86.77	85.92	86.23	86.38	99.24	0.01
LSTM	93.61	89.60	91.69	91.31	90.11	90.43	90.87	125.28	0.05

4.2.5 ผลการทดลองด้วยคุณลักษณะ TF-IDF Weighting

จากผลการทดลองในตารางที่ 27 จากการสกัดคุณลักษณะแบบ TF Weighting จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้องเท่ากับ 93.61% ด้วยค่า SD เท่ากับ 0.04 นอกจากนี้ยังคงให้ค่าความระลึกมากที่สุดในการทำนายข้อความประชดประชัน ซึ่งให้ค่าระลึกเท่ากับ 89.91% และเมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุดคือ 92.48% ทั้งนี้ขั้นตอนวิธี Naive Bayes ยังคงใช้เวลาในการประมวลผลน้อยที่สุด จากการใช้คุณลักษณะที่สกัดแบบ TF-IDF Weighting แต่อย่างไรก็ตาม LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นอันดับที่ 3

ตารางที่ 27 ผลการทดลองด้วยคุณลักษณะ TF-IDF Weighting

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	71.57	95.96	47.28	64.56	92.24	77.18	62.46	48.85	0.02
SVM	84.81	89.16	82.61	83.71	88.37	86.34	85.38	3563.93	0.01
Decision-Tree	74.63	78.88	76.94	77.39	78.43	78.12	77.67	594.39	0.02
Naïve Bayes	83.61	78.10	89.13	87.79	80.28	82.65	84.46	14.52	0.01
DNN	84.44	86.31	85.30	85.51	86.20	85.87	85.70	87.16	0.01
LSTM	93.61	95.22	89.31	89.91	95.01	92.48	92.06	125.36	0.04

4.2.4 ผลการทดลองด้วยคุณลักษณะ Word Embedding

จากผลการทดลองในตารางที่ 28 จากการสกัดคุณลักษณะแบบ Word Embedding ซึ่งทดลองกับเทคนิค LSTM จะเห็นได้ว่าให้ค่าความถูกต้องเท่ากับ 91.39% ด้วยค่า SD เท่ากับ 0.03 และให้ค่าความแม่นยำในการทำนายข้อความประชดประชันได้เท่ากับ 90.45% เมื่อเปรียบเทียบการ

ใช้คุณลักษณะ Boolean, TF, TF-IDF Weighting ในเทคนิควิธี LSTM ให้ค่าประสิทธิภาพโดยรวม F-measure อยู่ในอันดับที่ 3 ซึ่งเท่ากับวิธี Boolean Weighting เท่ากับ 90.03% และใช้เวลาในการประมวลผลเท่ากัน

ตารางที่ 28 ผลการทดลองด้วยคุณลักษณะ Word Embedding

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
LSTM	91.39	90.45	89.58	89.70	90.38	90.03	89.94	132.60	0.03

จากผลการทดลองด้วยคุณลักษณะของข้อมูลจากบริบทในข้อความ โดยการสกัดคุณลักษณะทั้ง 4 แบบจากตารางที่ 26 ตารางที่ 27 และตารางที่ 28 สรุปผลการทดลองได้ว่าวิธีที่ให้ค่าความถูกต้องในการทำนายคือคุณลักษณะที่สกัดด้วย TF และ TF-IDF ซึ่งให้ค่าความถูกต้องเท่ากันคือ 93.61% แต่ทั้งนี้คุณลักษณะ TF-IDF ให้ค่าความแม่นยำในการทำนายข้อความประชดประชันมากกว่าเท่ากับ 95.22% และให้ประสิทธิภาพโดยรวมมากที่สุด ซึ่งขั้นตอนวิธีที่ให้ประสิทธิภาพสูงที่สุดคือเทคนิควิธี LSTM ในทุกคุณลักษณะ

4.3 ผลการทดลองจำแนกข้อมูลจากเนื้อหาในข้อความ

การทดลองโดยการใช้คุณลักษณะจากเนื้อหาในข้อความ ซึ่งประกอบด้วย การมีข้อความที่แสดงอารมณ์เชิงบวก การมีข้อความที่แสดงอารมณ์เชิงลบ การใช้คำที่ไม่ปกติ จำนวนของการใช้ปรัศนี จำนวนของการแสดงคำหัวเราะ การมีคำที่แสดงถึงการประชดประชันโดยทั่วไป จำนวนของการใช้เครื่องหมายอัศเจรีย์ จำนวนของการใช้มหัพภาค จำนวนของคำหยาบคาย จำนวนของการใช้บุพสัญญา จำนวนของการใช้ไม้ยมก ไอคอนแสดงอารมณ์เชิงบวก ไอคอนแสดงอารมณ์เชิงลบ จำนวนของการใช้เครื่องหมายบวก และจำนวนของการใช้เครื่องหมายลบ ด้วยเทคนิคการเรียนรู้ของเครื่องและ ด้วยเทคนิคการเรียนรู้เชิงลึก โดยแสดงผลการทดลองดังในตารางที่ 29 จะเห็นได้ว่าขั้นตอนวิธี DNN ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้องเท่ากับ 80.86% ด้วยค่า SD เท่ากับ 0.02 และให้ค่าประสิทธิภาพโดยรวมเป็นอันดับที่ 2 เท่ากับ 78.78% แต่อย่างไรก็ตาม DNN ใช้เวลาในการประมวลผลมากที่สุดเป็นอันดับที่ 2 ส่วนขั้นตอนวิธี Decision Tree ใช้เวลาในการประมวลผลน้อยที่สุด

ตารางที่ 29 ผลการทดลองจำแนกข้อมูลจากเนื้อหาในข้อความ

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	49.17	99.89	0.23	50.03	43.81	66.65	0.45	1.37	0.02
SVM	80.00	72.09	88.98	86.72	76.13	78.71	82.03	13.06	0.01

ตารางที่ 29 ผลการทดลองจำแนกข้อมูลจากเนื้อหาในข้อความ (ต่อ)

Decision-Tree	80.00	72.04	89.01	86.75	76.10	78.69	82.03	0.20	0.01
Naïve Bayes	78.15	57.66	97.57	96.01	69.73	71.93	81.27	0.21	0.01
DNN	80.86	71.80	89.74	87.26	76.47	78.78	82.57	76.18	0.02
LSTM	80.37	80.67	81.56	81.46	80.91	80.99	81.16	7550.22	0.01

4.4 ผลการทดลองจำแนกข้อมูลจากบริบทในข้อความและเนื้อหาในข้อความ

การทดลองโดยการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหาในข้อความ ด้วยการวัดค่าน้ำหนัก 4 แบบคือ 1) Boolean Weighting 2) TF Weighting 3) TF-IDF Weighting และ 4) Word Embedding ด้วยเทคนิคการเรียนรู้ของเครื่องและด้วยเทคนิคการเรียนรู้เชิงลึก โดยคุณลักษณะจากบริบทในข้อความนั้นเป็นคำที่สกัดได้จากชุดข้อความ และคุณลักษณะจากเนื้อหาในข้อความสกัดได้จากชุดข้อความซึ่งประกอบด้วย การมีข้อความที่แสดงอารมณ์เชิงบวก การมีข้อความที่แสดงอารมณ์เชิงลบ การใช้คำที่ไม่ปกติ จำนวนของการใช้ปรัศนี จำนวนของการแสดงคำหัวเราะ การมีคำที่แสดงถึงการประชดประชันโดยทั่วไป จำนวนของการใช้เครื่องหมายอัศเจรีย์ จำนวนของการใช้มหัพภาค จำนวนของคำหยาบคาย จำนวนของการใช้บุพสัญญา จำนวนของการใช้ไม้ยมก ไอคอนแสดงอารมณ์เชิงบวก ไอคอนแสดงอารมณ์เชิงลบ จำนวนของการใช้เครื่องหมายบวก และจำนวนของการใช้เครื่องหมายลบ ผลการทดลองการรวมคุณลักษณะจากบริบทในข้อความและคุณลักษณะจากเนื้อหาในข้อความ สรุปผลการทดลองได้ดังนี้

4.4.1 ผลการทดลองด้วยคุณลักษณะ Boolean Weighting + Content

การทดลองโดยการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหา โดยการให้ค่าน้ำหนักด้วยวิธี Boolean weighting ดังแสดงในตารางที่ 30 จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้อง 95.46% ด้วยค่า std = 0.04 นอกจากนี้ยังให้ค่าระลึกดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าระลึกเท่ากับ 92.44% ส่วน KNN ให้ค่าความแม่นยำดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าความแม่นยำในการทำนายข้อความประชดประชันถึง 92.77% แต่เมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุด คือ 91.42% ส่วนขั้นตอนวิธี KNN ใช้เวลาในการประมวลผลน้อยที่สุด อย่างไรก็ตามขั้นตอนวิธี LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นลำดับที่ 3

ตารางที่ 30 ผลการทดลองด้วยคุณลักษณะ Boolean Weighting + Content Feature

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	77.59	92.77	64.63	77.45	89.94	81.34	75.16	37.22	0.02

ตารางที่ 30 ผลการทดลองด้วยคุณลักษณะ Boolean Weighting + Content Feature (ต่อ)

SVM	87.87	87.69	87.43	87.48	87.63	87.58	87.52	1902.80	0.01
Decision-Tree	81.67	84.06	83.32	83.46	83.93	83.75	83.62	306.79	0.01
Naïve Bayes	88.06	88.11	89.37	89.27	88.38	88.68	88.81	44.22	0.01
DNN	89.81	88.66	89.20	89.14	88.69	88.89	88.94	111.92	0.01
LSTM	95.46	90.47	92.63	92.44	90.74	91.42	91.66	144.64	0.04

4.4.2 ผลการทดลองด้วยคุณลักษณะ TF Weighting + Content

การทดลองโดยการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหา โดยการให้ค่าน้ำหนักด้วยวิธี TF weighting ดังแสดงในตารางที่ 31 จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้อง 95.37% ด้วยค่า std = 0.03 นอกจากนี้ยังให้ค่าระลึกดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าระลึกเท่ากับ 94.08% ส่วน KNN ให้ค่าความแม่นยำดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าความแม่นยำในการทำนายข้อความประชดประชันถึง 95.28% แต่เมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุด คือ 93.15% ส่วนขั้นตอนวิธี Naïve Bayes ใช้เวลาในการประมวลผลน้อยที่สุด อย่างไรก็ตามขั้นตอนวิธี LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นลำดับที่ 2

ตารางที่ 31 ผลการทดลองด้วยคุณลักษณะ TF Weighting + Content Feature

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	76.11	95.28	58.08	69.49	92.58	80.34	71.29	70.04	0.01
SVM	87.22	87.59	86.29	86.49	87.43	87.02	86.84	4242.45	001
Decision-Tree	82.68	83.76	83.48	83.53	83.71	83.64	83.59	730.66	0.01
Naïve Bayes	87.31	88.78	87.46	87.66	88.62	88.21	88.03	51.67	0.01
DNN	89.81	88.32	90.26	90.08	88.55	89.18	89.38	107.86	0.01
LSTM	95.37	92.27	94.29	94.08	92.54	93.15	93.39	146.56	0.03

4.4.3 ผลการทดลองด้วยคุณลักษณะ TF-IDF Weighting + Content

การทดลองโดยการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหา โดยการให้ค่าน้ำหนักด้วยวิธี TF weighting ดังแสดงในตารางที่ 32 จะเห็นได้ว่าขั้นตอนวิธี LSTM ให้ค่าความถูกต้องมากที่สุด โดยให้ค่าความถูกต้อง 96.67% ด้วยค่า std = 0.05 นอกจากนี้ยังให้ค่าระลึกดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าระลึกเท่ากับ 92.05% และยังให้ค่าความแม่นยำดีที่สุดในการทำนายข้อความประชดประชัน โดยให้ค่าความแม่นยำในการทำนายข้อความประชด

ประชันถึง 93.91% แต่เมื่อพิจารณาค่า F-measure จะเห็นได้ว่า LSTM ให้ค่า F-measure สูงสุด คือ 92.96% ส่วนขั้นตอนวิธี Naïve Bayes ใช้เวลาในการประมวลผลน้อยที่สุด อย่างไรก็ตามขั้นตอนวิธี LSTM ใช้เวลาในการประมวลผลมากที่สุดเป็นลำดับที่ 2

ตารางที่ 32 ผลการทดลองด้วยคุณลักษณะ TF Weighting + Content Feature

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
KNN	83.52	77.51	91.06	89.71	80.19	83.14	85.26	33.65	0.02
SVM	88.15	87.61	89.80	89.60	87.85	88.85	88.801	2105.68	0.01
Decision-Tree	81.57	82.75	83.46	83.34	82.88	83.04	83.16	353.71	0.01
Naïve Bayes	88.55	77.17	94.75	93.64	80.58	84.60	87.09	23.05	0.01
DNN	90.00	87.48	89.66	89.48	87.76	88.44	88.67	90.78	0.01
LSTM	96.67	93.91	91.97	92.05	93.90	92.96	92.91	159.02	0.05

4.4.4 ผลการทดลองด้วยคุณลักษณะ Word Embedding + Content

การทดลองโดยการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหา โดยการให้ค่าน้ำหนักด้วยวิธี Word Embedding ดังแสดงในตารางที่ 33 ด้วยขั้นตอนวิธี LSTM โดยให้ค่าความถูกต้องมากที่สุดเมื่อเปรียบเทียบกับทุกวิธีที่กล่าวมาข้างต้น ซึ่งมีค่าเท่ากับ 96.79% ด้วยค่า std = 0.01 นอกจากนี้ยังให้ค่าระลึกดีที่สุดในการทำนายข้อความประชดประชันมากที่สุด โดยให้ค่าระลึกเท่ากับ 95.08% และเมื่อพิจารณาค่า F-measure จะเห็นได้ว่าให้ค่า F-measure สูงสุด คือ 96.88% และเมื่อพิจารณาผลการทดลองโดยรวมทั้งหมดวิธีนี้ให้ประสิทธิภาพสูงที่สุด

ตารางที่ 33 ผลการทดลองด้วยคุณลักษณะ Word Embedding

	Accuracy	Precision		Recall		F-measure		times	STD
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm		
LSTM	96.79	98.80	94.72	95.08	98.79	96.88	96.68	802.58	0.01

4.5 ผลการทดลองด้วยคุณลักษณะ Word Embedding ด้วยวิธี Hold out

จากการทดลองสรุปได้ว่าคุณลักษณะที่ดีที่สุดที่ใช้ในการทดลองที่ให้ประสิทธิภาพสูงที่สุดคือ Word Embedding + Content Feature ด้วยขั้นตอนวิธี LSTM ซึ่งใช้วิธีการทดลองในการแบ่งชุดข้อมูลแบบ 10-Fold Cross Validation ดังนั้นการทดลองนี้จึงใช้การทดลองที่ให้ผลดีที่สุดมาทำการแบ่งชุดการทดลองเป็นชุดสร้างโมเดล 70% และแบ่งเป็นชุดทดสอล 30% เพื่อเปรียบเทียบผลการทดลองว่าผลการทดลองด้วย 10-Fold Cross Validation ซึ่งผลการทดลองดังแสดงในตารางที่ 34

ซึ่งผลการทดลองแสดงให้เห็นว่า ค่าความถูกต้องเมื่อเปรียบเทียบกับการทดลองด้วยวิธี 10-Fold ได้ค่าความถูกต้องลดลงจาก 96.79% เป็น 86.33% ดังนั้นจากผลการทดลองที่ลดลงแสดงให้เห็นว่าชุดข้อมูลที่ใช้สร้างโมเดลส่วนมากที่เมื่อนำมาสร้างโมเดลแล้วทำให้สามารถทำนายคลาสได้ดีหรือกล่าวอีกนัยหนึ่งคือชุดข้อมูลสามารถแบ่งแยกแต่ละคลาสได้ง่ายนั่นเอง การแก้ไขปัญหาจะต้องเก็บข้อมูลเพิ่มมากขึ้น ซึ่งเป็นข้อเสนอแนะเพิ่มเติมในการวิจัยครั้งต่อไป

ตารางที่ 34 ผลการทดลองจำแนกข้อมูลจากบริบทในข้อความและเนื้อหาในข้อความ Word Embedding ด้วยวิธี Hold out

	Accuracy	Precision		Recall		F-measure		times
		Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	Sarcasm	Non-Sarcasm	
LSTM	86.60	88.43	84.74	85.50	87.81	86.94	86.25	44.75

จากผลการทดลองทำให้ทราบว่าการใช้คุณลักษณะจากบริบทในข้อความรวมกับคุณลักษณะจากเนื้อหา ให้ประสิทธิภาพที่ดีกว่าการใช้คุณลักษณะจากบริบทของข้อความอย่างเดียว หรือการใช้คุณลักษณะจากเนื้อหาในข้อความอย่างเดียว เนื่องจากการใช้แค่คุณลักษณะจากบริบทในข้อความไม่สามารถจำแนกข้อความประชดประชันที่เป็นข้อความที่ไม่ใช่ข้อความที่สื่อสารออกมาแบบข้อความทั่วๆ ไป บางข้อความที่สื่อออกมามีความหมายตรงกันข้ามกับความหมายที่แท้จริงของผู้สื่อสาร ทำให้การใช้คุณลักษณะจากบริบทในข้อความจำแนกได้ไม่ดี ทั้งนี้การใช้คุณลักษณะจากเนื้อหาในข้อความอย่างเดียวจะสามารถจำแนกข้อความประชดประชันได้ดีเนื่องจากคุณลักษณะจากเนื้อหาในข้อความถูกสกัดจากข้อความที่เป็นข้อความประชดประชันทำให้จำแนกข้อความประชดประชันได้ไม่ดี ดังนั้นการรวมคุณลักษณะทั้งสองแบบจึงเป็นวิธีที่เหมาะสมในการจำแนกข้อความประชดประชัน เนื่องจากมีคุณลักษณะที่เป็นตัวแทนของการเป็นตัวแทนของข้อความประชดประชันที่ดีกว่า จึงทำให้ประสิทธิภาพการจำแนกมีประสิทธิภาพมากขึ้น ทั้งนี้เทคนิคอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอนที่ให้ประสิทธิภาพสูงที่สุดคือ SVM ส่วนเทคนิคอัลกอริทึมที่ให้ประสิทธิภาพสูงที่สุดและเหมาะสมที่จะนำไปใช้งานมากที่สุดคือการใช้คุณลักษณะจากการแปลงข้อความเป็นเวกเตอร์ด้วย Word Embedding ซึ่งมีข้อดีในการเรียนรู้รูปแบบของข้อความและสามารถวิเคราะห์ความคล้ายคลึงของรูปแบบการเกิดข้อความ และนำมาเรียนรู้ด้วยเทคนิคการเรียนรู้เชิงลึกด้วยอัลกอริทึม Bi-LSTM ซึ่งเป็นการเรียนรู้ข้อความแบบสองทิศทางทำให้เรียนรู้ข้อความรอบข้างได้ไกลขึ้น จึงเหมาะสมมากที่สุดในการนำมาใช้เป็นโมเดลสำหรับการจำแนกข้อความประชดประชันบนเครือข่ายสังคมออนไลน์ภาษาไทยในการวิจัยนี้

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผล

งานวิจัยนี้เป็นการทดลองการตรวจจับข้อความประชดประชันบนเครือข่ายสังคมอออนไลน์ ชุดข้อมูลที่ใช้ในการทดลองเป็นการเก็บข้อมูลจากผู้ใช้งานเครือข่ายสังคมออนไลน์เฟสบุ๊ค ซึ่งใช้วิธีการรวบรวมข้อมูลจากเว็บไซต์เฟสบุ๊คโดยการค้นหาจากคำค้นโดยใช้แฮสแท็ก #ประชดประชัน จากนั้นทำการศึกษาคุณลักษณะ 2 แบบเพื่อใช้ในการตรวจสอบข้อความประชดประชัน คุณลักษณะแรกที่ศึกษา คือ คุณลักษณะที่สกัดจากบริบทของข้อความ (Context-based features) ซึ่งเป็นคุณลักษณะที่ใช้คำในข้อความเพื่อเป็นถุงคำของคุณลักษณะ แบบที่สอง คือ คุณลักษณะที่สกัดจากเนื้อหาในข้อความ (Content-based feature) จากนั้นใช้คุณลักษณะแบบแรกและแบบที่สองรวมกันเพื่อใช้เป็นคุณลักษณะในการตรวจจับข้อความประชดประชัน ซึ่งสามารถอภิปรายผลการใช้งานคุณลักษณะทั้งสองแบบได้ดังนี้ 1) คุณลักษณะจากบริบทของข้อความ คุณลักษณะจากบริบทของข้อความใช้หลักการวิธีในการจำแนกข้อความ ซึ่งประกอบด้วยวิธีการดังนี้ 1) การทำความสะอาดข้อความ โดยการตัดข้อความภาษาอังกฤษ ตัวเลข สัญลักษณ์พิเศษต่าง ๆ ออกจากข้อความ 2) การตัดคำภาษาไทย ซึ่งใช้วิธีการตัดคำจากการวิจัย [63, 64] มาใช้ในการตัดคำ 4) การกำหนดค่าน้ำหนักของคุณลักษณะใช้การคำนวณด้วยวิธี Boolean Weighting TF Weighting TF-IDF Weighting และ Word Embedding 2) คุณลักษณะจากเนื้อหาในข้อความ เป็นคุณลักษณะที่ดึงจากองค์ประกอบจากเนื้อหาในข้อความ เช่น การมีข้อความที่แสดงอารมณ์เชิงบวก-ลบ การใช้คำที่ไม่ปกติ การใช้ปรัศนี การแสดงคำหัวเราะ การมีคำที่แสดงถึงการประชดประชันโดยทั่วไป การใช้เครื่องหมายอัศเจรีย์ การใช้มหัพภาค คำหยาบคาย การใช้บุพสัญญา การใช้ไม้ยมก ไอคอนแสดงอารมณ์เชิงบวก-ลบ การใช้เครื่องหมายบวก และการใช้เครื่องหมายลบ ซึ่งได้ผลการศึกษาที่สามารถสรุปได้ ดังนี้

1. คุณลักษณะจากบริบทในข้อความ ซึ่งจากการทดลองในงานวิจัยนี้เลือกใช้คุณลักษณะทั้งหมดที่สกัดได้ และใช้เทคนิควิธีการเรียนรู้ด้วยอัลกอริทึมสองแบบคือ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก ผลการทดลองสำหรับการเรียนรู้จากคุณลักษณะด้วยบริบทในข้อความนั้นได้ค่าความถูกต้องเท่ากับ 93.61% ด้วยคุณลักษณะ TF Weighting ซึ่งเท่ากับ TF-IDF Weighting แต่คุณลักษณะการใช้ค่าน้ำหนักด้วย TF-IDF ให้ประสิทธิภาพโดยรวมมากว่าคือเท่ากับ 92.48% ด้วยขั้นตอนวิธี LSTM

2. คุณลักษณะจากในข้อความ ซึ่งจากการทดลองในงานวิจัยนี้เลือกใช้คุณลักษณะที่สกัดได้จากเนื้อหาในข้อความซึ่งประกอบด้วย การมีข้อความที่แสดงอารมณ์เชิงบวก การมีข้อความที่แสดงอารมณ์เชิงลบ การใช้คำที่ไม่ปกติ จำนวนของการใช้ปรัศนี จำนวนของการแสดงคำหัวเราะ การมีคำที่

แสดงถึงการประชดประชันโดยทั่วไป การใช้เครื่องหมายอัศเจรีย์ การใช้มหัพภาค คำหยาบคาย การใช้บุพสัญญา การใช้ไม้ยมก ไอคอนแสดงอารมณ์เชิงบวก ไอคอนแสดงอารมณ์เชิงลบ การใช้เครื่องหมายบวก และการใช้เครื่องหมายลบ และใช้เทคนิควิธีการเรียนรู้ด้วยอัลกอริทึมสองแบบคือ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก ผลการทดลองสำหรับการเรียนรู้จากคุณลักษณะด้วยเนื้อหาในข้อความนั้นได้ค่าความถูกต้องเท่ากับ 80.86% ด้วยวิธีการเทคนิคอัลกอริทึม DNN

3. การรวมกันระหว่างคุณลักษณะบริบทในข้อความและคุณลักษณะจากเนื้อหาในข้อความโดยการเลือกใช้คุณลักษณะจากบริบทในข้อความเลือกคุณลักษณะที่สกัดได้ทั้งหมด รวมกับคุณลักษณะจากเนื้อหาในข้อความ ซึ่งประกอบด้วย การมีข้อความที่แสดงอารมณ์เชิงบวก การมีข้อความที่แสดงอารมณ์เชิงลบ การใช้คำที่ไม่ปกติ จำนวนของการใช้ปรัศนี จำนวนของการแสดงคำหัวเราะ การมีคำที่แสดงถึงการประชดประชันโดยทั่วไป จำนวนของการใช้เครื่องหมายอัศเจรีย์ จำนวนของการใช้มหัพภาค จำนวนของคำหยาบคาย จำนวนของการใช้บุพสัญญา จำนวนของการใช้ไม้ยมก ไอคอนแสดงอารมณ์เชิงบวก ไอคอนแสดงอารมณ์เชิงลบ จำนวนของการใช้เครื่องหมายบวก และจำนวนของการใช้เครื่องหมายลบ และใช้เทคนิควิธีการเรียนรู้ด้วยอัลกอริทึมสองแบบคือ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก ขั้นตอนวิธี การเรียนรู้เชิงลึกด้วยอัลกอริทึม LSTM ให้ประสิทธิภาพการจำแนกข้อความประชดประชันสูงที่สุดคือ 96.79%

5.2 อภิปรายผล

จากการศึกษาวิจัยการตรวจจับข้อความประชดประชันบนเครือข่ายสังคมออนไลน์สามารถอภิปรายผลได้ดังนี้

1. การเรียนรู้ข้อความประชดประชันบนเครือข่ายสังคมออนไลน์ ซึ่งโดยทั่วไปเป็นข้อความที่มีลักษณะเนื้อหาที่แตกต่างจากการใช้ข้อความปกติ ลักษณะการสื่อสารส่วนมากบ่งบอกถึงการแสดงออกที่ตรงข้ามกับความหมายที่แท้จริงของของเจ้าของข้อความ การเลือกคุณลักษณะที่จะเป็นตัวแทนของข้อความประชดประชันที่ดีนั้นจึงเป็นสิ่งสำคัญในการศึกษางานด้านการจำแนกข้อความประชดประชัน

2. การวิจัยนี้แสดงให้เห็นว่า การเลือกคุณลักษณะจากบริบทในข้อความอย่างเดียว หรือใช้คุณลักษณะจากเนื้อหาในข้อความอย่างเดียวนั้นไม่เพียงพอต่อการจำแนก แต่การรวมกันระหว่างคุณลักษณะบริบทในข้อความและคุณลักษณะที่สกัดจากเนื้อหาในข้อความ สามารถให้ประสิทธิภาพที่สูงขึ้นและการเรียนรู้เชิงลึกด้วยอัลกอริทึม LSTM ให้ผลการทดลองที่ดีที่สุดเนื่องเป็นเทคนิควิธีที่เหมาะกับข้อมูลในรูปแบบข้อความซึ่งเป็นการข้อมูลรูปแบบต่อเนื่อง

5.3 ข้อเสนอแนะ

จากการทดลองพบว่า ผลการจำแนกด้วยแบบจำลองที่จะทำให้มีประสิทธิภาพการจำแนกที่สูงนั้น การเลือกใช้อัลกอริทึมที่เหมาะสมเป็นเพียงปัจจัยหนึ่งในการทำการวิจัย หากแต่กระบวนการวิจัยทั้งหมดมีความสำคัญทั้งสิ้น ไม่ว่าจะเป็นการเก็บข้อมูลให้ได้จำนวนที่มากพอเพื่อให้ได้แบบจำลองที่น่าเชื่อถือ การสร้างคุณลักษณะที่สามารถเป็นตัวแทนของข้อมูลที่นำมาใช้ในการวิเคราะห์ อีกข้อสำคัญในการทำการวิจัยด้านการวิเคราะห์ข้อความที่เป็นภาษาไทย คือ การเลือกใช้วิธีการตัดคำที่มีประสิทธิภาพ ทั้งนี้ปัญหาการตัดคำนั้นยังเป็นปัญหาในภาษาไทย เนื่องจากข้อมูลที่นำมาใช้ทดลองนั้นมาจากสื่อสังคมออนไลน์ ที่ผู้ใช้งานพิมพ์ผิด การใช้ภาษาที่ไม่ตรงกับคำศัพท์ การหลีกเลี่ยงการพิมพ์ข้อความที่สามารถสื่อความหมายได้โดยตรง ดังนั้น การวิจัยทางการด้านการวิเคราะห์ภาษาไทย ยังเป็นความท้าทายสำหรับนักวิจัยที่ทำงานทางด้านการประมวลผลภาษาธรรมชาติสำหรับภาษาไทย ซึ่งงานวิจัยในอนาคตมีดังนี้

1. การรวมรวมข้อความประชดประชันจำนวนมากขึ้นซึ่งข้อความจำนวนมากจะเรียนรู้การสร้างโมเดลที่ดียิ่งขึ้น

2. การศึกษาข้อความประชดประชันจากข้อความหลากหลายประเภท เช่น ความคิดเห็นด้านการเมือง ความคิดเห็นด้านรีวิวสินค้า เป็นต้น

3. การทดลองการเลือกใช้ตัวตัดคำภาษาไทยที่มีประสิทธิภาพมากขึ้นในอนาคตเนื่องจากข้อความประชดประชันมีลักษณะการพิมพ์ข้อความที่มีรูปแบบที่ไม่ปกติทำให้การตัดคำได้ไม่ดีเท่าที่ควรทำให้ต้องทำความสะอาดข้อมูลให้ดีขึ้น

4. การศึกษาวิธีการเลือกคุณลักษณะจากบริบทในข้อความเพื่อค้นหาตัวแทนของชุดข้อมูลที่แท้จริงและช่วยลดเวลาในการประมวลผล

5. การใช้วิธีการเพิ่มหรือสกัดคุณลักษณะด้วยวิธีอื่น เช่น การใช้หน้าที่ของคำ การเพิ่มคุณลักษณะจากเนื้อหาในข้อความ

6. การเลือกเทคนิคอัลกอริทึมการเรียนรู้แบบอื่นที่เหมาะสมกับการเรียนรู้ข้อมูลชนิดข้อความ เช่น Gated Recurrent Unit (GRU), Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU), Recurrent Convolutional Neural Networks (RCNN), Random Multi-model Deep Learning (RMDL) และ Hierarchical Deep Learning for Text (HDLTex)

บรรณานุกรม

บรรณานุกรม

[1] Chan SWK, Chong MWC. Sentiment analysis in financial texts. Decision Support Systems 2017; 9453-64.

[2] Tartir S, Abdul-Nabi I. Semantic Sentiment Analysis in Arabic Social Media. Journal of King Saud University - Computer and Information Sciences 2017; 29229-233. [2017-08-24 16:34:03]

[3] Gitto S, Mancuso P. Improving airport services using sentiment analysis of the websites. Tourism Management Perspectives 2017; 22132-136. [2017-08-24 16:34:37]

[4] Jasso G, Meza I. Character and Word Baselines Systems for Irony Detection in Spanish Short Texts. Procesamiento Del Lenguaje Natural 2016; [56]: 41-48.

[5] Bouazizi M, Otsuki T. A Pattern-Based Approach for Sarcasm Detection on Twitter. Ieee Access [Article] 2016; 45477-5488.

[6] Razali MS, Halin AA, Ye L, Doraisamy S, Norowi NM. Sarcasm Detection Using Deep Learning With Contextual Features. IEEE Access 2021; 968609-68618.

[7] Bharti SK, Babu KS, Jena SK. Parsing-based Sarcasm Sentiment Recognition in Twitter Data. 2015; 1373-1380.

[8] Bouazizi M, Ohtsuki T. Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis. 2015; 1594-1597.

[9] Kumar AA, S.Chandrasekhar. Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. International Journal of Engineering Research & Technology July 2012; 1[5]: 1-6.

[10] Shivaprasad TK, Shetty J. Sentiment analysis of product reviews: A review. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT); March 2017; 298-301.

[11] C.Ramasubramanian, R.Ramya. Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering December 2013; 2[12]: 4536-4537.

[12] Nasim Z, Rajput Q, Haider S. Sentiment analysis of student feedback using machine learning and lexicon based approaches. 2017 International Conference on Research and Innovation in Information Systems (ICRIIS); July 2017; 1-6.

[13] Qin Z, Petrounias I. A Semantic-Based Framework for Fine Grained Sentiment Analysis. 2017 IEEE 19th Conference on Business Informatics (CBI); July 2017; 295-301.

[14] Wicana SG, İbisoglu TY, Yavanoglu U. A Review on Sarcasm Detection from Machine-Learning Perspective. 2017 IEEE 11th International Conference on Semantic Computing (ICSC); January 2017; 469-476.

[15] Bharti SK, Vachha B, Pradhan RK, Babu KS, Jena SK. Sarcastic sentiment detection in tweets streamed in real time: a big data approach. Digital Communications and Networks 2016; 2108-121. [2017-09-14 22:48:29]

[16] Mukherjee S, Bala PK. Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. Technology in Society 2017; 4819-27. [2018-06-01 11:22:36]

[17] Fersini E, Pozzi FA, Messina E. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 2015; 1-8.

[18] Belwal RC, Rai S, Gupta A. Text summarization using topic-based vector space model and semantic measure. Information Processing & Management 2021; 58[3]: 102536. [2022-02-22 04:21:16]

[19] ภูมิรพี ภูมิค้า. เทคนิคการเรียนรู้เชิงลึกเพื่อวิเคราะห์ความรู้สึกจากผู้ใช้ผลิตภัณฑ์. Thesis: สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี; 2562.

[20] Pasupa K, Ayutthaya TSN. Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features. Sustainable Cities and Society 2019; 50

[21] Ahuja R, Sharma SC. Transformer-Based Word Embedding With CNN Model to Detect Sarcasm and Irony. Arabian Journal for Science and Engineering 2021;

[22] Khan FH, Qamar U, Bashir S. Senti-CS: Building a lexical resource for sentiment analysis using subjective feature selection and normalized Chi-Square-based feature weight generation. Expert Systems 2016; 33[5]: 489-500. [2018-06-01 11:21:15]

[23] Vilares D, Alonso MA, Gómez-Rodríguez C. Supervised sentiment analysis in multilingual environments. Information Processing & Management 2017; 53595-607. [2017-08-24 16:32:31]

[24] Huang F, Zhang S, Zhang J, Yu G. Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing 2017; 253144-153. [2017-08-24 16:30:21]

[25] Dave AD, Desai NP. A comprehensive study of classification techniques for sarcasm detection on textual data. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT); March 2016; 1985-1991.

[26] ปริญญา สงวนสัตย์. *Artificial Intelligence with Machine Learning, AI* สร้างได้ด้วยแมชชีนเลิร์นนิ่ง. บริษัท ซีเอ็ดยูเคชั่น จำกัด (มหาชน); 2019.

[27] วรรษา เงินดี, วีรยุทธ เจริญเรืองกิจ. BOOK RECOMMENDATION WITH DATA MINING USING RAPIDMINER. 2021; [2022-02-19 02:00:49]

[28] ปิยวรรณ นิลถนอม, ธนพร มาลัย, สายชล สินสมบูรณ์ทอง. การเปรียบเทียบประสิทธิภาพการทำนายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. Thai Journal of Science and Technology 2021; 10[1]: [2022-02-19 02:08:58]

[29] อัชฌาพร กว้างสวาสดิ์, เพียงฤทัย หนูสวัสดิ์, วราลี คงเหมาะ, ปวีณา ทิพยากุลรักษ์, บุษกร สังขนันท์. ระบบทำนายระดับความเครียด ด้วยเทคนิคต้นไม้การตัดสินใจ. Rattanakosin Journal of Science and Technology 2019; 1[2]: 13-26. [2022-02-19 12:06:11]

[30] Pasupa K, Seneewong Na Ayutthaya T. Hybrid Deep Learning Models for Thai Sentiment Analysis. Cognitive Computation 2021;

[31] ทรงกรด พิมพิศาล ณัฐวุฒิ ศรีวิบูลย์. การประมวลผลภาพสำหรับการจำแนกรูปภาพฟันดัสโดยใช้การเรียนรู้เชิงลึก. JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY 2020; 10[2]: 19-25. [2022-02-15 01:14:41]

[32] Chunyan Y, Chen Y, Zuo W. Multi-Task Deep Neural Networks for Joint Sarcasm Detection and Sentiment Analysis. Pattern Recognition and Image Analysis 2021; 31[1]: 103-108. https://dx.doi.org/10.1134/s105466182101017x

[33] Onan A, Tocoglu MA. A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification. IEEE Access 2021; 97701-7722. https://dx.doi.org/10.1109/ACCESS.2021.3049734

[34] Ayutthaya TSN, Pasupa K. Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features. 2018 International Joint

Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP); 2018-11; 1-6.

[35] ภานุพงษ์ ร่องอ้อ และ นุวีย์ วิวัฒนวัฒนา. THE IMPACT OF FEATURE EXTRACTION AND DATA IMPUTATION ON PM2.5 FORECASTING MODEL FOR BANGKOK AREA. Thesis: Srinakharinwirot University; 2020.

[36] B P, K.p. S, Kumar MA. A deep learning approach for Malayalam morphological analysis at character level. Procedia Computer Science 2018; 13247-54. [2022-02-22 07:21:55]

[37] Eke CI, Norman AA, Shuib L. Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model. IEEE Access 2021; 948501-48518. https://dx.doi.org/10.1109/ACCESS.2021.3068323

[38] Rajadesingan A, Zafarani R, Liu H. Sarcasm Detection on Twitter: A Behavioral Modeling Approach. 2015; New York, NY, USA: ACM; 97–106.

[39] Mukherjee S, Bala PK. Detecting sarcasm in customer tweets: an NLP based approach. Industrial Management & Data Systems 2017; 117[6]: 1109-1126. [2018-06-01 11:22:15]

[40] Boudad N, Faizi R, Oulad Haj Thami R, Chiheb R. Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal 2017; [2017-08-24 16:30:43]

[41] Hiai S, Shimada K. A Sarcasm Extraction Method Based on Patterns of Evaluation Expressions. 2016; 31-36.

[42] Salas-Zárate MdP, Paredes-Valverde MA, Rodriguez-García MÁ, Valencia-García R, Alor-Hernández G. Automatic detection of satire in Twitter: A psycholinguistic-based approach. Knowledge-Based Systems 2017; 12820-33. [2018-06-01 13:15:07]

[43] Bouazizi M, Ohtsuki T. Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); August 2015; 1594-1597.

[44] Mladenović M, Krstev C, Mitrović J, Stanković R. Using Lexical Resources for Irony and Sarcasm Classification. 2017; New York, NY, USA: ACM; 13:11–13:18.

[45] de Freitas LA, Vanin AA, Hogetop DN, Bochernitsan MN, Vieira R. Pathways for Irony Detection in Tweets. 2014; New York, NY, USA: ACM; 628–633.

[46] Reganti AN, Maheshwari T, Kumar U, Das A, Bajpai R. Modeling Satire in English Text for Automatic Detection. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW); December 2016; 970-977.

[47] Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F. Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts. 2017 8th International Conference on Information Technology (ICIT); May 2017; 703-709.

[48] Justo R, Corcoran T, Lukin SM, Walker M, Torres MI. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems 2014; 69124-133. [2017-08-24 16:11:06]

[49] Kunneman F, Liebrecht C, van Mulken M, van den Bosch A. Signaling sarcasm: From hyperbole to hashtag. Information Processing & Management 2015; 51[4]: 500-509. [2017-08-24 16:14:03]

[50] Schifanella R, de Juan P, Tetreault J, Cao L. Detecting Sarcasm in Multimodal Social Platforms. 2016; New York, NY, USA: ACM; 1136–1145.

[51] Taslioglu H, Karagoz P. Irony Detection on Microposts with Limited Set of Features. 2017; New York, NY, USA: ACM; 1076–1081.

[52] Al-Ghadhban D, Alnkhilan E, Tatwany L, Alrazgan M. Arabic sarcasm detection in Twitter. 2017 International Conference on Engineering MIS (ICEMIS); May 2017; 1-7.

[53] Jain T, Agrawal N, Goyal G, Aggrawal N. Sarcasm detection of tweets: A comparative study. 2017 Tenth International Conference on Contemporary Computing (IC3); August 2017; 1-6.

[54] Razali MS, Halin AA, Norowi NM, Doraisamy SC. The importance of multimodality in sarcasm detection for sentiment analysis. 2017 IEEE 15th Student Conference on Research and Development (SCOReD); December 2017; 56-60.

[55] Gidhe P, Ragha L. Sarcasm detection of non # tagged statements using MLP-BP. 2017 International Conference on Advances in Computing, Communication and Control (ICAC3); December 2017; 1-4.

[56] Chaudhari P, Chandankhede C. Literature survey of sarcasm detection. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET); March 2017; 2041-2046.

[57] Bhan N, D'silva M. Sarcasmometer using sentiment analysis and topic modeling. 2017 International Conference on Advances in Computing, Communication and Control (ICAC3); December 2017; 1-7.

[58] Manohar MY, Kulkarni P. Improvement sarcasm analysis using NLP and corpus based approach. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS); June 2017; 618-622.

[59] Karoui J, Zitoune FB, Moriceau V. SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media. Procedia Computer Science 2017; 117[Supplement C]: 161-168. [2017-12-10 15:32:05]

[60] Liebrecht CC, Kunneman FA, Bosch APJvd. The perfect solution for detecting sarcasm in tweets #not. http://aclweborg/anthology/W/W13/W13-1605pdf 2013; [2017-09-18 14:05:40]

[61] Ahmad T, Akhtar H, Chopra A, Akhtar MW. Satire Detection from Web Documents Using Machine Learning Methods. 2014 International Conference on Soft Computing and Machine Intelligence; Sept 2014; 102-105.

[62] Vateekul P, Koomsubha T. A study of sentiment analysis using deep learning techniques on Thai Twitter data. 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE); July 2016; 1-6.

[63] Limkonchotiwat P, Phatthiyaphaibun W, Sarwar R, Chuangsuwanich E, Nutanong S. Handling Cross- and Out-of-Domain Samples in Thai Word Segmentation. Findings 2021; 2021-08; Online: Association for Computational Linguistics; 1003–1016.

[64] Phatthiyaphaibun W. OSKut (Out-of-domain StacKed cut for Word Segmentation). 2022-01-23 12:02:59]; https://github.com/mrpeerat/OSKut.

[65] Kittinaradorn R. A Thai word tokenization library using Deep Neural Network. 2022-03-05 05:06:20]; https://github.com/rkcosmos/deepcut.

[66] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:13013781 [cs] 2013; [2022-03-05 05:17:17]

[67] Phatthiyaphaibun W. Thai Sentiment Analysis Toolkit. 2022-03-01T13:12:59Z]; https://www.kaggle.com/datasets/rtatman/thai-sentiment-analysis-toolkit.

[68] Karthik E, Sethukarasi T. Sarcastic user behavior classification and prediction from social media data using firebug swarm optimization-based long short-term

memory. The Journal of Supercomputing 2021; https://dx.doi.org/10.1007/s11227-021-04028-4

ภาคผนวก ก

ตารางแสดงคำอารมณ์เชิงบวก

กตเวที	กตัญญู	กล้า	กรุณา	กันเอง	ขำ
แข็ง	แข็งแกร่ง	แข็งขัน	แข็งแรง	ครื้นครึก	ครึกครื้น
แจ่ม	แจ่มใส	ชัดเจน	ชุ่มชื้น	เชี่ยวชาญ	เด่น
เต็ม	ทันสมัย	ทรหด	แถม	แท้	นวล
น่ารัก	นุ่ม	นูน	เน้น	แน่	แน่ชัด
บริบูรณ์	บวก	เบา	เบาใจ	เบิกบาน	ปราดเปรียว
ประณีต	อดทน	ทน	ฉลาด	ชัดเจน	ตรง
ตงฉิน	โต	แท้	จริง	เน้น	แน่
แน่ชัด	แน่แน่ว	แน่ว	แน่วแน่	บริบูรณ์	บ้องแบ๊ว
เบาใจ	เบิกบาน	ประณีต	ประดิษฐ	ประดิษฐ์	ประหลาด
ปราดเปรียว	ปลั่ง	แปล้	ผ่อง	ผ่องใส	ถูก
พริ้งเพรา	สวยงาม	พะงา	พิทักษ์	พูน	เพราพริ้ง
เพราเพริศ	เพริศแพร้ว	เพริศพราย	เพา	มหา	มอน
มนุญ	มั่งคั่ง	มั่งมี	มั่นคง	ยอดเยี่ยม	เยี่ยม
เยี่ยมยอด	ยุ้ย	รจนา	รอบคอบ	ร่ำรวย	รุ่ง
รุ่งเรือง	รุ่งโรจน์	เร็ว	เร้า	เรือง	ลอยด
ละเอียด	ละมุน	ละมุนละม่อม	ทันสมัย	ล้ำเลิศ	ลึกซึ้ง
เล่อ	แล้	วาว	วิรุฬห์	ศุกร	สดใส
สนิท	สนิทสนม	สบาย	สมจริง	สมบูรณ์	สร้างสรรค์
สวย	สวยงาม	สห	สะดวก	สะอาด	สัณห์
สำรวม	สำเร็จ	สุก	สุข	สุขี	สุภาพ
สูง	สูงส่ง	เสมอ	ใส	หนาแน่น	หนุน
หนุ่มแน่น	เหมาะเจาะ	เหมาะสม	อร่อย	อธิ	ใหญ่โต
อติ	อ่อนโยน	อัศจรรย์	อิฏฐ	อึด	โอชะ
ฮ้อ	เฮง	ฮา	ดี	ดูแลตัวเอง	ดูแลรักษา
ดูแลเอาใจใส่	ดีใจ	หอม	กรวด	อุ่นใจ	อุ้มชู
เข้าตา	เข้าตา กรรมการ	กรวดน้ำ	รัก	ชอบ	กราบ

กลับใจ	กอด	การันตี	เหลือเชื่อ	เด่น	งดงาม
ก้าว	ก้าวหน้า	เรียน	กรุ่ม	กล้าหาญ	การันตี
พบ	เจอ	ช่วย	เล่น	อนุมัติ	กตเวที
กตัญญู	กล้า	กรุณา	กันเอง	ขำ	แข็ง
แข็งแกร่ง	แข็งขัน	ปรองดอง	แข็งแรง	ครื้นครึก	รวดเร็ว
ครึกครื้น	แจ่ม	แจ่มใส	ชัดเจน	ชัดแจ้ง	ชัดแจ๋ว
ชัดถ้อยชัดคำ	ชุ่มชื้น	โปรด	นิรโทษ	หวาน	น่ารื่นรมย์
ปาฏิหาริย์	หายาก	เชี่ยวชาญ	เด่น	ไม่เคย	น่ากอด
เต็ม	ทันสมัย	ทรหด	แถม	ความเป็นกลาง	เป็นกลาง
แท้	นวล	ขอบใจ	ขอบคุณ	น่ารัก	นุ่ม
นูน	เน้น	แน่	แน่ชัด	บริบูรณ์	บวก
เบา	เบาใจ	เบิกบาน	ปราดเปรียว	ประณีต	ศานติ
อดทน	ทน	ฉลาด	ประหยัด	คิดถึง	อบอุ่น
ความสุข	สุข	เคารพ	นับถือ	รักษา	ตรง
คำนึงถึง	สรรเสริญ	ติดตาม	ประทับใจ	บำรุง	ซาบซึ้ง
เชื่อฟัง	ตงฉิน	ไม่หวั่น	รอยยิ้ม	ยิ้ม	ประทับใจ
ระวัง	โต	นมัสการ	แท้	จริง	เน้น
แน่	แน่ชัด	แน่แน่ว	แน่ว	ภูมิใจ	มงคล
มงคลสมรส	มหัศจรรย์	มหัศจรรย์ใจ	มีความรู้	มีความยินดี	มีความหวัง
แน่วแน่	บริบูรณ์	บ้องแบ๊ว	เบาใจ	เบิกบาน	ประณีต
ประดิษฐ	ประดิษฐ์	น่าประหลาด	ประหลาด	ปราดเปรียว	ปลั่ง
แปล้	ผ่อง	อดออม	อนุโมทนา	คว้ารางวัล	คุ้มครอง
คุ้มครองรักษา	ทำลายสถิติ	ทำสำเร็จ	ที่เชื่อถือได้	ไม่เป็นอันตราย	ผ่องใส
ถูก	พริ้งเพรา	สวยงาม	พะงา	พิทักษ์	พูน
เพราพริ้ง	เพราเพริศ	เพริศแพร้ว	เพริศพราย	เพา	มหา
มอน	มนุญ	มั่งคั่ง	มั่งมี	มั่นคง	ยอดเยี่ยม
เยี่ยม	เยี่ยมยอด	ยุ้ย	รจนา	เงียบสงบ	ปลอดภัย
ที่ปลอดภัย	ทำบุญ	ทำบุญทำทาน	หยุดนิ่ง	เงียบสงบ	รักสงบ

มากมาย	ตกลง	ซื่อสัตย์	เป็นกันเอง	รอบคอบ	ร่ำรวย
โดดเด่น	ตามควร	รุ่ง	รุ่งเรือง	รุ่งโรจน์	ขยันหมั่นเพียร
มีประสิทธิภาพ	เร็ว	เร้า	เรือง	ลอยด	ละเอียด
ละมุน	ละมุนละม่อม	ทันสมัย	ล้ำเลิศ	ลึกซึ้ง	เล่อ
แล้	ความชื่นชม	ความชื่นชมยินดี	ความชื่นชอบ	ความชื่นบาน	ความดี
ความดีงาม	ความดีใจ	ความทัดเทียม	วาว	วิรุฬห์	เสถียรภาพ
ชนะ	เสริมสร้าง	ความแข็งแรง	สร้าง	เพลิดเพลิน	ความรัก
ศุกร	สมดุล	ยังชีพ	น่านับถือ	สด	สดใส
สนใจไยดี	เกียรติ	สนิท	สนิทสนม	สบาย	สมจริง
สมบูรณ์	สร้างสรรค์	สวย	สวยงาม	สห	สะดวก
สะอาด	สัณห์	สำรวม	สำเร็จ	สุก	สุข
สุขี	สุภาพ	สูง	สูงส่ง	เสมอ	ใส
หนาแน่น	หนุน	หนุ่มแน่น	เหมาะเจาะ	เหมาะสม	มีประโยชน์
อร่อย	อธิ	มีเหตุผล	ทุ่มเท	โรแมนติก	ถูกกาลเทศะ
มีชีวิตชีวา	มีชีวิตรอด	มีชื่อ	มีชื่อเสียง	ตื่นเต้น	กระตือรือร้น
ตื่นเต้น	พร้อม	ใหญ่โต	มั่นอกมั่นใจ	ลงตัว	คล่องแคล่ว
คล่องมือ	คล่องตัว	ใส่ใจ	ไม่ได้ยอมแพ้	เมตตา	มีประสิทธิผล
มีประสิทธิภาพ	มีประโยชน์	มีอิสรภาพ	มีเงินทอง	มีเงินมีทอง	สนใจ
สุขสันต์	ชอบด้วยกฎหมาย	ชอบธรรม	เห็นด้วย	อติ	สนุก
อ่อนโยน	คบค้า	คบหา	คบหา	คบหาสมาคม	อ่อนโยน
มนุญ	คณิกา	ความสงบ	สงบ	อัศจรรย์	ประมาณ
อิฏฐ	อึด	ประชาธิปไตย	โอชะ	ฮ้อ	เฮง
ฮา	คุณค่า	นับถือ	อย่า	ฟรี	ให้กำลังใจ
กำลังใจ	ให้ความร่วมมือ	ให้ความสนใจ	ขอพร	ใจรัก	เชื่อมั่น
ความเชื่อมั่น	ความเชื่อมั่น	เอาใจใส่	เอาใจเขามา	ความเอาใจใส่	ความเอื้ออาทร

	ในตนเอง		ใส่ใจเรา		
ความเอื้อเฟื้อ	พอใจ	น่าพอใจ	น่าประทับใจ	น่าพึงพอใจ	น่าฟัง
น่าภาคภูมิใจ	น่ามหัศจรรย์	น่ายกย่อง	น่ายินดี	สนุกสนาน	ดี
ดีใจ	กรวด	กรวดน้ำ	รัก	ชอบ	กราบ
กลับใจ	กอด	การันตี	ก้าว	ก้าวหน้า	เรียน
กรุ่ม	กล้าหาญ	การันตี	พบ	เจอ	ช่วย
เล่น	อนุมัติ				

ภาคผนวก ข

ตารางแสดงคำอารมณ์เชิงลบ

กระจุกกระจิก	กระดาน	แก่	เกรียน	เกเร	เกิน
ขวางโลก	ขัดสมาธิ	ขุ่น	คมคาย	กระเซอะกระเซิง	เคร่งขรึม
เครียด	งอ	งอแง	โง่	จังไร	จัญไร
เจือ	เจือจาง	ต่ำ	ต่ำต้อย	ฉูดฉาด	ชั่ว
ชู้สาว	ซวย	ซ้ำ	ซ้ำซ้อน	ซ้ำร้าย	เซ่อ
ดุดัน	ดุร้าย	ตกม้าตาย	ทราม	ทะลุ	ทุเรศ
เทียม	นักเลง	ลบ	เบื่อ	ปรัมปรา	เท็จ
ตอแหล	ตาขาว	ฉุนเฉียว	เฉโก	เฉื่อย	ชะงัด
ชู้สาว	ซวย	ซ้ำซ้อน	ซ้ำร้าย	ซุกซน	เซ
เซ่อ	เซา	ดื้อ	ดุร้าย	เดียรดาษ	ตกม้าตาย
ตรงข้าม	ตระหนี่ถี่เหนียว	ตอแหล	ตาขาว	ต่ำ	ต่ำต้อย
ติดเชื้อ	เตี้ย	ถวัล	ทรหด	ทราม	ทะลุ
ทื่อ	ทุเรศ	เท็จ	เทียม	น้อย	นักเลง
น่าย	บิ่น	บุ๋ม	เบี่ยง	เบื่อ	ป้อ
ปากเสีย	ปากหมา	ปาราชิก	เป็นลม	เปล่าเปลี่ยว	เปื่อย
แปร่ง	ผรุส	ผิด	ผิดหวัง	เผ็ดร้อน	เผล
แผลง	พด	พล่าน	พิการ	พี	ฟก
เพี้ยน	แพง	ฟก	ฟาง	แฟบ	มัวหมอง
โมเม	ยับ	ยืด	ยุ่งเหยิง	โย้	รกเรื้อ
ร่อแร่	รัดกุม	ร้าย	ร้ายกาจ	ริบ	รุนแรง
รุกรุย	รุ่ยร่าย	ล้ม	ลัก	ล้าสมัย	ลำบาก
ล้ำ	เลว	โลเล	ว่ายาก	วิปโยค	วิประโยค
วิเวก	สลัว	สวก	สวะ	สะเหล่อ	สั้น
สันหลังยาว	สับปลับ	สัพเพเหระ	สิ้นสติ	สูญ	เสีย
เสื่อมโทรม	เสื่อมเสีย	โสโครก	โสมม	หงุดหงิด	หน้าด้าน
หนาว	หมองมัว	หมิ่น	หยาบ	หยอก	หรี่

หลง	ห่วย	หวิว	หวะ	หึง	หัวสูง
เหงา	เหม็น	เหี้ย	เหียก	เหี้ยมโหด	แหย
โหดร้าย	โหดเหี้ยม	ไหม้	อกตัญญู	อกแตก	อลัชชี
อ้วก	อ้วน	อ๋อง	อ่อนเพลีย	อี๋อฉาว	เอน
ฆ่า	ตัด	ข่ม	ข่มขืน	ขโมย	ตาย
ตก	ล้ม	เสียใจ	กดดัน	กัก	กดขี่
กักขัง	กระชาก	กระวน กระวาย	กระแทก	กระทืบ	ก่อกวน
ก่อกรรมทำเข็ญ	กระทำชำเรา	กระทำอนาจาร	กระแทกกระทั้น	กระตุกหนวดเสือ	กระอัก
กระหึ่ม	กริ่งใจ	โกหก	กราดเกรี้ยว	กวาดล้าง	กรรโชก
กดราคา	ร้องไห้	กลัว	กลั่นแกล้ง	เศร้า	ร้องไห้
แพง	กำเริบ	แกว่ง	ขัดใจ	กล่าวหา	ก้าวก่าย
กำจัด	กังขา	กอบโกย	กวน	กวนตีน	กัด
กังวล	เกร็ง	เกรงกลัว	เกรงใจ	เกลียด	เกลื่อน
โกหก	แก่งแย่ง	ข่มขวัญ	ข่มขี่	ข่มขู่	ขวาง
ขัดขืน	ขัดข้อง	ขัดขวาง	ร้องเรียน	ขัง	ขาดแคลน
ขาดทุน	ขายชาติ	ขายขี้หน้า	ขายหน้า	ขาลง	ขี้
ขีดฆ่า	ขี้ไม้	ขืน	ขืนใจ	ขู่	ขู่เข็ญ
ขูดรีด	เขี่ย	เขวี้ยง	คต	คดโกง	คดงอ
ครอบงำ	คลื่นไส้	คลาดเคลื่อน	คว้าน้ำเหลว	คว่ำ	คว่ำบาตร
คอร์รัปชัน	คะนอง	คับ	คับคั่ง	คัด	คัดค้าน
ค้าง	ค้างคา	ค้างคาใจ	คาใจ	คุกเข่า	คุ้มเกรง
คุมขัง	คุมตัว	เค้น	เคร่ง	เคร่งเครียด	เคราะห์ซ้ำกรรมซัด
เคือง	แค้น	ใคร่ครวญ	ฆ่าแกง	ฆาต	ฆ่าตัดตอน
ฆ่าฟัน	ฆ่าไม่ตาย ขายไม่	ฆ่ายกครัว	ฆ่าล้างโคตร	เฆี่ยน	งง

	ขาด				
งงวย	งก	งด	เงี่ยน	จน	จม
จำโซ่	จิก	จำนำ	จิตตก	จับกุม	จำกัด
เจ็บ	เจ็บไข้	เจ็บใจ	ฉ้อ	ฉี่	ฉีก
ชก	ชดใช้	ช่วงชิง	ช็อก	ชะงัก	ชะลอ
ชัก	ชักกระตุก	ชักใย	ชัง	ช้ำใจ	ชิง
ชิงชัง	ชิงดีชิงเด่น	ชิงนรกเกิด	ชิงสุกก่อนห่าม	ชิงหมาเกิด	ซ่อน
ซึม	ซึมซาบ	ซ้ำเติม	ซุก	ซุกซ่อน	แซว
แซะ	ไซ้	เณรคุณ	ด้อย	ดัดสันดาน	ดับ
ด่า	ด่าทอ	ดิ้น	ดื้อ	ดุ	ดูแคลน
ดูถูก	ดูหมิ่น	ดูหมิ่นถิ่นแคลน	เดา	เดือด	เดือดดาล
เดือดร้อน	แดก	แดกดัน	โดน	ได้เสีย	ตก
ตกงาน	ตกใจ	ตกนรกทั้งเป็น	ตกหลุม	ตกหลุมพราง	ตั้งแง่
ตายดาบหน้า	ตรม	ตบ	ตบตา	ตราหน้า	ต่อต้าน
ตอบโต้	ต่อย	ต่อสู้	ตอแหล	ตะลุมบอน	ตัดขาด
ต้าน	ตาย	ตายด้าน	ตาลาย	ตำหนิ	ติ
ติเตียน	ติดเชื้อ	ติดขัด	ตัดพ้อ	ตัดพ้อต่อว่า	ตี
ตีกิน	ถอน	ถีบ	ทรยศ	ถึงฆาต	ทรุด
ท่วม	ท้อ	ทะลึ่ง	ทักท้วง	ทุบ	ทุบตี
เท	เทกระจาด	แทง	แทงใจดำ	นอกใจ	เนรคุณ
บกพร่อง	บด	บ่น	บ่อน	บ่อนทำลาย	บาด
บาดหมาง	บานปลาย	บิด	บิดเบือน	เบน	เบียด
เบียดเบียน	เบื่อ	เบื่อหน้า	โบ้ย	ปฏิเสธ	ปด
ปน	ประจาน	ประจาร	ประชด	ประชดประชัน	ประณาม

ประท้วง	ประมาท	ปรับ	ปราบ	ปราบปราม	ปลงใจ
ป่วย	ปลงชีวิต	ปลง	ปล้น	ปล้นสะดม	ปลอม
ปรำ	ปวด	ปลอมปน	ปะทะ	ปะปน	ปาด
เป็นลม	เปลือง	โป๊	ผลัก	ผัด	ผัดวันประกันพรุ่ง
ผ่า	เผ่น	ลืม	หลบหนี	ความขมขื่น	ความขมุกขมัว
ความขุ่นเคือง	ความข้องใจ	ความฉิบหาย	ความชั่วช้า	ความซ้ำซ้อน	ความดื้อ
ความทรุดโทรม	ฟุ่มเฟือย	ฟุ้งเฟ้อ	กระจุกกระจิก	กระดาน	แก่
เกรียน	คลุมเครือ	คลุมถุงชน	เกเร	เกิน	ช้า
วิกฤติ	เจ็บปวด	ความลึกลับ	เผด็จการ	ขัดแย้ง	องควิการ
รังเกียจ	อกหัก	ไร้สาระ	คงขาด	คบชู้	ครหา
คบชู้สู่ชาย	กำกวม	มืด	ฮือ	ลึกลับ	กลุ้มใจ
น่ากลัว	เกียจคร้าน	ขวางโลก	เศร้าใจ	ไม่มีทางรักษา	สิ้นหวัง
ไร้ผล	อัมพาต	เศร้าซึม	สลดใจ	ท้อแท้	น่าเบื่อ
กระสับกระส่าย	ขัดสมาธิ	ขุ่น	คมคาย	กระเซอะกระเซิง	เคร่งขรึม
คระหาย	กระหาย	คระเมิม	เครียด	คบไม่ได้	งอ
งอแง	โง่	จังไร	ไม่สน	จัญไร	เจือ
เจือจาง	ต่ำ	ต่ำต้อย	ฉูดฉาด	ชั่ว	ชู้สาว
ซวย	ซ้ำ	ซ้ำซ้อน	ซ้ำร้าย	เซ่อ	ดุดัน
ดุร้าย	ตกม้าตาย	ทราม	ทะลุ	ทุเรศ	เทียม
นักเลง	ลบ	เบื่อ	ปรัมปรา	เท็จ	ตอแหล
ตาขาว	ฉุนเฉียว	เฉโก	เฉื่อย	ชะงัด	ชู้สาว
ซวย	ซ้ำซ้อน	ซ้ำร้าย	ซุกซน	เซ	เซ่อ
เซา	ดื้อ	ดุร้าย	เดียรดาษ	ตกม้าตาย	ตรงข้าม
ตระหนี่ถี่	ตอแหล	ตาขาว	ต่ำ	ต่ำต้อย	ติดเชื้อ

เหนียว					
เตี้ย	ถวัล	ทรหด	ทราม	ทะลุ	ทื่อ
ทุเรศ	เท็จ	เทียม	น้อย	นักเลง	น่าย
บิ่น	บุ๋ม	เบี่ยง	เบื่อ	ป้อ	ปากเสีย
ปากหมา	ปาราชิก	เป็นลม	เปล่าเปลี่ยว	เปื่อย	แปร่ง
ผรุส	ผิด	ผิดหวัง	เผ็ดร้อน	เผล	แผลง
พด	พล่าน	พิการ	พี	ฟก	เพี้ยน
แพง	ฟก	ฟาง	แฟบ	มัวหมอง	โมเม
ยับ	ยืด	ยุ่งเหยิง	โย้	รกเรื้อ	ร่อแร่
รัดกุม	ร้าย	ร้ายกาจ	ริบ	รุนแรง	รุกรุย
รุ่ยร่าย	ล้ม	ลัก	ล้าสมัย	ลำบาก	ล้ำ
เลว	โลเล	ยาก	ว่ายาก	วิปโยค	วิประโยค
วิเวก	สลัว	สวก	สวะ	สะเหล่อ	สั้น
สันหลังยาว	สับปลับ	สัพเพเหระ	สิ้นสติ	สูญ	โทรม
เสีย	เสื่อมโทรม	เสื่อมเสีย	โสโครก	โสมม	หงุดหงิด
หน้าด้าน	หนาว	หมองมัว	หมิ่น	หยาบ	หยอก
หรี่	หลง	ห่วย	หวิว	หวะ	หึง
หัวสูง	เหงา	เหม็น	ไม่เชื่อ	เหี้ย	บรรลัย
เหียก	เหี้ยมโหด	แหย	โหดร้าย	ทำสงคราม	ทำอันตราย
ทำเป็นทองไม่รู้ร้อน	ทำเป็นเล่น	ทำเล่นๆ	ทำโดยพลการ	ทำเอาเจ็บ	ทิ้งขว้าง
ทิ้งๆ ขว้างๆ	ไม่รู้คุณ	ไม่รู้จบรู้สิ้น	ทำใจไม่ได้	ดูไม่ได้	ด่า
ด่ากลับ	ไม่รู้ไม่ชี้	ไม่ลงตัว	วุ่นวาย	วุ่นวายใจ	ก่อความวุ่นวาย
สับสนวุ่นวาย	สับสน	ไม่รู้สึกตัว	ไม่ลงรอย	ไม่ลงรอยกัน	ไม่สบายใจ
ไม่สมบูรณ์	ไม่สมฐานะ	ไม่สม่ำเสมอ	ไม่สมควร	ไม่อร่อย	ไม่เข้ากัน
ไม่เจียมกะลาหัว	ไม่เชื่อ	ไม่เหมาะสม	ไม่เห็นค่า	ไม่เห็นด้วย	ละเลย

มีโทษ	มอมเมา	มอมเหล้า	มากเกินไป	มีความผิด	มีชู้
มีปัญหา	ไม่เอาไหน	ไม่ใช่	ไม่ได้เรื่อง	ไม่ไว้วางใจ	ไม่ไว้หน้าใคร
ไม่ไว้ใจ	ไม่ไหว	อดไม่ไหว	ทนไม่ไหว	สละสิทธิ์	ขับไล่
ขับไล่ไสส่ง	ไล่ออก	เมื่อย	ทำร้าย	ไม่สวย	ไม่สะไม่สวย
โหดเหี้ยม	ไหม้	อกตัญญู	อกแตก	อลัชชี	อ้วก
อ้วน	ทำมิดีมิร้าย	ทำลายขวัญ	ทำลายล้าง	อ๋อง	อ่อนเพลีย
อื้อฉาว	เอน	ฆ่า	ตัด	ข่ม	ข่มขืน
ขโมย	ตาย	ตก	ทุจริต	ล้ม	เสียใจ
กดดัน	กัก	กดขี่	กักขัง	กระชาก	กระวนกระวาย
กระแทก	กระทืบ	ก่อกวน	ก่อกรรมทำเข็ญ	กระทำชำเรา	กระทำอนาจาร
กระแทกกระทั้น	กระตุกหนวดเสือ	กระอัก	กระหึ่ม	กริ่งใจ	โกหก
กราดเกรี้ยว	กวาดล้าง	กรรโชก	กดราคา	ร้องไห้	กลัว
กลั่นแกล้ง	เศร้า	ร้อนตัว	ร้าว	ร้าวราน	แพง
เมา	ลดเกียรติ	ลาตาย	ลำบากใจ	ล่มจม	ล่มสลาย
ล่วงประเวณี	ล่วงละเมิด	ล่อแหลม	กดหัว	กำเริบ	แกว่ง
ขัดใจ	กล่าวหา	ก้าวก่าย	กำจัด	กังขา	กอบโกย
กวน	กวนตีน	กัด	กังวล	ไฟไหม้	น้ำท่วม
สีนามิ	ทำบาป	ทำบาปทำกรรม	ไอ้เวร	ไอ้บ้า	ไอ้
คิดมาก	คิดมิดีมิร้าย	คิดร้าย	ถุย	ถุยน้ำลาย	เกร็ง
เกรงกลัว	ปาดคอ	รุม	รุมล้อม	ชั่งแม่ง	แม่ง
มั่ว	มั่วซั่ว	มั่วนิ่ม	มั่วสุม	ติดคุก	คุก
เกรงใจ	เกลียด	ไม่มีแรง	ไม่มีใคร	ไม่มีสติ	ไม่มีน้ำใจ

ไม่มี	ไม่มีการศึกษา	เกลื่อน	โกหก	แก่งแย่ง	แย่
คอรัปชัน	เหลื่อมล้ำ	ข่มขวัญ	ข่มขี่	ข่มขู่	ขวาง
ขัดขืน	ขัดข้อง	ขัดขวาง	ร้องเรียน	ขัง	ขาดแคลน
ขาดทุน	ขายชาติ	ขายขี้หน้า	ขายหน้า	ขาลง	เปียก
ขี้	ขีดฆ่า	ขี้โม้	ขืน	ขืนใจ	ขู่
ขู่เข็ญ	ขูดรีด	เขี่ย	ทรมาน	เขวี้ยง	โกง
คต	คดโกง	คดงอ	ครอบงำ	คลื่นไส้	คลาดเคลื่อน
คว้าน้ำเหลว	คว่ำ	คว่ำบาตร	คอร์รัปชัน	คะนอง	คับ
คับคั่ง	ค้างชำระ	ชอบกล	คัด	คัดค้าน	ค้าง
ค้างคา	ค้างคาใจ	คาใจ	คุกเข่า	คุ้มเกรง	คุมขัง
คุมตัว	เค้น	ชัดช้า	ความหดหู่	หดหู่	หดหู่ใจ
ความสูญเสีย	ความหงุดหงิด	เคร่ง	อย่างเคร่งเครียด	ตึงเครียด	ความเคร่งเครียด
ผิดจารีตประเพณี	ผิดจังหวะ	ผิดคำพูด	ผิดคำสัตย์	เคร่งเครียด	เคราะห์ซ้ำกรรมซัด
เคือง	แค้น	ใคร่ครวญ	ฆ่าแกง	มดเท็จ	ฆาต
ฆ่าตัดตอน	ฆ่าฟัน	ฆ่าไม่ตายขายไม่ขาด	ฆ่ายกครัว	ฆ่าล้างโคตร	ม้วนหน้า
เฆี่ยน	งง	จก	งงงวย	งก	งด
เงี่ยน	จน	จม	จำโซ่	จิก	จำนำ
จิตตก	จับกุม	จำกัด	เจ็บ	เจ็บไข้	เจ็บใจ
ฉ้อ	ฉี่	ฉีก	ชก	ชดใช้	ช่วงชิง
ช็อก	ชะงัก	ชะลอ	ชัก	ชักกระตุก	ชักใย
ชัง	ช้ำใจ	ชิง	เชิง	ชิงชัง	ชิงดีชิงเด่น
ชิงนรกเกิด	ชิงสุกก่อนห่าม	ชิงหมาเกิด	ซ่อน	ซึม	ซึมซาบ
ซ้ำเติม	ซุก	ซุกซ่อน	แซว	แซะ	ไซ้

เณรคุณ	แดก	ด้อย	ล่มสลาย	ช่างมัน	ทาส
ดัดสันดาน	ดับ	ภาวะมลพิษ	ภาวะวิกฤติ	ภาวะสงคราม	ด่าทอ
ดิ้น	ละเหี่ยใจ	ดื้อ	ดุ	ดูแคลน	แคบ
ดูถูก	ดูถูกดูแคลน	ดูหมิ่น	ดูหมิ่นถิ่นแคลน	อิจฉาริษยา	เดา
เดือด	เดือดดาล	เดือดร้อน	แดก	แดกดัน	โดน
ได้เสีย	สงสาร	โมโห	โรคจิต	อีห่า	แม่ง
ไม่เกรงใจ	แทงคอ	แทง	เหนื่อย	ตก	ตกงาน
โหยหา	ฝืน	ตกใจ	ตกนรกทั้งเป็น	ตกหลุม	ตกหลุมพราง
ตั้งแง่	ตายดาบหน้า	ตรม	ตบ	ตบตา	ตราหน้า
ต่อต้าน	ตอบโต้	ต่อย	ต่อสู้	ตอแหล	ตะลุมบอน
ตัดขาด	ต้าน	ตาย	ตายด้าน	ตาลาย	ตำหนิ
ติ	ติเตียน	ติดเชื้อ	ติดขัด	ตัดพ้อ	ตัดพ้อต่อว่า
ตี	ตีกิน	ถอน	ถีบ	ทรยศ	ถึงฆาต
ทรุด	ท่วม	ท้อ	ทะลึ่ง	ทักท้วง	ทุบ
ทุบตี	เท	เทกระจาด	แทง	แทงใจดำ	นอกใจ
เนรคุณ	บกพร่อง	ร่วงโรย	บด	บ่น	อำเภอใจ
บ่อน	บ่อนทำลาย	บาด	บาดหมาง	บานปลาย	บิด
บิดเบือน	เบน	เบียด	เบียดเบียน	ขี้ข้า	เบื่อ
เบื่อหน้า	โบ้ย	ปฏิเสธ	ปด	ปน	ประจาน
ประจาร	ประชด	ประชดประชัน	ประณาม	ประท้วง	อันตราย
ประมาท	ปรับ	ความผิดพลาด	ถอยหลัง	ปราบ	ปราบปราม
ปลงใจ	ป่วย	ปลงชีวิต	ปลง	ปล้น	ปล้นสะดม
อกตัญญู	อดตาย	อดนอน	อดหลับอดนอน	ความไม่แน่นอน	ค่ำบาตร

คับอกคับใจ	คับแค	คับแคบ	คับแค้น	คับแค้นใจ	คับใจ
คาดโทษ	ปลอม	ปรำ	ยอมแพ้	ปวด	ปลอมปน
ปะทะ	บ้า	ปะปน	เคียดแค้น	ชั่วร้าย	ไม่ทน
หวงห้าม	ปาด	ล้มเหลว	เศร้าโศก	ไม่ยุติธรรม	น่าเวทนา
ความฉิบหาย	ขม	ไม่มีอารมณ์	ทุกข์ยาก	ยากจน	ไม่เพียงพอ
อึดอัด	ตกใจ	ไม่ลงรอยกัน	ขี้ขลาด	เงอะงะ	ลังเล
ฉุนเฉียว	ดื้อดึง	หายไป	เป็นลม	เยาะเย้ย	ปราบปราม
ดูหมิ่น	ชัง	เปลือง	เสแสร้ง	โป๊	ซับซ้อน
ใจร้อน	ผลัก	พังทลาย	พังพินาศ	พังยับเยิน	อกคราก
ขี้อาย	ย่ำแย่	แย่จัง	แย่มาก	ใจลอย	ใจสลาย
เสี่ยง	ขอโทษ	ผัด	อกจะแตก	ผัดวันประกันพรุ่ง	ผ่า
เผ่น	เป็นไปไม่ได้	ลืม	ฟุ้งซ่าน	ดรามา	ดราม่า
ไม่ดี	หมดกำลัง	หมดกำลังใจ	หมดความรู้สึก	หมดค่า	หมดจิตหมดใจ
หมดตัว	ให้การเท็จ	เสียกำลัง	เสียกำลังใจ	ความเดือดร้อน	เอาแต่ใจตัวเอง

พหูน ปณฺ ฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	ปราโมชย์ นามวงศ์
วันเกิด	วันที่ 27 พฤษภาคม 2525
สถานที่เกิด	ศรีสะเกษ
สถานที่อยู่ปัจจุบัน	39/145 หมู่ 13 ต.แสนสุข อ.วารินชำราบ จ.อุบลราชธานี 34190
ตำแหน่งหน้าที่การงาน	พนักงานมหาวิทยาลัย สายวิชาการ
สถานที่ทำงานปัจจุบัน	คณะบริหารธุรกิจและการจัดการ มหาวิทยาลัยราชภัฏอุบลราชธานี เลขที่ 2 ต.ในเมือง อ.เมือง จ.อุบลราชธานี 34000
ประวัติการศึกษา	พ.ศ. 2544 มัธยมศึกษา โรงเรียนสายธารวิทยา ตำบลสวนกล้วย อำเภอกันทรลักษ์ จังหวัดศรีสะเกษ พ.ศ. 2548 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยราชภัฏอุบลราชธานี พ.ศ. 2553 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศการเกษตรและพัฒนาชนบท มหาวิทยาลัยอุบลราชธานี พ.ศ. 2566 ปริญญาปรัชญาดุษฎีบัณฑิต (ปร.ด.) สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม