การวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ

วิทยานิพนธ์

ของ

วสวัตติ์ อินทร์แปลง

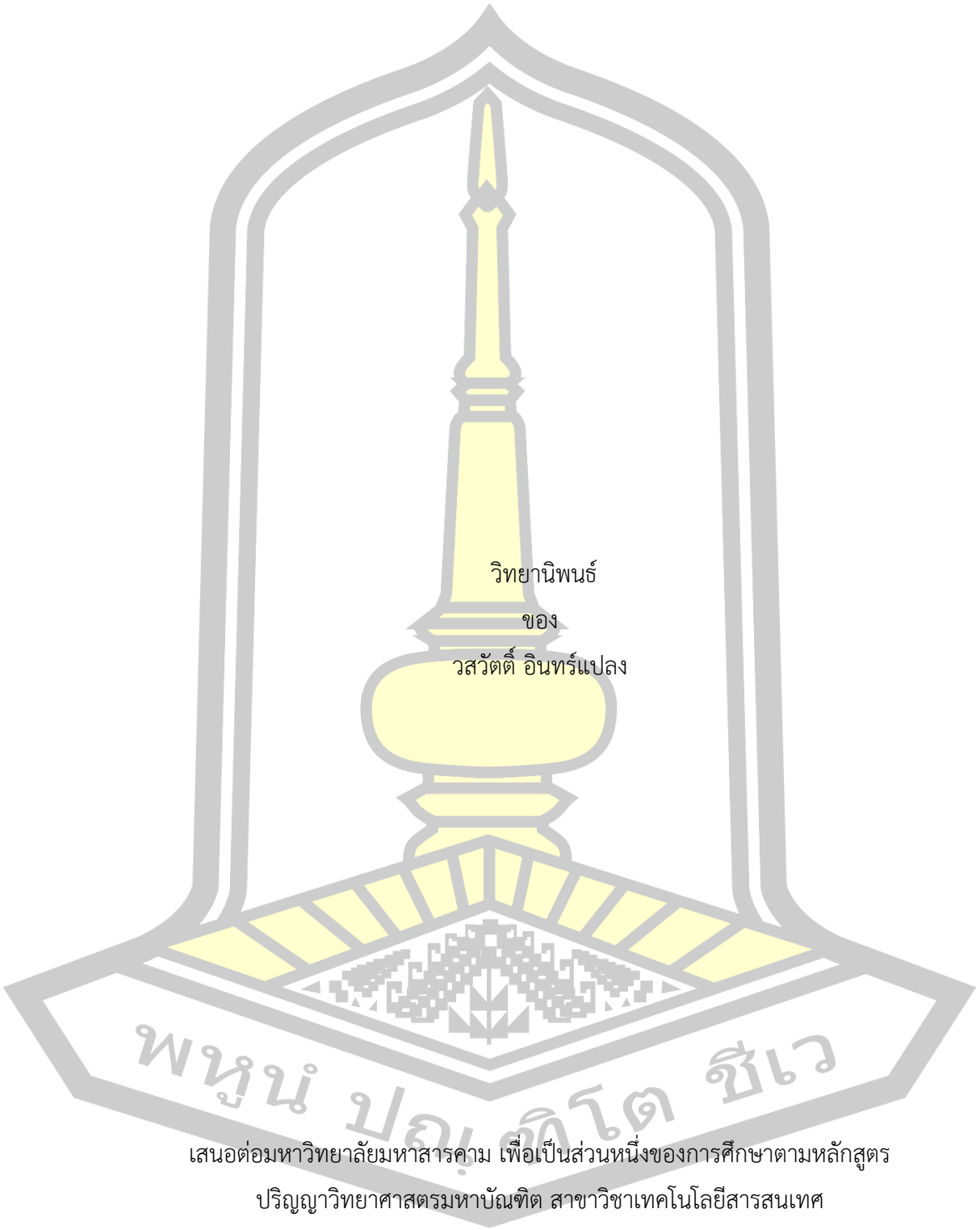
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

พฤษภาคม 2563

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

การวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ



วิทยานิพนธ์

ของ

วสวัตติ์ อินทร์แปลง

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

พฤษภาคม 2563

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Opinion Analysis on PUBG Mobile Game Using Text Mining

Wossawat Inplang

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

May 2020

Copyright of Mahasarakham University

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายวสวัตติ์ อินทร์แปลง แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. สิทธิชัย บุษหมั่น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จารี ทองคำ)

.....กรรมการ
(ผศ. ดร. แกมกาญจน์ สมประเสริฐศรี)

.....กรรมการ
(ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

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

พหุน ปญฺ ฑิโต ชีวเ

ชื่อเรื่อง การวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ
ผู้วิจัย วสวัตติ์ อินทร์แปลง
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จารี ทองคำ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2563

บทคัดย่อ

เหมืองข้อความเป็นกระบวนการวิเคราะห์ข้อมูลตัวอักษรเพื่อสกัดข้อมูลที่เป็นประโยชน์จากแหล่งข้อมูล ปัจจุบันเทคนิคในการจำแนกเหมืองข้อความมีหลายวิธี งานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคการจำแนก จาก 4 เทคนิคที่มีประสิทธิภาพ คือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine (SVM) เทคนิค K-Nearest Neighbor และเทคนิค C4.5 โดยเก็บรวบรวมข้อมูลความคิดเห็นต่อเกมมือถือผับจีจำนวน 3,798 ข้อความ ในกระบวนการคัดเลือกคำบ่งชี้เพื่อใช้ในการแยกคุณลักษณะได้เลือกใช้คำวิเศษณ์ และคำสแลงบางคำที่ความหมายของคำเป็นคำวิเศษณ์เพื่อทำการแยกคุณลักษณะเชิงบวกและเชิงลบ ผลการศึกษาพบว่ามีความไม่สมดุลของคลาสในข้อมูล โดยมีจำนวนคลาสหนึ่งมากกว่าอีกคลาสหนึ่งเป็นจำนวนมาก ผู้วิจัยจึงได้ทำการแก้ปัญหาโดยการปรับความสมดุลของข้อมูลด้วยวิธี SMOTE (Synthetic Minority Over-sampling Technique) และใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพการจำแนกของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) เมื่อทำการทดสอบและวัดประสิทธิภาพของโมเดลพบว่า เทคนิค Support Vector Machine ให้ผลดีที่สุดในการวิเคราะห์ความคิดเห็น โดยให้ค่าความถูกต้อง 99.65% ค่าความไว 99.30% และค่าความจำเพาะ 100%

คำสำคัญ : เหมืองข้อความ, ความคิดเห็น, เกมมือถือ

TITLE Opinion Analysis on PUBG Mobile Game Using Text Mining
AUTHOR Wossawat Inplang
ADVISORS Assistant Professor Jaree Thongkam , Ph.D.
DEGREE Master of Science **MAJOR** Information Technology
UNIVERSITY Mahasarakham University **YEAR** 2020

ABSTRACT

Text mining is a one of the most effective data analysis process by using alphabetically method. Currently, text mining techniques are used to classify a variety of ways. This research aims to fine the most effectiveness techniques from 4 techniques including Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor and C4.5. The data collected were all made by the viewers, in total of 3,798 messages. The categorization process has divided the data into 2 groups: positive character and negative character. Interestingly, the process has only indicated to selects adverbs and slangs as a core division to produce positive and negative characters. After analyzing the data, to problems were found class imbalanced. SMOTE were used to filtering and increase the minority class. 10-fold cross validation is applied to segment the data into training and testing sets. Moreover, accuracy sensitivity and specificity are used as the criteria for selecting the most effective model. The results presented the Support Vector Machine produce the most accuracy in categorizing the messages with accuracy of 99.65% sensitivity of 99.30% and specificity score of 100%

Keyword : Text mining, Opinion, Mobile game

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จลงได้ด้วยดี เนื่องจากได้รับความกรุณาอย่างสูงจาก ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ที่กรุณาให้คำแนะนำปรึกษาตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยตระหนักถึงความตั้งใจจริงและความทุ่มเทของอาจารย์ และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ผู้วิจัยมีความหวังว่า งานวิจัยฉบับนี้จะมีประโยชน์อยู่ไม่น้อย จึงขอมอบส่วนดีทั้งหมดนี้ ให้แก่เหล่าคณาจารย์ ที่ได้ให้คำแนะนำและแนวทางในการแก้ไขปัญหาและขอมอบความกตัญญูกตเวทิตาคุณแด่มารดาและครอบครัวผู้เป็นแรงผลักดันและกำลังใจให้สามารถผ่านพ้นปัญหาไปได้ด้วยดี และขอขอบคุณนางสาวฐิติกานต์ ภูอาราม ที่เป็นกำลังใจคอยช่วยเหลือสนับสนุนผู้วิจัยในทุกด้าน สำหรับข้อบกพร่องต่าง ๆ ที่อาจจะเกิดขึ้นนั้น ผู้วิจัยก็ขออภัย และยินดีที่จะรับฟังคำแนะนำจากทุกท่านที่ได้เข้ามาศึกษา เพื่อเป็นประโยชน์ในการพัฒนางานวิจัยต่อไป

วสวัตติ์ อินทร์แปลง

พหูน ปณฺฑิโต ชีเว

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญภาพ ..... ฌ

บทที่ 1 บทนำ 1

1.1 หลักการและเหตุผล 1

1.2 วัตถุประสงค์ของการวิจัย 3

1.3 ขอบเขตของการวิจัย 3

1.4 ความสำคัญของการวิจัย 3

1.5 นิยามศัพท์เฉพาะ 3

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4

2.1 เอกสารที่เกี่ยวข้อง 4

2.1.1 ความหมายและประเภทของเกม 4

2.1.2 PlayerUnknown's : Battlegrounds (PUBG) 7

2.1.3 การประมวลผลภาษาธรรมชาติ 9

2.1.4 Google Play 15

2.1.5 เทคนิคการทำเหมืองข้อความ 16

2.1.6 Synthetic Minority Over-sampling Technique 20

2.1.7 เทคนิคที่ใช้ในงานวิจัย 20

2.1.8 การวัดประสิทธิภาพของแบบจำลอง 26

2.2 งานวิจัยที่เกี่ยวข้อง 28
บทที่ 3 วิธีดำเนินการวิจัย 31
3.1 การเก็บรวบรวมข้อมูล 31
3.2 ขั้นตอนก่อนการสร้างแบบจำลอง 32
3.2.1 กระบวนการกลั่นกรองข้อมูล 33
3.2.2 กระบวนการตัดคำและกำจัดคำหยุด 33
3.2.3 กระบวนการเลือกคำบ่งชี้ 34
3.3 การสร้างแบบจำลอง 37
3.4 การวัดประสิทธิภาพของแบบจำลอง 38
บทที่ 4 ผลการวิจัยและการอภิปราย 39
4.1 วิธีการวัดผลการวิจัย 39
4.2 ผลการวิจัย 39
บทที่ 5 สรุปผล อภิปรายผล และข้อเสนอแนะ 44
5.1 สรุปผล 44
5.2 อภิปรายผล 45
5.3 ข้อเสนอแนะ 46
บรรณานุกรม 47
ภาคผนวก ก คำแสดงคุณลักษณะที่ใช้ในงานวิจัย 54
ประวัติผู้เขียน 58

พหุน ปณ ฑิโต ชีเว

สารบัญภาพ

ภาพที่ 2.1 สถาปัตยกรรมของการประมวลผลภาษาธรรมชาติ 11

ภาพที่ 2.2 ตัวอย่างการสร้างตัวแทนเอกสาร 14

ภาพที่ 2.3 กระบวนการเหมืองข้อความ 17

ภาพที่ 2.4 การวางตัวของข้อมูลในลักษณะเชิงเส้น 22

ภาพที่ 2.5 การจัดกลุ่มของเทคนิค k-NN 23

ภาพที่ 2.6 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation 26

ภาพที่ 2.7 Confusion Matrix 27

ภาพที่ 3.1 การเก็บข้อมูลผ่านโปรแกรม WebHarvy 31

ภาพที่ 3.2 ตัวอย่างการดึงข้อความความคิดเห็น 32

ภาพที่ 3.3 ข้อมูลที่อยู่ในรูปแบบไฟล์ CSV ที่ถูกเปิดด้วย Microsoft Excel 32

ภาพที่ 3.4 ตัวอย่างข้อความที่ถูกตัดคำ 33

ภาพที่ 3.5 ตัวอย่างประเภทของคำตามความหมายในพจนานุกรม 34

ภาพที่ 3.6 ตัวอย่างของคำวิเศษณ์และคำสแลงที่ใช้ในการบ่งชี้คุณลักษณะ 35

ภาพที่ 3.7 ตัวอย่างการแปลงข้อความให้เป็นฟีเจอร์เวกเตอร์ ด้วยค่า TF-IDF 36

ภาพที่ 3.8 ตัวอย่างการกำหนดคุณลักษณะ 36

ภาพที่ 3.9 ตัวอย่างของชุดข้อมูล 36

ภาพที่ 3.10 ข้อมูลที่ผ่านวิธี SMOTE แก้ปัญหาข้อมูลไม่สมดุล 37

ภาพที่ 4.1 ค่าความถูกต้อง 40

ภาพที่ 4.2 ค่าความไว 41

ภาพที่ 4.3 ค่าความจำเพาะ 42

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ปัจจุบันโทรศัพท์มือถือทำให้คนทุกเพศทุกวัยสามารถเข้าถึงเกมได้ง่ายและมากขึ้น การที่เกมถูกผสมผสานเข้ากับเทคโนโลยีต่าง ๆ ทำให้เกมบนมือถือได้รับความนิยมและมีความน่าสนใจมากขึ้น ซึ่งอุตสาหกรรมเกมเป็นอุตสาหกรรมหนึ่งที่มีมูลค่าทางเศรษฐกิจสูง ทำให้ผู้ผลิตและพัฒนาเกมต้องแข่งขันกันพัฒนาเกมมือถือออกมาให้มีคุณภาพและตรงกับความต้องการของผู้เล่นอยู่เสมอ PlayerUnknown's Battlegrounds : Mobile (PUBG Mobile) หรือเรียกสั้น ๆ ว่า ผับจี โมบาย เป็นเกมมือถือที่ดาวน์โหลดได้ฟรี ซึ่งได้รับความนิยมอย่างมากในปัจจุบัน [1] แม้ว่าเกมมือถือผับจีจะสามารถดาวน์โหลดได้ฟรี แต่ถ้าหากผู้เล่นต้องการซื้ออุปกรณ์หรือไอเทมภายในเกมเพิ่มเติม ก็จะต้องเสียค่าใช้จ่ายผ่านการเติมเงินเข้าระบบเพื่อนำมาซื้อไอเทมนั้น การพัฒนาเกมให้มีคุณภาพ ไม่มีข้อบกพร่อง (Bug) และตรงกับความต้องการของผู้เล่นจึงเป็นสิ่งจำเป็นอย่างยิ่งของผู้ผลิตและพัฒนาเกม ซึ่งแหล่งที่มาสำคัญเพื่อช่วยระบุคุณลักษณะของเกมคือ ความคิดเห็น (Opinion) ที่มาจากผู้เล่น ซึ่งผู้เล่นจะกล่าวถึงความพึงพอใจที่มีต่อเกม รวมถึงปัญหาที่ผู้เล่นพบเจอจากการเล่นเกม และความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจียังเป็นตัวแปรสำคัญที่มีผลต่อผู้เล่นเกมอื่น ๆ ในการตัดสินใจเลือกดาวน์โหลดเกมมือถือผับจีมาเล่น หรือตัดสินใจเติมเงินเข้าระบบเพื่อซื้อไอเทมภายในเกม [2] หากผู้ผลิตและพัฒนาเกมสามารถนำความคิดเห็นของผู้เล่นมาสร้างความพึงพอใจหรือนำมาปรับปรุงคุณภาพของเกมด้วยการสกัดข้อมูลจากความคิดเห็นของผู้เล่นที่มีต่อเกมได้ ก็จะสามารถพัฒนาเกมให้มีคุณภาพที่ดียิ่งขึ้น เป็นไปตามความต้องการของผู้เล่น และยังเป็นอีกช่องทางหนึ่งในการเข้าถึงผู้เล่นอีกทั้งยังสามารถแย่งชิงความได้เปรียบในการแข่งขันกับคู่แข่งทางการตลาดอีกด้วย อย่างไรก็ตาม เนื่องจากข้อความความคิดเห็นของผู้เล่นเกมมือถือผับจีนั้นมีจำนวนมาก และมีลักษณะเป็นภาษาธรรมชาติ (Natural Language) ทำให้ข้อความความคิดเห็นมีความแตกต่างกัน ซึ่งมีการใช้ภาษาที่ไม่ถูกต้องตามหลักไวยากรณ์ และมักไม่มีการสรุปว่าประโยคที่ผู้เล่นกล่าวถึงนั้นมีความคิดเห็นเป็นเชิงบวกหรือเชิงลบ ซึ่งหากต้องการทราบต้องเสียเวลาในการจำแนกเป็นอย่างมาก

การประมวลผลภาษาธรรมชาติ คือการแปลความจากภาษาธรรมชาติที่มนุษย์ใช้สื่อสารกันให้อยู่ในรูปแบบที่เป็นโครงสร้าง (Structured Data) ที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ 2 แนวทาง [3] คือ แนวทางการศึกษาและทำความเข้าใจกับโครงสร้างทางภาษาศาสตร์ และอีกแนวทางคืออาศัยความรู้ด้านปัญญาประดิษฐ์ โดยการแทนความรู้ (Knowledge Representation) ด้วยคลังคำ (Corpus) ซึ่งการวิเคราะห์ความคิดเห็นที่เป็นภาษาไทยเป็นเรื่องท้าทาย เนื่องจากข้อความแสดงความคิดเห็นส่วนใหญ่ที่อยู่บนอินเทอร์เน็ตนิยมใช้ภาษาที่มีโครงสร้างประโยคที่ไม่แน่นอน (Unstructured Data) หรือเป็นภาษาธรรมชาติ ไม่ถูกต้องตามหลักไวยากรณ์ทางภาษา ซึ่งภาษาไทยมีรูปแบบการ

เขียนคำยาวต่อกัน ไม่มีการเว้นวรรคระหว่างคำดังเช่นภาษาอังกฤษ ทำให้ยากต่อการวิเคราะห์ ทำให้ข้อความความคิดเห็นที่เป็นภาษาไทยนั้นจะต้องตัดประโยคออกเป็นคำก่อนทำการวิเคราะห์ เหมืองข้อความ (Text Mining) เป็นกระบวนการเพื่อสกัดเอาความรู้จากภาษาธรรมชาติที่มีลักษณะของข้อมูลแบบไม่มีโครงสร้าง อีกทั้งไม่มีการกำหนดรูปแบบไว้ล่วงหน้า [4] การประมวลผลภาษาธรรมชาติ คือการแปลความจาก

ในการทำเหมืองข้อความมีนักวิจัยหลายท่านนำเทคนิคเหมืองข้อมูลมาประยุกต์ในการวิเคราะห์ความคิดเห็น ซึ่งเป็นอีกกระบวนการของการวิเคราะห์เหมืองข้อความ [5] โดยการนำเอาความคิดเห็นมาทำการวิเคราะห์เพื่อให้ทราบถึงความพึงพอใจที่มีต่อสิ่งนั้น ๆ และการสกัดคำตามคุณลักษณะที่แตกต่างกันก็อาจได้มาซึ่งประโยชน์เพื่อการพิจารณาวิเคราะห์ที่หลากหลายและแม่นยำมากขึ้น เช่น ประพัฒน์ พรมน้ำอ่างและคณะ [6] ได้ทำการแบ่งกลุ่มข้อความจากข้อความรีวิว โดยใช้เทคนิคเหมืองข้อมูล ซึ่งประกอบด้วยเทคนิค SVM เทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naïve Bayes จากการทดลองพบว่าโดยเทคนิค SVM ได้ค่าความถูกต้องสูงที่สุดอยู่ที่ 86.26 % ส่วนพัชรนิ กานต์ พงษ์ธนู และคณะ [7] นำเสนอการวิเคราะห์เหมืองข้อความจากการเก็บข้อมูลการแสดงความพึงพอใจของลูกค้าบนเว็บไซต์ เพื่อหาแนวทางในการปรับปรุงการบริการของเว็บไซต์ผู้ให้บริการโรงแรมให้มีประสิทธิภาพมากขึ้น โดยใช้เทคนิควิธีต้นไม้ตัดสินใจ ID3 และเทคนิค Naive Bayes จากการทดลองพบว่าเทคนิคต้นไม้ตัดสินใจให้ค่าเฉลี่ยมากที่สุดที่ 95.50% Peiman Barnaghi และ John G. Breslin [8] ได้นำเสนอกระบวนการสร้างโมเดลในการวิเคราะห์ความคิดเห็นเพื่อค้นหาภาพสะท้อนความเชื่อมั่นของประชาชนต่อเหตุการณ์ต่าง ๆ เพื่อนำมาประยุกต์ใช้เป็นพื้นฐานสำหรับการคาดคะเนเหตุการณ์ในอนาคตผ่านการใช้งานบน Twitter โดยการใช้ความคิดเห็นเกี่ยวกับการจัดงาน FIFA World Cup 2014 โดยทำการพิจารณาความคิดเห็นที่เป็นเชิงบวกและเชิงลบ จากนั้นทำการเปรียบเทียบการสร้างโมเดลด้วยเทคนิค Bayesian Logistic Regression (BLR) และเทคนิค Naïve Bayes จากการทดลองพบว่าเทคนิค BLR ให้ผลที่ดิกว่าการใช้เทคนิค Naïve Bayes % แต่เทคนิคเหล่านี้ไม่สามารถสร้างแบบจำลองที่มีประสิทธิภาพได้ถ้ามีความไม่สมดุลของข้อมูล [9] โดยมีนักวิจัยหลายท่านได้ทำการแก้ปัญหาโดยใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มของข้อมูล SMOTE (Synthetic Minority Oversampling Technique:) เช่น งานวิจัยของ เชาวนันท์ โสโท [10] ได้สร้างแบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม และใช้ SMOTE มาทำการปรับสมดุลของข้อมูลแล้วทำการสร้างแบบจำลอง ผลการทดลองโครงข่ายประสาทเทียมที่มีการปรับความสมดุลของข้อมูลด้วยวิธีการ SMOTE มีประสิทธิภาพในการทำนายที่ดีกว่าวิธีอื่นโดยมีค่า ความถูกต้องเท่ากับ 81.70% ค่าความไวเท่ากับ 94.47% และค่าความจำเพาะเท่ากับ 55.47%

ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะพัฒนาแบบจำลองในการวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ โดยการนำข้อความความคิดเห็นของผู้เล่นเกมมือถือผับจีมาจำแนกเป็น 2 กลุ่ม คือ ความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ และปรับความไม่สมดุลของข้อมูลด้วยเทคนิค SMOTE และเปรียบเทียบเทคนิคที่ใช้ในการสร้างแบบจำลองด้วยเทคนิค Naïve Bayes เทคนิค SVM (Support Vector Machine) เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 ในการวัดประสิทธิภาพของแบบจำลองผู้วิจัยจะใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity)

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาถึงเทคนิคการทำเหมืองข้อความร่วมกับการประมวลผลภาษาธรรมชาติที่เป็นภาษาไทย

1.2.2 เพื่อพัฒนาแบบจำลองในการวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ

1.2.3 เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองในการวิเคราะห์ความคิดเห็นต่อเกมมือถือผับจีด้วยเหมืองข้อความ

1.3 ขอบเขตของการวิจัย

1.3.1 เก็บรวบรวมข้อความความคิดเห็นต่อเกมมือถือผับจี จำนวน 3,798 ข้อความ บน Google Play ระหว่างวันที่ 1 กุมภาพันธ์ 2562 – 15 กุมภาพันธ์ 2562

1.3.2 วิเคราะห์ข้อความความคิดเห็นที่เป็นภาษาไทยเท่านั้น

1.3.3 สร้างแบบจำลองด้วยเทคนิคในเหมืองข้อความในการวิเคราะห์ความคิดเห็นในรูปแบบของภาษาไทย ด้วยเทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5

1.3.4 วัดประสิทธิภาพการวิเคราะห์ของเทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 ด้วยค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity)

1.4 ความสำคัญของการวิจัย

1.4.1 ได้แบบจำลองในการจำแนกความคิดเห็นต่อเกมมือถือผับจีด้วยเทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5

1.4.2 ได้แบบจำลองที่มีประสิทธิภาพในการจำแนกความคิดเห็นที่มีต่อเกมมือถือผับจีของคนไทย

1.5 นิยามศัพท์เฉพาะ

1.5.1 ข้อความความคิดเห็น คือ ข้อความแสดงความคิดเห็นที่ประกอบด้วยข้อมูลอันเป็นข้อเท็จจริงกับการแสดงความคิดเห็นต่อเกมมือถือผับจี

1.5.2 การวิเคราะห์ความคิดเห็น หมายถึง การวิเคราะห์ข้อความที่แสดงความรู้สึกหรืออารมณ์ที่มีต่อการเล่นเกมมือถือผับจี โดยแบ่งออกเป็นความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้แบ่งออกได้เป็นเอกสารที่เกี่ยวข้องโดยประกอบด้วย ความหมายและประเภทของเกม PlayerUnknown's : Battlegrounds (PUBG) การประมวลผลภาษาธรรมชาติ Google Play การทำเหมืองข้อความ เทคนิคในเหมืองข้อความ และการวัดประสิทธิภาพของเทคนิคในเหมืองข้อความ และงานวิจัยที่เกี่ยวข้อง

2.1 เอกสารที่เกี่ยวข้อง

2.1.1 ความหมายและประเภทของเกม

เกม [11] หมายถึงการเล่นที่ผู้เล่นอยู่ภายใต้เงื่อนไขการควบคุมและมีจุดเริ่มต้นและจุดจบที่ชัดเจน ในเงื่อนไขที่จำกัดและอิสระ การเล่นเกมเป็นการมุ่งใช้ความสามารถในการแสดงความเหนือกว่าเพื่อเอาชนะหรือแข่งขันบนเงื่อนไขข้อจำกัดต่าง ๆ โดยเกมบนอุปกรณ์อิเล็กทรอนิกส์หรือคอมพิวเตอร์ อาจจะมีการนำมาประยุกต์เข้ากับระบบเครือข่ายคอมพิวเตอร์หรืออินเทอร์เน็ต ซึ่งจะทำให้มีการเชื่อมต่อโดยผ่านการให้บริการของบริษัท 2 ผู้ดูแลเกมออนไลน์ซึ่งจะใช้ซอฟต์แวร์ในการประมวลผลเช่นกัน แต่จะมีการส่งและรับข้อมูลจากเซิร์ฟเวอร์ผู้ให้บริการเกมออนไลน์ (Host Server) อยู่ตลอดเวลา จุดเด่นของเกมคอมพิวเตอร์ออนไลน์คือการที่บุคคลสามารถเข้ามาอยู่ในที่เดียวกันหรือเซิร์ฟเวอร์เดียวกันได้ ซึ่งจะถูกจำลองโลกในจินตนาการหนึ่งขึ้นโดยผู้เล่นสามารถเข้าไปพบปะผู้คนมากมายจากทั่วโลกได้ เกมออนไลน์ คือ เกมที่ต้องใช้ผ่านการเชื่อมต่อผ่านอินเทอร์เน็ตอยู่ตลอดเวลาเพื่อให้สามารถติดต่อกับเซิร์ฟเวอร์ของผู้ให้บริการ เกมออนไลน์ และลักษณะของเกมที่ออกแบบมาเพื่อให้ผู้เล่นนั้นสามารถเล่นได้แบบเป็นกลุ่ม และเล่นเดี่ยว [12] เกมออนไลน์มีส่วนที่คล้ายคลึงกับเกมหลายผู้เล่นผ่านระบบอินเทอร์เน็ต (Multiplayer) เกมออนไลน์ส่วนมากจะเป็นเกมแบบ MMO (Massive Multiplayer Online) หรือก็คือเกมหลายผู้เล่นที่รับจานวนผู้เล่นได้มหาศาลในพื้นที่ที่หนึ่ง [13] โดยแบ่งประเภทของเกมออกเป็น 10 ประเภท [14] ดังนี้

1. เกมแอคชั่น (Action Game) เป็นแนวเกมที่มีความหลากหลายในรูปแบบการเล่นแนวหนึ่ง หลัก ๆ คือการควบคุมตัวละครภายในเกมเพื่อดำเนินตามเนื้อเรื่องหรือแก้ไขปัญหาต่าง ๆ เพื่อให้ผ่านในแต่ละฉากหรือแต่ละระดับการเล่น โดยแนวเกม Action จะเป็นแนวที่มีการผสมผสานแนวเกมอื่น ๆ เข้าไว้ด้วยเพื่อเพิ่มอรรถรสและความสมจริงให้กับผู้เล่นได้มากยิ่งขึ้น ตัวอย่างแนวเกม Action ที่สามารถพบได้ในปัจจุบันนี้ คือ

1.1 Action Adventure โดดเด่นด้วยเนื้อเรื่องที่มีความใกล้เคียงกับภาพยนตร์หนึ่งเรื่องและคอมพิวเตอร์กราฟิกที่สวยงามและมีความสมจริง นอกจากการควบคุมตัวละครเพื่อบรรลุ

จุดมุ่งหมายต่าง ๆ ของตัวเกมแล้ว บางเกมอาจมีการพัฒนาตัวละครเพื่อเพิ่มความสามารถต่าง ๆ รวมถึงการแก้ไขปริศนาภายในเกมอีกด้วย ตัวอย่างของเกมในแนวนี้คือ Uncharted หรือ Tomb Raider

1.2 Action Horror มีความคล้ายคลึงกับแนว Action Adventure แต่เพิ่มความระทึกขวัญให้กับผู้เล่น ตัวอย่างของเกมแนวนี้คือ Biohazard (Resident Evil) หรือ Outlast

1.3 Action Side Scrolling เป็นแนว Action ที่ยังได้รับความนิยมมาจนถึงปัจจุบัน โดยต้องควบคุมตัวละครเพื่อผ่านอุปสรรคในแต่ละด่าน ในปัจจุบันอาจมีการเพิ่มการแก้ปริศนาในแต่ละด่านเข้าไปด้วย ตัวอย่างของเกมแนวนี้คือ Rockman (Mega Man) หรือ INSIDE

1.4 Action Open World โดดเด่นด้วยโลกเสมือนจริงภายในเกม ที่มีลักษณะเป็นเมืองสมมติขึ้น ทำให้สามารถสำรวจและทำภารกิจต่าง ๆ ภายในเกมได้อย่างหลากหลาย ตัวอย่างของเกมแนวนี้คือ Grand Thrift Auto (GTA) หรือ Watchdog

2. เกมเล่นตามบทบาท (Role-Playing Game) หรือ อาร์พีจี (RPG) เป็นเกมที่มีความโด่งดังในวงการเกม รู้จักกันในอีกชื่อหนึ่งคือ “เกมภาษา” เพราะมีชื่อเสียงมาจากเกมทางฝั่งของประเทศญี่ปุ่น ซึ่งต้องใช้ความรู้และความเข้าใจทางภาษาญี่ปุ่นในระดับหนึ่งจึงจะได้อรรถรสจากเกมแนวนี้เพิ่มมากขึ้น จนในภายหลังได้มีเกมแนวนี้ในรูปแบบภาษาอังกฤษออกมามากขึ้น จุดเด่นหนึ่งของเกมแนวนี้คือระบบการเล่น ลักษณะของเกมเป็นแบบ Open World และกราฟิกที่สวยงาม ตัวอย่างของเกมแนวนี้คือ Final Fantasy หรือ The Witcher ซึ่งเกมแนวนี้ก็จะแบ่งออกได้อีกสองแบบหลัก ๆ คือ

2.1 Turn Based RPG ลักษณะการต่อสู้จะเป็นรอบ โดยที่รอบของเราจะต้องวางแผนการโจมตีคู่ต่อสู้หรือเลือกใช้ความสามารถต่าง ๆ ได้เฉพาะในรอบของเราเท่านั้น จนจบรอบจึงสลับเป็นของคู่ต่อสู้ เป็นเช่นนี้ต่อเนื่องจนกว่าจะมีฝ่ายที่ได้รับชัยชนะ

2.2 Action RPG มีความยืดหยุ่นในการควบคุมตัวละครที่สูงขึ้น ไม่ต้องรอการโจมตีแบบเป็นรอบ สามารถบังคับและเลือกคำสั่งต่าง ๆ ให้กับตัวละครที่ควบคุมได้เลยเช่นเดียวกับเกมแนว Action แต่จะมีความซับซ้อนของคำสั่งต่าง ๆ มากกว่า

3. เกมปริศนา (Puzzle Game) เป็นเกมแนวลับสมองที่ฝึกความคิดให้มีความว่องไว โดยเงื่อนไขหลัก ๆ ของเกมแนวนี้คือการใช้ความคิดที่ว่องไวในเวลาที่กำจัด ยิ่งระดับของเกมที่สูงมากขึ้นเท่าไหร่ก็จะมีความท้าทายมากขึ้นเท่านั้น ซึ่งรูปแบบของการคิดเพื่อแก้ปริศนานี้ก็สามารถไปปรากฏในเกมแนว Action Adventure ได้เช่นเดียวกัน ตัวอย่างของเกมแนวนี้คือ Tetris หรือ Magical Drop

4. เกมจำลองการขับขี่ยานพาหนะ (Racing/Flight Simulation Game) เป็นแนวเกมที่จำลองการควบคุมยานพาหนะอย่างรถยนต์หรือเครื่องบิน ซึ่งบางเกมก็มีระบบที่สมจริงโดยใช้การ

อ้างอิงจากการควบคุมรถยนต์หรือเครื่องบินจริง ๆ ในปัจจุบันเกมแนวนี้ก็ได้รับการพัฒนาทางด้านภาพให้สมจริงมากยิ่งขึ้นด้วยเช่นกัน ตัวอย่างของเกมแนวนี้คือ Gran Turismo หรือ Ace Combat

5. เกมวางแผนการรบ (Strategy Game) เป็นเกมที่สามารถฝึกความคิดและการตัดสินใจรูปแบบของเกมจะมีเงื่อนไขที่แตกต่างกันออกไปในแต่ละด่าน หัวใจสำคัญของเกมแนวนี้คือการวางแผนกลยุทธ์ต่าง ๆ ให้สอดคล้องกับทรัพยากรที่ควบคุมอยู่ เช่น เงื่อนไขของเวลาหรือเงื่อนไขของทรัพยากรที่มีอย่างจำกัด เพื่อให้การเล่นมีความท้าทายมากยิ่งขึ้น ตัวอย่างของเกมแนวนี้คือ Romance of The Three Kingdoms หรือ SD Gundam G Generation ซึ่งเกมแนวนี้ก็จะแบ่งออกได้อีกสองแบบหลักๆคือ Real Time Strategy มีเงื่อนไขหลักคือเวลาและการบริหารทรัพยากรต่าง ๆ ที่มีอยู่ให้เพียงพอและต้านทานคู่ต่อสู้ให้ได้ และ Turn Based Strategy มีลักษณะที่ใกล้เคียงกับ Turn Based RPG คือข้อจำกัดเรื่องรอบในการวางแผนและความคุมทรัพยากรที่มีอยู่

6. เกมกีฬา (Sport Game) คือ เกมแนวกีฬาที่อ้างอิงกฎและกติกาจากกีฬานั้นเพื่อความสมจริง เกมกีฬาบางเกมก็อาจเป็นเกมแนว Action ได้ หรือเกมกีฬาบางเกมก็ให้สวมบทบาทเป็นผู้จัดการทีม ควบคุมทีมและบริหารทีมตามบทบาท ซึ่งเป็นเกมแนวกีฬาที่สอดแทรกแนว Strategy รวมเข้าไว้ด้วย ตัวอย่างของเกมแนวนี้คือ PES (Pro Evolution Soccer) หรือ NBA

7. เกมอาเขต (Arcade Game) เป็นแนวเกมที่มีความหลากหลายอยู่ภายในตัว ส่วนมากจะเป็นเกมที่ใช้เวลาไม่นานมากในการเล่น บางเกมก็สามารถเล่นได้หลายผู้เล่นพร้อม ๆ กัน จุดมุ่งหมายหลักของเกมแนวนี้คือการทำคะแนนให้ได้มากที่สุดซึ่งบางเกมก็อาจจะมีเรื่องของเวลาเข้ามาเป็นอุปสรรคด้วย แนวเกมนี้มีที่มาจากการเล่นเกมที่ Game Center ในสมัยก่อน บางเกมก็ได้รับการพัฒนากราฟิกให้ร่วมสมัยมากขึ้นด้วย ตัวอย่างของเกมแนวนี้คือ Space Invader หรือ Pac-man

8. เกมต่อสู้ (Fighting Game) คือเกมที่เป็นลักษณะการควบคุมตัวละครและมีการกดคำสั่งเพื่อปล่อยท่าทางต่าง ๆ รวมไปถึงการใช้จังหวะเพื่อสร้างความต่อเนื่องในการสร้างความเสียหายและลดพลังของคู่ต่อสู้ โดยเกมแนวนี้มีทั้งรูปแบบสองมิติและสามมิติ ตัวอย่างของเกมแนวนี้คือ Street Fighter หรือ Tekken

9. เกมยิงปืน (Shooting Game) ในปัจจุบันได้เสริมความเป็น Action และมีเนื้อเรื่องที่สมจริงมากขึ้น จากแต่เดิมที่ต้องยิงปืนตามเป้าหมายที่แตกต่างกันไปในแต่ละด่านเพียงอย่างเดียว ซึ่งเกมแนวนี้ก็จะแบ่งออกได้อีกสองแบบหลัก ๆ คือ First Person Shooting (FPS) หรือเกมยิงปืนในรูปแบบมุมมองบุคคลที่หนึ่ง จุดเด่นคือเนื้อเรื่องที่มีความสมจริงโดยอ้างอิงจากเหตุการณ์ต่าง ๆ ที่เกิดขึ้นจริงหรืออ้างอิงจากนิยายวิทยาศาสตร์ มุมมองภายในเกมจะเหมือนมองเห็นโดยผ่านจากสายตาของตัวละครที่ถูกควบคุมอยู่ ตัวอย่างของเกมแนวนี้คือ Call of Duty หรือ Battlefield และ Third Person Shooting (TPS) ที่ตัวเกมมีลักษณะเช่นเดียวกับ FPS แต่จะเป็นมุมมองจากบุคคลที่สามซึ่งจะทำให้ตัวตัวละครที่ควบคุมอยู่ได้เต็มตัว ตัวอย่างของเกมแนวนี้คือ Fortnite

10. เกมดนตรี (Music Game) จุดเด่นของเกมแนวนี้คือเสียงเพลงหรือการประกอบจังหวะกับเพลงภายในเกม บางเกมก็เป็นการจำลองรูปแบบของเครื่องดนตรีบางอย่าง หรือการกำหนดให้ออกท่าทางตามจังหวะของเพลง จุดมุ่งหมายคือการทำคะแนนให้สูงที่สุดโดยใช้ทักษะจากการจดจำหรือจับจังหวะ ตัวอย่างของเกมแนวนี้คือ Guitar Hero หรือ DJMAX

2.1.2 PlayerUnknown's : Battlegrounds (PUBG)

PlayerUnknown's : Battlegrounds (PUBG) เป็นเกมเอาชีวิตรอด (Survival Game) แบบออนไลน์หลายผู้เล่น (Online Multiplayer) จัดอยู่ในประเภทเกมแอคชั่น (Action Game) พัฒนาและเผยแพร่โดย PUBG Corporation ซึ่งเป็นบริษัทย่อยของ Bluehole บริษัทวิดีโอเกมของเกาหลีใต้ เกมดังกล่าวมีพื้นฐานมาจาก mods ก่อนหน้านี้ที่สร้างขึ้นโดย Brendan Greene ผู้ใช้ชื่อในโลกอินเทอร์เน็ตว่า "PlayerUnknown" ซึ่งได้แรงบันดาลใจจากภาพยนตร์เรื่อง Battle Royale ของญี่ปุ่นปี 2000 และขยายเป็นเกมเดี่ยว (Standalone) ภายใต้ทิศทางการสร้างสรรค์ของ Greene ในเกม PUBG จะมีผู้เล่นถึง 100 คนกระโดดลงมาจากเครื่องบินพร้อมร่มชูชีพลงบนแผนที่ที่เป็นเกาะร้างและวิ่งหาอาวุธและอุปกรณ์เพื่อฆ่าคนอื่นในขณะที่ต้องหลีกเลี่ยงไม่ให้ตัวเองถูกฆ่าตาย พื้นที่ปลอดภัย (Safe Zone) ที่มีอยู่ในแผนที่ของเกมจะลดขนาดลงเรื่อย ๆ เมื่อเวลาผ่านไป ซึ่งจะบีบผู้เล่นที่รอดชีวิตมาสู่พื้นที่ที่จำกัดมากขึ้นเพื่อบังคับให้เผชิญหน้ากัน ผู้เล่นหรือทีมใดอยู่รอดเป็นทีมสุดท้ายก็คือผู้ชนะ

PUBG เปิดตัวครั้งแรกในเดือนมีนาคม 2560 บนระบบปฏิบัติการ Microsoft Windows ผ่านทางโปรแกรม Steam โดยตัวเกมอยู่ระหว่างการพัฒนาและเปิดให้ผู้เล่นได้ทดลองเล่นไปด้วย (Early access beta) PUBG เปิดตัวเต็มรูปแบบในวันที่ 20 ธันวาคม 2560 โดย Microsoft Studios สำหรับ Xbox One ผ่านทางโปรแกรม Xbox Game Preview และเปิดตัวอย่างเป็นทางการในเดือนกันยายน 2561 สำหรับเวอร์ชั่นมือถือที่เล่นได้ฟรีบน Android และ iOS PUBG เป็นหนึ่งในวิดีโอเกมที่ขายดีที่สุดและมีผู้เล่นมากที่สุดตลอดกาล มียอดขายมากกว่าห้าสิบล้านไลเซน (License) ทั่วโลกภายในเดือนมิถุนายน 2561 โดยมีผู้เล่นมากกว่า 400 ล้านคน

PUBG คือเกมยิง (Shooter Game) ซึ่งมีผู้เล่นมากถึงหนึ่งร้อยคนต่อสู้ประจัญบาน (Battle Royale) กันในแผนที่ที่กว้างใหญ่ โดยผู้เล่นสามารถเลือกที่จะเล่นโหมดการแข่งขันแบบเดี่ยว, คู่, หรือทีมสี่คนได้ ผู้เล่นหรือทีมสุดท้ายที่ยังมีชีวิตอยู่คือผู้ชนะการแข่งขัน

การแข่งขันแต่ละครั้งเริ่มต้นด้วยผู้เล่นจะกระโดดร่มลงมาจากเครื่องบินไปยังหนึ่งในสี่แผนที่ โดยมีพื้นที่ประมาณ 8 × 8 กิโลเมตร (5.0 × 5.0 ไมล์), 6 × 6 กิโลเมตร (3.7 × 3.7 ไมล์) และ 4 × 4 กิโลเมตร (2.5 × ขนาด 2.5 ไมล์) [15] เส้นทางการบินของเครื่องบินบนแผนที่จะแตกต่างกันไปในแต่ละรอบทำให้ผู้เล่นต้องกำหนดเวลาที่ดีที่สุดในการดีดออกและกระโดดร่มลงสู่พื้น [16] ผู้เล่นเริ่มต้นด้วยการที่ไม่มีอุปกรณ์ใด ๆ รวมถึงเสื้อผ้า (ผู้เล่นสามารถปรับแต่งเครื่องแต่งกายได้เองจากหน้า Lobby ก่อนเริ่มเกมซึ่งไม่มีผลต่อการเล่นเกม) ผู้เล่นสามารถค้นหาอาวุธ ยานเกราะ ชุดเกราะ และ

อุปกรณ์อื่น ๆ ในอาคาร เมืองร้าง และพื้นที่อื่นอื่น ๆ ซึ่งไอเท็มเหล่านี้จะกระจายทั่วทั้งแผนที่เมื่อเริ่มต้นการแข่งขันโดยโซนที่มีความเสี่ยงสูงมักจะมีอุปกรณ์ที่ดีกว่า [16] เมื่อผู้เล่นสังหารผู้เล่นอื่นจะสามารถเก็บอุปกรณ์ของผู้เล่นที่ตายแล้วได้เช่นกัน [16] ผู้เล่นสามารถเลือกเล่นจากมุมมองบุคคลที่หนึ่งหรือบุคคลที่สาม ซึ่งแต่ละอย่างมีข้อดีและข้อเสียของตัวเองในการต่อสู้ และการรับรู้สถานการณ์สภาพแวดล้อมรอบตัว ทั้งนี้การตั้งค่าเซิร์ฟเวอร์สามารถบังคับผู้เล่นทุกคนให้ใช้มุมมองเดียวได้เพื่อกำจัดข้อได้เปรียบบางอย่างออกไป

ทุก ๆ สองถึงสามนาทีพื้นที่ที่สามารถเล่นได้ของแผนที่จะเริ่มหดตัวลง โดยสุ่มตำแหน่งซึ่งโดยผู้เล่นคนใดก็ตามที่ติดอยู่นอกพื้นที่ปลอดภัยจะได้รับความเสียหายเพิ่มขึ้น และในที่สุดก็จะถูกกำจัดหากไม่มีการเข้ามายังโซนปลอดภัย (Safe Zone) ในเกมผู้เล่นจะมองเห็นขอบเขตเป็นกำแพงสีฟ้าที่ส่องแสงระยิบระยับ เมื่อเวลาผ่านไป [17] โซนปลอดภัยจะถูกจำกัดมากขึ้นเพื่อบีบให้ผู้เล่นเผชิญหน้ากัน [16] และในระหว่างการแข่งขันจะมีการไฮไลต์บริเวณที่สุ่มของแผนที่เป็นสีแดงและถูกวางระเบิดซึ่งเป็นการคุกคามผู้เล่นที่ยังคงอยู่ในพื้นที่นั้น [18] โดยทั้งสองกรณีผู้เล่นจะได้รับการเตือนล่วงหน้าก่อน 2-3 นาที เพื่อให้ผู้เล่นมีเวลาที่จะย้ายไปยังที่ปลอดภัย [19] นอกจากนี้ยังมีเครื่องบินบินข้ามส่วนต่าง ๆ ของแผนที่ที่สามารถเล่นได้ และทิ้งหีบห่อบรรจุสิ่งของที่ไม่สามารถหาได้ในระหว่างการเล่นเกมปกติ (Special Item) โดยหีบห่อเหล่านี้จะปล่อยควันสีแดงที่มองเห็นได้ชัดเจนดึงดูดผู้เล่นที่สนใจเข้ามาใกล้ และทำให้ผู้เล่นต้องเผชิญหน้ากันเพื่อแย่งชิงสิ่งของเหล่านี้ [16] โดยเฉลี่ยแล้วใช้เวลาไม่เกิน 30 นาทีต่อรอบ [19] เมื่อเสร็จสิ้นแต่ละรอบผู้เล่นจะได้รับสกุลเงินในเกมตามประสิทธิภาพ สกุลเงินที่ใช้ในการซื้อลังซึ่งบรรจุรายการเครื่องแต่งกายสำหรับปรับแต่งตัวละครหรืออาวุธ [20]

ตามข้อตกลงการเผยแพร่ของจีนในเวอร์ชั่น Windows Tencent Games และ PUBG Corporation ได้ประกาศเพิ่มเติมว่าพวกเขาวางแผนที่จะปล่อยเกมเวอร์ชันโทรศัพท์มือถือในประเทศ 2 เวอร์ชัน คือ[21] [22] 1) PUBG: *Exhilarating Battlefield* เป็นฉบับย่อของเกมดั้งเดิมและได้รับการพัฒนาโดย Lightspeed & Quantum Studio แผนกภายในของ Tencent Games [23] 2) PUBG: Army Attack มีองค์ประกอบแบบอาร์เคดมากกว่ารวมถึงฉากแอ็คชั่นบนเรือรบและได้รับการพัฒนาโดย Timi Studio ของ Tencent ทั้งสองเวอร์ชันเปิดให้เล่นฟรีและเปิดตัวสำหรับอุปกรณ์ Android และ iOS ในวันที่ 9 กุมภาพันธ์ 2561 [24] [25] เกมดังกล่าวมียอดรวมการลงทะเบียนล่วงหน้า 75 ล้านครั้งและติดอันดับที่หนึ่งและสองในชาร์ตดาวน์โหลด iOS ของจีนเมื่อเปิดตัว [26] หลังจากการเปิดตัวที่ในแคนาดา PUBG: *Exhilarating Battlefield* ก็ถูกเรียกตามภาษาปากอย่างง่าย ๆ ว่า PUBG Mobile และได้เปิดตัวทั่วโลกอย่างเป็นทางการในวันที่ 19 มีนาคม 2561 [26] [27] [28]

2.1.3 การประมวลผลภาษาธรรมชาติ

การประมวลภาษาธรรมชาติ (Natural Language Processing : NLP) เป็นกระบวนการที่จะทำให้คอมพิวเตอร์เข้าในภาษาของมนุษย์ โดยใช้วิธีการรับอินพุตเป็นข้อความในภาษาใด ๆ แล้วสามารถโต้ตอบกลับมาได้ด้วยภาษาของผู้ใช้งานเองได้ [29] ภาษาธรรมชาติ (Natural Language) เป็นเครื่องมือที่มนุษย์ใช้ในการติดต่อสื่อสารเพื่อถ่ายทอดความรู้สึกต่าง ๆ ในชีวิตประจำวัน ภาษาธรรมชาติเป็นภาษาที่ไม่สามารถควบคุมให้มีลักษณะจำกัดยอมให้มีความผิดพลาดได้ มีความกำกวมและเป็นภาษาที่สื่อความหมายได้โดยไม่จำเป็นต้องมีความชัดเจน [30] ซึ่งภาษาธรรมชาติมี 2 รูปแบบ คือ ภาษาพูดและภาษาเขียนโดยได้มีการแบ่งแยกหน่วยย่อยของภาษาออกเป็น 3 ส่วนด้วยกัน คือ

ตัวอักษร คือ สัญลักษณ์แทนเสียง แทนความรู้สึก

คำ คือ กลุ่มของตัวอักษรที่เรียงต่อกันเป็นคำ มีความหมายของคำ

ประโยค คือ กลุ่มคำที่นำมาเรียงต่อกันเพื่อแทนความหมาย หรือเป็นข้อความสื่อสาร

ซึ่งจากลักษณะของภาษาธรรมชาติทำให้การประมวลผลนั้นเกิดปัญหาที่หลากหลายในกระบวนการวิเคราะห์และมีความซับซ้อนตามลักษณะของกลุ่มคำและรูปประโยค และปัญหาที่พบมากในการประมวลผลภาษาธรรมชาตินั้น มีอยู่ 3 ลักษณะ คือ

1. ความไม่มีขอบเขตจำกัด กล่าวคือ ภาษาธรรมชาตินั้นมีการคิดค้นหรือเกิดคำใหม่อยู่เสมอ ซึ่งได้แก่ ชื่อเฉพาะ ศัพท์เทคนิค คำประสมและคำย่อ

2. การยอมให้มีข้อผิดพลาด ซึ่งมนุษย์สามารถที่จะเข้าใจภาษาได้โดยง่าย ถึงแม้ว่าในภาษาที่ใช้นั้นจะมีความผิดพลาดหรือการผิดไปจากหลักการใช้งานของภาษานั้น ๆ

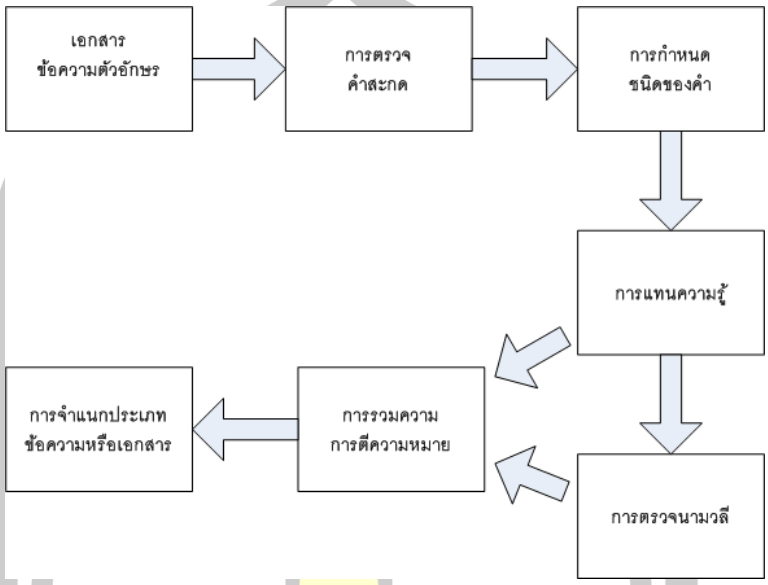
3. ความกำกวมของภาษา เนื่องจากว่าภาษาธรรมชาติมีความกำกวมในทุกระดับปะเภทของการใช้งานไม่ว่าจะเป็นในระดับคำ ระดับประโยค และระดับความหมาย

นอกจากนี้ยังมีเครื่องมือทางภาษาที่ทำให้ภาษาธรรมชาติมีความกระชับมากขึ้น เช่น การใช้คำอ้างอิงหรือการละคำประโยค ซึ่งในการประมวลผลด้วยคอมพิวเตอร์จำเป็นจะต้องมีเทคนิคและกลไกต่าง ๆ รวมทั้งการใช้ฐานความรู้ที่แตกต่างกันทางภาษาเพื่อให้สามารถประมวลผลทางภาษาได้อย่างมีประสิทธิภาพ ซึ่งจากลักษณะของการประมวลผลภาษาธรรมชาติในลักษณะของภาษาไทยนั้นจะมีความซับซ้อนที่แตกต่างจากภาษาอังกฤษ กล่าวคือ คำหรือวลีที่ใช้ในประโยคอาจมีหลายความหมายและมีความกำกวมของคำรวมอยู่ด้วย เป็นต้น

มาลินี นาคใหญ่ [31] ได้พัฒนาตัวแบบของการวิเคราะห์ประโยคข้อมูลด้วยการใช้เทคนิคของการประมวลผลภาษาธรรมชาติร่วมกับเทคนิคการจำแนกแบบต้นไม้เพื่อวิเคราะห์กลุ่มคำ ชนิดของกลุ่มคำ ความหมายของประโยค และใช้เทคนิคการทำเหมืองข้อมูลสร้างความสัมพันธ์ระหว่างกลุ่มคำ ความหมายและรูปประโยค ซึ่งจากผลการทดลองพบว่า กฎที่สร้างขึ้นด้วยเทคนิคจากการประมวลผลภาษาธรรมชาติสำหรับการวิเคราะห์กลุ่มคำ ชนิดของกลุ่มคำ ความหมายของประโยค สามารถวิเคราะห์ประโยคได้ถูกต้องซึ่งผลการวิจัยดังกล่าวจะสามารถนำไปประยุกต์ใช้กับการวิเคราะห์ประโยคข้อมูลประวัติศาสตร์ของไทยได้ สายัณห์ เทพแดง [32] ได้ศึกษาถึงการเปรียบเทียบการใช้คำหรือไม่ใช้คำนำหน้าชื่อ ซึ่งใช้วิธีการจดจำนิพจน์ระบุนามในภาษาไทย ซึ่งเป็นอีกลักษณะวิธีการประมวลผลภาษาธรรมชาติ ซึ่งได้เสนอวิธีการสร้างโมเดลเพื่อทำการเปรียบเทียบการใช้และการไม่ใช้คำนำหน้าชื่อหรือคำต่อท้าย โดยการใช้เทคนิควิธีจดจำนิพจน์ระบุนามในภาษาไทย โดยเปรียบเทียบโมเดล โดยการสร้างกฎในการเรียนรู้ ด้วยวิธีคอนดิชันนอลแรนดอมฟิลด์ และสร้างโมเดลโดยใช้คลังข้อความ จำนวน 250,000 คำ ผลการทดสอบประสิทธิภาพของโมเดลนั้น มีค่าสูงสุดคือชื่อองค์กร 86.80 % และมีค่าต่ำสุดชื่อบุคคล 85.36% ส่วนโมเดลที่ไม่ใช้ มีค่าสูงสุดคือชื่อองค์กร 84.70 % มีค่าต่ำสุดคือชื่อบุคคล 81.81% สรวัตร ประภานิติเสถียร และ ไกรศักดิ์ เกษร [33] ได้นำเสนอการนำหลักไวยากรณ์และการสร้างตารางความน่าจะเป็นแบบ 5 และ 3 แกรม (N-gram) ที่สร้างจากตัวแบบด้วยเทคนิคต้นไม้ตัดสินใจในการปรับปรุงโครงสร้างของประโยคและการใช้เทคนิค Semantic Role Labeling ร่วมกับการให้ค่าน้ำหนักของคำในการเปรียบเทียบเชิงความหมาย จากการทดลองพบว่าเมื่อทำการปรับปรุงโครงสร้างของประโยคแล้ว ทำให้มีประสิทธิภาพในการตรวจจับมากยิ่งขึ้นเมื่อเทียบกับการใช้เทคนิค Semantic Role Labeling ร่วมกับการให้ค่าน้ำหนักของคำเพียงอย่างเดียว กรมวุฒิ นงนุช และคณะ [34] ได้นำเสนอระบบช่วยสรุปบทความวิจัยที่มีการตีพิมพ์แบบอัตโนมัติ โดยใช้วิธีการดึงเอาบทความจากแหล่งเผยแพร่บทความที่มีการตีพิมพ์แบบสาธารณะเช่น Science Direct PubMed IEEE เป็นต้น ใช้วิธีทาง XML Call API Web Service และ Web Content จากการเขียนโปรแกรมแบบ Crontab Daily ซึ่งจากทดลองพบว่าการใช้กระบวนการภาษาธรรมชาติ ในการสรุปข้อมูลของเอกสารที่สำคัญจะช่วย ลดระยะเวลาในการศึกษา งานวิจัยที่มีความยากในการอ่านหรือแปลความหมายของบทความนั้น

ในการประมวลผลด้วยภาษาธรรมชาตินั้น มีกระบวนการตามโครงสร้างของสถาปัตยกรรมการประมวลผลภาษาธรรมชาติ ซึ่งประกอบด้วย การทำความเข้าในภาษา (Natural

Language Understanding) และการสร้างภาษา (Natural Language Generation) ดังแสดงในภาพที่ 2.1



ภาพที่ 2.1 สถาปัตยกรรมของการประมวลผลภาษาธรรมชาติ

ในการประมวลผลด้วยภาษาธรรมชาตินั้น มีกระบวนการตามโครงสร้างของสถาปัตยกรรมการประมวลผลภาษาธรรมชาติ ซึ่งประกอบด้วย อินพุตซึ่งมีหน้าที่ในการรับเอาข้อมูลที่มีลักษณะเป็นข้อความหรือเอกสารข้อความ ซึ่งเมื่อรับข้อมูลเอกสารแล้วจะผ่านเข้าสู่กระบวนการตรวจคำสะกดต่าง ๆ เป็นการลักษณะของการวิเคราะห์องค์ประกอบของคำว่าคำคำนั้นสามารถแบ่งออกเป็นคำใดได้บ้างโดยยังไม่สนใจความหมาย และเมื่อเข้าสู่กระบวนการกำหนดชนิดของคำ คำที่ได้จะมีการตรวจสอบลักษณะความหมายตามหลักของไวยากรณ์เพื่อทำการกำหนดหน้าที่ของคำและชนิดของคำเหล่านั้นต่อประโยคว่ามีลักษณะเป็นคำกริยา คำนามหรือคำสรรพนามเป็นต้น ในส่วนของกระบวนการแทนความรู้จะมีการนำมาใช้เมื่อประโยคมีความซับซ้อนกล่าวคือมีความกำกวมของคำหรือความหมายโดยลักษณะของการแทนความรู้ในภาษาธรรมชาตินั้นจะใช้วิธีการตีความจากสถานการณ์จากนั้นจะมีการให้เหตุผลต่อสถานการณ์นั้น ๆ ว่าควรมีความในลักษณะใด ซึ่งกระบวนการแทนความรู้ในภาษาธรรมชาตินั้นจะมีวิธีการที่ใช้ในการตัดสินสถานการณ์เพื่อแทนความรู้อยู่หลากหลายวิธีด้วยกัน กระบวนการตรวจนามวลี เป็นกระบวนการตรวจสอบความหมายของคำที่มีลักษณะเป็นกลุ่มคำ โดยในแต่ละวลีนั้นจะประกอบไปด้วยคำที่มีความหมายแตกต่างกันแต่เมื่อนำมาประกอบกันจะมีลักษณะทางความหมายที่แตกต่างออกไปจากเดิม จากนั้นเมื่อเข้าสู่กระบวนการรวม

ความหรือการตีความหมายในส่วนนี้จะทำหน้าที่ในการรับเอาองค์ความรู้ที่ได้มีการกำหนดไว้เพื่อทำการตีความหมายตามลักษณะของโครงสร้างของคำหรือตามลักษณะของไวยากรณ์ที่ได้มีการกำหนดไว้แล้ว และจากกระบวนการที่ผ่านมาทั้งหมดนี้เมื่อเข้าสู่กระบวนการสุดท้ายนั้นก็คือการจำแนกประเภทข้อความหรือเอกสารตามหมวดหมู่ที่ได้กำหนดไว้ โดยในการทำความเข้าใจในการประมวลผลภาษาธรรมชาตินั้น มีขั้นตอนดังต่อไปนี้ คือ

1. การวิเคราะห์ทางองค์ประกอบ เป็นกระบวนการวิเคราะห์ในหน่วยคำ ว่าคำนั้น ๆ สามารถแยกย่อยออกเป็นคำอื่นใดได้บ้าง

2. การวิเคราะห์ทางวากยสัมพันธ์ เป็นการวิเคราะห์ทางไวยากรณ์ ซึ่งเป็นวิธีที่ต้องการดูว่าประโยคที่ได้มีการรับเข้ามานั้นซึ่งประกอบด้วยคำหลายๆ คำ การเรียงต่อกันนั้นมีโครงสร้างเชิงวากยสัมพันธ์อย่างไร โดยจำทำการวิเคราะห์ตามลำดับของคำให้อยู่ในโครงสร้างตามที่ได้กำหนดไว้

3. การวิเคราะห์ทางความหมาย เป็นการวิเคราะห์ความหมายเมื่อได้โครงสร้างทางไวยากรณ์แล้วจากนั้นจะทำการกำหนดค่าของคำแต่ละคำว่าหมายถึงอะไร

4. บูรณาการทางวจนิพนธ์ เป็นการพิจารณาความหมายของประโยคซึ่งดูได้จากประโยคข้างเคียง เนื่องจากลักษณะของภาษานั้นจะสามารถทำความเข้าใจในคำในประโยคได้เมื่อมีประโยคเพื่อโยงความหมายทั้งประโยคก่อนหน้าหรือประโยคตามหลัง

5. การวิเคราะห์ทางปฏิบัติ เป็นการแปลความหมายของประโยคใหม่อีกครั้งเพื่อให้ทราบถึงความหมายของการสื่อภาษาที่ชัดเจน

จากกระบวนการวิเคราะห์เหมืองข้อความดังกล่าวได้นำมาประยุกต์ใช้ในงานวิจัยนี้เพื่อการวิเคราะห์ความคิดเห็นในรูปแบบของภาษาไทย ซึ่งจำเป็นต้องใช้วิธีการประมวลผลภาษาธรรมชาติโดยอาศัยความรู้ในการเข้าใจภาษาธรรมชาติอยู่ด้วยกัน 6 อย่างด้วยกัน คือ

2.1.3.1 การแทนความรู้

การแทนความรู้ เป็นกระบวนการหนึ่งในการวิเคราะห์ข้อความโดยจะอาศัยการวิเคราะห์จากคำเป็นหลัก เรียกว่าการแทนข้อความด้วยถุงคำ (Bag of word: Bow) การแทนความรู้ด้วยถุงคำเป็นการสร้างตัวแทนอย่างง่ายเพื่อใช้ในการประมวลผลภาษาธรรมชาติ ในรูปแบบข้อความประโยคหรือเอกสาร โดยจะแสดงเป็นชุดของคำในถุงคำนั้นโดยไม่คำนึงถึงไวยากรณ์ รูปแบบของถุงคำเป็นที่นิยมใช้ในวิธีการของการจำแนกประเภทเอกสารโดยการใช้วิธีการเกิดขึ้นของแต่ละคำโดยมีวิธีที่ใช้ในการคำนวณค่าเพื่อกำหนดให้เป็นตัวแทนข้อความการสกัดคุณลักษณะ ได้แก่

1. การแทนคำด้วยค่าการเกิดขึ้นหรือไม่เกิดขึ้นของคำ (Binary weighting) โดย

$$binary\ weighting = \begin{cases} 1 \\ 0 \end{cases}$$

เมื่อ 1 คือ แทนคำที่กำหนดโดยเกิดขึ้นในเอกสาร

0 คือ แทนคำที่กำหนดโดยไม่เกิดในเอกสาร

2. การแทนด้วยค่าความถี่ของคำ (Term Frequency: TF) คือ การหาคำที่เกิดขึ้นในเอกสารจากความถี่ของคำนั้น ๆ ในเอกสารทั้งหมดซึ่งเป็นวิธีการที่ได้รับความนิยมและมีประสิทธิภาพพอสมควร แต่วิธีการนี้ยังถือว่ามีข้อบกพร่องอยู่บ้าง เพราะเมื่อเกิดค่าความถี่ของคำสูง นั่นก็คือ คำ ๆ นั้นมีโอกาสเกิดขึ้นในหลายข้อความพร้อมกัน ซึ่งอาจเกิดข้อผิดพลาดต่อการวิเคราะห์เพื่อการจำแนกข้อความ

3. การแทนด้วยค่าความถี่คำและค่าส่วนกลับความถี่เอกสารที่เกิดคำ (Term Frequency – Inverse Document Frequency: TF-IDF) ซึ่งวิธีการแทนค่าข้อความด้วยค่า TF-IDF เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์ที่มีความน่าเชื่อถือและได้รับความนิยมเป็นอย่างมาก โดยจะมีการหาค่าความถี่ของคำที่ปรากฏในเอกสารและจากนั้นจะทำการหาค่าส่วนกลับของเอกสารหรือระบบน้ำหนักของความถี่เอกสารผกผัน โดยสามารถคำนวณได้จากสมการ 2-1

$$\mathrm{idf} = 1 + \log(\mathrm{N}|\mathrm{df}) \qquad (2\text{-}1)$$

เมื่อ N คือ จำนวนเอกสารทั้งหมดที่อยู่ในกลุ่ม

df คือ จำนวนเอกสารที่มีคำ ๆ นั้นปรากฏอยู่

ดังนั้นจะได้ดังสมการที่ 2-2

$$\mathrm{tf} - \mathrm{idf} = \mathrm{tf} \times \mathrm{idf} \qquad (2\text{-}2)$$

2.1.3.2 การสกัดคุณลักษณะ

กระบวนการในการสกัดคุณลักษณะเป็นกระบวนการเบื้องต้นในการคัดเลือกคำสำคัญ ซึ่งเป็นวิธีการพื้นฐานที่นิยมใช้ในการประมวลผลภาษาธรรมชาติ โดยเทคนิคที่นิยมใช้กันมากคือ การตัดคำที่ไม่มีนัยสำคัญหรือคำหยุด (stop word) ออกไป เนื่องจากว่าเป็นกลุ่มคำที่ไม่สามารถนำมาใช้ในการสลัดคุณลักษณะได้ ซึ่งได้แก่ คำบุพบท คำสันธาน คำสรรพนาม คำลักษณะนาม จำนวนนับ ตัวเลข และคำลงท้ายประโยค เป็นต้น คำหยุดมักเป็นคำที่เกิดขึ้นบ่อยในเอกสารและการกำจัดคำหยุดเหล่านี้จะไม่ทำให้ความหมายของเอกสารนั้นเปลี่ยนไป

2.1.3.3 การสร้างตัวแทนเอกสาร

การสร้างตัวแทนเอกสารเป็นวิธีการเตรียมการวิเคราะห์ข้อความที่มีโครงสร้างไม่แน่นอนโดยจำเป็นต้องแปลงให้อยู่ในรูปแบบที่มีโครงสร้างก่อน เพื่อให้คอมพิวเตอร์สามารถนำไปประมวลผลได้ เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกกลุ่มของข้อความนั้น ๆ จากภาษาธรรมชาติได้ ดังนั้นจึงต้องทำการแปลงข้อความให้อยู่ในรูปที่คอมพิวเตอร์สามารถเข้าใจได้ การแทนข้อความให้อยู่ในรูปเวกเตอร์สเปซโมเดล (Vector Space Model: VSM) เป็นวิธีการหนึ่งในการแทนข้อความให้มีโครงสร้างแบบฟีเจอร์เวกเตอร์ (Feature vector) สามารถเลือกคุณลักษณะของข้อความมาแทนได้หลายวิธี เช่น คำ วลี หรือหน้าที่ของคำ เป็นต้น ซึ่งคำที่นำมาแสดงในเวกเตอร์นั้นได้มาจากคำที่มีการกำหนดในถุงคำ โดยวิธีการในการสร้างเวกเตอร์นั้นจะขึ้นอยู่กับว่าลักษณะของคำที่อยู่ในถุงคำนั้นใช้วิธีการในการสร้างถุงคำใด โดยการสร้างถุงคำส่วนใหญ่จะนิยมใช้วิธีการค่าความถี่และน้ำหนักของคำหรือวิธีการหาความถี่ของคำที่ใช้ในเอกสาร

	w_1	w_2	w_3	w_4	w_5	...	w_n
d_1	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	...	w_{1n}
d_2	w_{21}	w_{22}	w_{23}	w_{24}	w_{25}	...	w_{2n}
d_3	w_{31}	w_{32}	w_{33}	w_{34}	w_{35}	...	w_{3n}

ภาพที่ 2.2 ตัวอย่างการสร้างตัวแทนเอกสาร

จากภาพที่ 2.2 เป็นวิธีการสร้างตัวแทนเอกสารคือ ในชุดข้อมูลมีคำเกิดขึ้นจะนำคำที่ได้มาเรียงกันจากนั้นจะทำการนับความถี่ของการเกิดขึ้นของคำนั้น ๆ ตัวอย่างเช่น เมื่อมีคำบ่งชี้ของคุณลักษณะในชุดข้อมูลทั้งหมด 10 คำ โดยเมื่อมีคำบ่งชี้ w_1 ในเอกสารจำนวน 3 คำ w_4 จำนวน 1 คำ w_5 จำนวน 2 คำ w_6 จำนวน 1 คำ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และคำบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 ดังนั้น การสร้างเวกเตอร์ตัวแทนเอกสารจากความถี่ของคำจะได้คือ

$$d_1 = \{3, 0, 0, 1, 2, 1, 0, 0, 0, 0\}$$

2.1.3.4 การตรวจนามวลี

การตรวจนามวลี เป็นขั้นตอนของการตรวจความหมายของคำที่มีต่อประโยคซึ่งจะเป็นลักษณะที่เป็นวลี โดยคำในภาษาไทยนั้นเมื่อเขียนในลักษณะของคำอาจให้ความหมายที่มี

ความแตกต่างออกไปจากลักษณะของคำที่อยู่ในรูปแบบของวลี โดยในขั้นตอนนี้อาจมีการนำไปใช้หรือไม่ก็ได้

2.1.3.5 การรวมความหรือการตีความหมาย

กระบวนการรวมความหรือการตีความหมายในส่วนนี้จะทำหน้าที่ในการรับเอาองค์ความรู้ที่ได้มีการกำหนดไว้เพื่อทำการตีความหมายตามลักษณะของโครงสร้างของคำหรือตามลักษณะของไวยากรณ์ที่ได้มีการกำหนดไว้แล้ว

2.1.3.6 การจำแนกประเภทข้อความเอกสาร

การจำแนกประเภทข้อความหรือเอกสารตามหมวดหมู่ที่ได้กำหนดไว้ โดยในส่วนนี้นิยมใช้เทคนิคในการทำเหมืองข้อมูลเข้ามาใช้ในการวิเคราะห์เพื่อทำการจำแนกกลุ่มหรือประเภทของข้อความ

2.1.4 Google Play

Google Play [35] เป็นบริการในด้านของแอพพลิเคชันสำหรับผู้ที่ใช้ระบบปฏิบัติการ Android โดยเปิดให้ผู้ใช้สามารถโหลดแอพพลิเคชัน มาติดตั้งเอาไว้ในสมาร์ทโฟนหรือแท็บเล็ตที่ลงทะเบียนเชื่อมต่อกับบัญชีอีเมล์ที่เป็นของ Gmail เดิมนั้นใช้ชื่อ เดิมชื่อ Android Market เป็นศูนย์รวมคอนเทนต์ต่าง ๆ ของกูเกิล เช่น เพลง แอพพลิเคชั่น หนังสือ เกมและอื่น ๆ ที่กูเกิลตั้งใจจะทำเป็นแหล่งดาวน์โหลดทุกอย่างที่กูเกิลมีให้บริการ โดยรวบรวมเอาบรรดา Application ต่าง ๆ เอาไว้มากมาย มีทั้งแบบที่เสียค่าใช้จ่ายและแบบฟรี

นอกจากจะช่วยให้ผู้ใช้โหลดมาติดตั้งผ่านทางอุปกรณ์โมบายต่าง ๆ ที่ใช้ระบบปฏิบัติการ Android แล้ว ยังสามารถดาวน์โหลดจาก Google Play ผ่านทาง Web browser ต่าง ๆ ได้อีกด้วย ไม่ว่าจะเป็น Chrome, Internet Explorer, Firefox หรือ Safari ที่ทำงานอยู่บนระบบปฏิบัติการอื่น ๆ ได้มากมาย ไม่ว่าจะเป็นการทำงานบนคอมพิวเตอร์ PC หรือ Mac ก็ตาม บริการ Google Play นี้ ออกแบบมาเพื่อให้ผู้ที่ใช้งานสมาร์ทโฟนที่เป็นระบบ Android สามารถที่จะโหลดแอพพลิเคชันต่าง ๆ มาปรับใช้ร่วมกับอุปกรณ์โมบายของตน ไม่ว่าจะเป็นเกม ซอฟต์แวร์สำหรับงานสำนักงาน การเชื่อมต่อ ความบันเทิงหรือแม้กระทั่งระบบโซเชียลเน็ตเวิร์กและการติดตามข่าวสาร ซึ่งล้วนแต่มีความจำเป็นในชีวิตประจำวันทั้งสิ้น

การดาวน์โหลดและติดตั้ง สิ่งที่สำคัญคือ ต้องมีบัญชีอีเมล์ของ Gmail สำหรับเป็นแอคเคาต์ของตัวเอง สำหรับการยืนยันตัวตนในการใช้งานและความปลอดภัย โดยที่แอคเคาต์หรือยูสเซอร์ที่ใช้อยู่นั้น จะติดตามไปยังอุปกรณ์อื่น ๆ ที่ผู้ใช้เปลี่ยนไป สำหรับการใช้งาน Google Play นั้นสามารถเข้าไปได้ที่เว็บไซต์ https://play.google.com/store ซึ่งจะมีการแบ่งหมวดหมู่ให้การเลือกใช้ง่ายขึ้น

2.1.5 เทคนิคการทำเหมืองข้อความ

การทำเหมืองข้อความเป็นอีกกระบวนการเพื่อสกัดเอาข้อมูลและความรู้ต่าง ๆ และเป็นอีกแนวทางในการที่จะแก้ปัญหาการเกินของข้อมูลการโดยการใช้เทคนิคการทำเหมืองข้อมูล โดยจากการเรียนรู้การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เทคนิคการดึงข้อมูล (Information retrieval : IR) ในการจัดการความรู้ในการทำเหมืองข้อความที่เกี่ยวข้องกับการประมวลผลเอกสาร โดยการพิจารณาจากการจัดหมวดหมู่ข้อความและข้อมูลจากการสกัดคำ การจัดเก็บข้อมูลของการแสดงผลของการประมวลผลในแต่ละเทคนิคที่ใช้ในการวิเคราะห์ เช่น การวิเคราะห์การกระจาย การจัดกลุ่ม การวิเคราะห์แนวโน้มและกฎความสัมพันธ์

เหมืองข้อความ เป็นกระบวนการค้นหาความรู้หรือข้อเท็จจริงที่แฝงอยู่ในชุดข้อความ การทำเหมืองข้อความมักจะนำไปใช้กับข้อมูลประเภทไม่มีโครงสร้าง (Unstructured Data) เช่น ข้อความอีเมล์ ข้อความบนเว็บบอร์ด หรือข้อมูลประเภทกึ่งโครงสร้าง (Semi-Structured Data) เช่น ข้อความรูปแบบ XML หรือ HTML เหมืองข้อความจึงเป็นเครื่องมือในการจัดการข้อมูลเหล่านี้ได้เป็นอย่างดี แต่การทำเหมืองข้อความมีความแตกต่างจากการทำเหมืองข้อมูลในด้านเครื่องมือโดยเครื่องมือการสร้างเหมืองข้อมูลจะออกแบบมาสำหรับข้อมูลที่มีโครงสร้าง ส่วนเครื่องมือการทำเหมืองข้อความจะออกแบบมาสำหรับข้อมูลที่ไม่มีโครงสร้าง กระบวนการทำเหมืองข้อความ มีกระบวนการคล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล ทั้งนี้หากผลจากการวิเคราะห์ในแต่ละขั้นตอนมีความถูกต้องหรือความน่าเชื่อถือต่ำเกินไป จะต้องกลับไปขั้นตอนที่ต่ำกว่า หรือทำการเลือกข้อมูลมาใหม่เพื่อทำให้กระบวนการทำเหมืองข้อความมีคุณภาพ โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ ดังภาพที่ 2.3 มีกระบวนการดังนี้

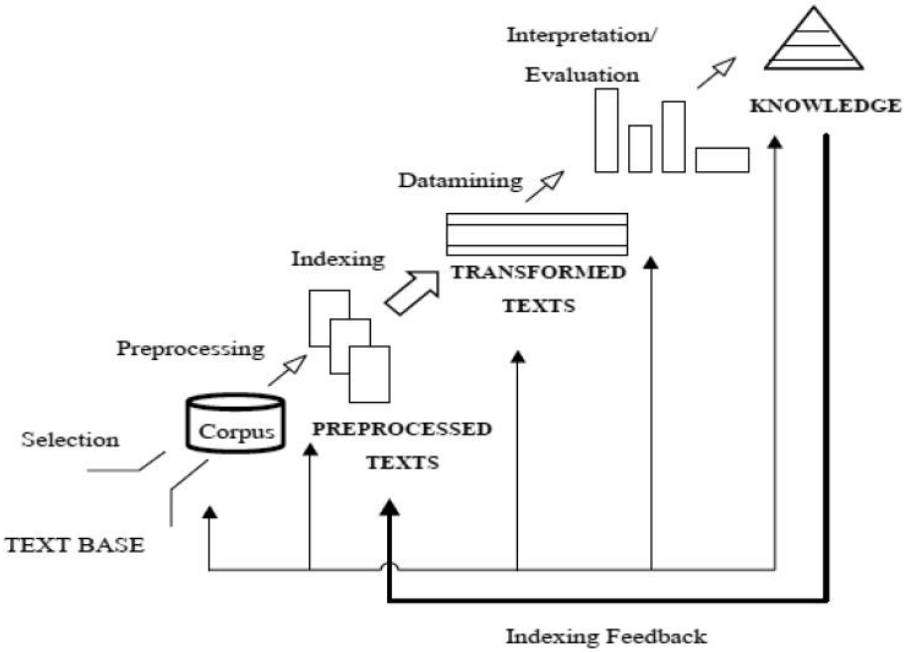
1. การเลือกข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อความ รวมถึงการนำข้อมูลที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการทำการศึกษา

2. การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนำข้อมูลที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขข้อมูลก่อนนำไปใช้งาน

3. การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป เช่น การตัดบางคอลัมน์ที่ไม่จำเป็นออก

4. การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนประมวลผลโดยใช้อัลกอริทึมต่าง ๆ เพื่อแก้ไขปัญหาหรือหารูปแบบของข้อมูลการจัดหมวดหมู่ (Classification) ด้วยเทคนิคต่าง ๆ เช่น เทคนิค Naïve Bayes เทคนิค SVM (Support Vector Machine) เทคนิคต้นไม้ตัดสินใจ (Decision tree) เทคนิค C4.5 เทคนิค K-Nearest Neighbor เป็นต้น

5. การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย



ภาพที่ 2.3 กระบวนการเหมืองข้อความ

อัจฉรา ภูระย้า [36] ได้ให้ความหมายของเหมืองข้อความว่า เป็นการค้นหารูปแบบที่ซ่อนอยู่ภายในข้อความที่มีอยู่จำนวนมาก โดยการใช้หลักการทางคณิตศาสตร์ การเรียนรู้ของเครื่อง ประมวลเอกสาร หลักการทางสถิติ การประมวลผลข้อความและการประมวลผลภาษาธรรมชาติ สมศักดิ์ วิชัยกิจ [37] ได้กล่าวว่า การทำเหมืองข้อความนั้นเป็นกระบวนการที่จัดการกับข้อความจำนวนมาก เพื่อค้นหารูปแบบ แนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความเอกสารนั้น ๆ โดยเป็นการใช้หลักการทางสถิติ การรู้จำ หลักการทางคณิตศาสตร์ การประมวลผลเอกสาร การประมวลผลข้อความและการประมวลผลภาษาธรรมชาติ เพื่อทำการนำเสนอผลลัพธ์ของการจัดการ

ข้อมูลให้เป็นองค์ความรู้ใหม่ซึ่งเป้าหมายของการทำเหมืองข้อความนั้นสามารถจำแนกออกได้เป็น 3 ประเภท คือ

1. Document Summarization คือ การสรุปเอกสารข้อความซึ่งเป็นวิธีการทำเหมืองข้อความที่ทำการลดรายละเอียดและความซับซ้อนของข้อมูลในเอกสารที่ไม่จำเป็นออกไป แต่จะยังคงมีเนื้อหาและสาระสำคัญของข้อมูลในเอกสารนั้น ๆ อย่างสมบูรณ์

2. Document Classification คือ การแบ่งประเภทเอกสารข้อความเป็นเทคนิควิธีการทำเหมืองข้อความโดยใช้การแยกประเภทของเอกสาร ซึ่งจำเป็นต้องมีการกำหนดจำนวนประเภทของเอกสารที่ต้องการแยกก่อน จากนั้นจึงทำการสอนเพื่อให้รู้จำรูปแบบเอกสารแต่ละประเภท

3. Document Clustering คือ การแบ่งกลุ่มเอกสารข้อความ วิธีการนี้จะทำการจัดแบ่งเอกสารข้อความออกเป็นกลุ่ม โดยการทำเหมืองข้อความด้วยวิธีนี้มีจุดประสงค์เพื่อที่ต้องการจัดกลุ่มของเอกสารที่มีจำนวนมากและข้อมูลเอกสารมีความคล้ายคลึงกัน

Nikhil R และคณะ [4] ได้กล่าวถึงลักษณะของข้อมูลในการทำเหมืองข้อความ โดยข้อมูลหรือข้อความนั้น ๆ เป็นลักษณะของข้อมูลที่ไม่มีโครงสร้างซึ่งไม่มีการกำหนดรูปแบบไว้ล่วงหน้า ซึ่งผลของข้อมูลที่ได้จึงไม่มีความแน่นอนของเหตุผลนั้น ๆ ที่แฝงอยู่ในกลุ่มของข้อความซึ่งมีลักษณะเป็นภาษาธรรมชาตินั่นเอง Ning Zhong และคณะ [38] ได้กล่าวว่าเทคนิคการทำเหมืองข้อมูลจำนวนมากนั้นจะเป็นลักษณะของการนำเสนอรูปแบบสำหรับการทำเหมืองข้อความที่ใช้ประโยชน์ในเอกสารและใช้ความรู้ที่ได้จากการสกัดข้อความเหล่านั้นมาเพื่อปรับปรุงรูปแบบและปัญหาต่าง ๆ ของสิ่งนั้น การค้นพบรูปแบบยังคงเป็นปัญหาเนื่องจากส่วนใหญ่แล้ววิธีการทำเหมืองข้อความมักจะประสบปัญหาของหลักภาษาและคำพ้องซึ่งมีลักษณะที่ซับซ้อนกันออกไป โดยได้มีการจัดสมมติฐานที่เกี่ยวกับกลุ่มคำหรือวลี ซึ่งวิธีดังกล่าวน่าจะได้ผลการทำงานที่ดีกว่า

เหมืองความคิดเห็น (Opinion Mining) เป็นการทำเหมืองข้อความอีกประเภทหนึ่งโดยเน้นการผสมผสานเทคนิคทางภาษา เพื่อพัฒนาเป็นระบบในส่วนเฉพาะที่เป็นความ คิดเห็น โดยจะพิจารณาถึงเซตของข้อความที่มีความคิดเห็นหรือทัศนคติที่เกี่ยวกับสิ่งใดสิ่งหนึ่งประกอบอยู่ด้วย ซึ่งลักษณะของความคิดเห็นนั้นจะเป็นถ้อยคำที่แสดงถึงเหตุผลและเป็นความคิดส่วนตัวที่สะท้อนให้เห็นทัศนคติหรือการรับรู้เกี่ยวกับสิ่งที่มีตัวตนและเหตุการณ์ของคน โดยการทำเหมืองข้อมูลความคิดเห็นมีวัตถุประสงค์เพื่อดึงคุณลักษณะและส่วนประกอบของสิ่งนั้น ๆ ที่ได้ถูกแสดงความคิดเห็นในรูปแบบของข้อความตัวอักษรหรือสัญลักษณ์แทนความรู้สึก เพื่อจะพิจารณาว่าความคิดเห็นนั้นเป็นไปในเชิงบวก เชิงลบ หรือเป็นกลาง การจัดการข้อความในลักษณะของข้อความคิดเห็นนั้นมีในแอพพลิเคชั่นที่

สำคัญหลายประเภทด้วยกัน ตัวอย่างเช่น การตัดสินความคิดเห็นของผู้วิจารณ์เกี่ยวกับสินค้าโดยแบ่งประเภทบทวิจารณ์สินค้าออนไลน์ หรือการติดตามการเปลี่ยนทัศนคติของประชาชนทั่วไปเกี่ยวกับการเมือง การทำเหมืองข้อมูลความคิดเห็นนั้นสามารถแบ่งออกเป็นงานย่อยหลายงาน ซึ่งประกอบด้วยคุณลักษณะของข้อความดังต่อไปนี้

1. การตัดสินขั้วความเป็นวัตถุหรือหัวข้อที่นำมาแสดงความคิดเห็นของข้อความเป็นการตัดสินว่าข้อความมีลักษณะเป็นความจริง โดยบรรยายถึงสถานการณ์หรือเหตุการณ์ โดยปราศจากการแสดงออกถึงความคิดเห็นเชิงบวกหรือเชิงลบในข้อความ หรือแสดงออกถึงความคิดเห็นเกี่ยวกับเรื่องที่สนใจ เพื่อจะสร้างการแบ่งประเภทข้อความแบบไบนารี่ภายใต้ประเภทของวัตถุหรือหัวข้อที่นำมาแสดงความคิดเห็นนั้น

2. การตัดสินขั้วความเป็นบวกหรือลบของข้อความ เป็นการตัดสินว่าข้อความของหัวข้อที่นามาแสดงความคิดเห็นที่ได้ แสดงออกถึงความคิดเห็นเชิงบวกหรือเชิงลบเกี่ยวกับเรื่องที่สนใจ

3. การตัดสินความแข็งแรงของขั้วบวกขั้วลบของข้อความ ตัวอย่างเช่น ความคิดเห็นเชิงบวกที่ข้อความแสดงออกมาเกี่ยวกับเรื่องที่สนใจนั้น เป็นเชิงบวกอย่างอ่อน เชิงบวก กลางๆ หรือเชิงบวกอย่างมาก

ได้มีการนำเทคนิคการทำเหมืองข้อความมาประยุกต์ในการทำเหมืองแสดงความคิดเห็นซึ่ง กานดา แผ่วัฒนากุล [5] ได้กล่าวถึงการวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายออนไลน์ว่าเป็นกระบวนการวิเคราะห์ข้อคิดเห็นที่ผู้บริโภคนำเสนอในรูปแบบของข้อความซึ่งพบได้บนสื่อออนไลน์เพื่อให้ทราบถึงความพึงพอใจที่มีต่อสิ่งนั้น ๆ และในการสกัดคำตามคุณลักษณะที่แตกต่างกันอาจได้มาซึ่งประโยชน์เพื่อการพิจารณาวิเคราะห์ที่หลากหลายและแม่นยำมากขึ้น พัชรนิกานต์ พงษ์ธนู [7] ได้นำเสนอโมเดลการวิเคราะห์เหมืองข้อความจากการเก็บข้อมูลการแสดงความพึงพอใจของลูกค้าจากข้อความคำแนะนำบนเว็บไซต์ โดยใช้การสกัดคำ ในความคิดเห็นด้านดีและด้านไม่ดีเพื่อสรุปการให้บริการของเว็บไซต์ บริการ ซึ่งได้เปรียบเทียบผลจากการสร้างโมเดลด้วยเทคนิควิธีต้นไม้ตัดสนใจและวิธีการแบบเบย์อย่างง่าย โดยมีวัตถุประสงค์เพื่อหาแนวทางในการปรับปรุงการบริการของเว็บไซต์ผู้ให้บริการโรงแรมให้มีประสิทธิภาพมากขึ้น

ในการศึกษาการทำเหมืองข้อความในครั้งนี้ ผู้วิจัยได้มุ่งเน้นการจำแนกประเภทเอกสารข้อความที่มีลักษณะเป็นข้อความแสดงความคิดเห็น ซึ่งจำเป็นต้องมีการกำหนดจำนวน

ประเภทของข้อความคิดเห็นที่ต้องการแยกก่อน จากนั้นจึงทำการสอนเพื่อให้รู้จำรูปแบบข้อความแต่ละประเภท

2.1.6 Synthetic Minority Over-sampling Technique

Synthetic Minority Over-sampling Technique (SMOTE) [9] เป็นเทคนิคในการสุ่มตัวอย่างของคลาสที่น้อย เพื่อแก้ปัญหาชุดข้อมูลที่มีคลาสไม่สมดุล ซึ่งข้อมูลมีจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาส เมื่อทำการจำแนกประเภท จะทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มที่มาก ผลที่ได้ก็จะจำแนกไปในข้อมูลที่มีกลุ่มมาก วิธี SMOTE เป็นวิธีการเพิ่มจำนวนข้อมูลประเภทที่มีข้อมูลน้อย ให้เพิ่มปริมาณข้อมูลใกล้เคียงกับประเภทที่มีมากที่สุด โดยสุ่มค่าขึ้นมาหนึ่งค่า และหาระยะห่างระหว่างค่าที่เลือกกับทุก ๆค่า แล้วเลือกค่าที่ใกล้เคียงที่สุด เช่น กำหนดไว้ 5 ค่า สุ่มค่าจากที่เลือก 1 ใน 5 หาค่าอยู่ระหว่างค่าที่เลือกตอนแรกและค่าที่สุ่มมาตอนหลัง เพื่อนำค่าที่ได้มาเพิ่มจำนวนข้อมูล ดังสมการที่ 2-3

$$X_{new} = X_i + (\hat{X_i} - X_i) \times \delta \qquad (2\text{-}3)$$

X_{new} คือ ข้อมูลใหม่

X_i คือ ข้อมูลที่สุ่มในตอนแรก

$\hat{X_i}$ คือ ข้อมูลที่สุ่มมาอีก เช่น สุ่มมาอีก 5 จุด

δ คือ ค่าสุ่มตั้งแต่ 0-1

ได้มีการนำ SMOTE มาใช้เพื่อแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยซึ่ง เชาวนันท์ โสโท [10] ได้สร้างแบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม และใช้ SMOTE มาทำการปรับสมดุลของข้อมูลแล้วทำการสร้างแบบจำลอง โดยผลการทดลองพบว่าการปรับความสมดุลของข้อมูลด้วยวิธีการ SMOTE มีประสิทธิภาพในการทำนายที่ดีกว่าวิธีอื่น นพมาศ อัครจันทโชติและดิเรก พนิตสุภากมล [39] ได้เปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยา ด้วยเทคนิคการถดถอยโลจิสติกและต้นไม้ตัดสินใจ โดยใช้วิธีการสุ่มตัวอย่างเพิ่มข้อมูลเริ่มต้นอย่างสุ่ม (Random over-sampling) และ SMOTE ผลการทดลองพบว่าการสุ่มตัวอย่างของคลาสที่น้อยหรือ SMOTE มีประสิทธิภาพสูงที่สุดในการจำแนก

2.1.7 เทคนิคที่ใช้ในงานวิจัย

ในงานวิจัยนี้เทคนิคที่นำมาใช้ในการสร้างแบบจำลองได้แก่ เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิคต้นไม้ตัดสินใจ C4.5

2.1.7.1 Naïve Bayes

นาอีฟเบย์ มีพื้นฐานมาจากกฎของเบย์ เป็นทฤษฎีทางด้านสถิติโดยน่าความน่าจะเป็นมาใช้ ประเมินความไม่แน่นอนให้เป็นตัวเลขได้ กล่าวถึง ความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น (A) ถ้ามี เหตุการณ์อีกเหตุการณ์หนึ่งเกิดมาแล้ว (B) สามารถ เขียนให้อยู่ในรูปอย่างง่าย ดังสมการที่ 2-4

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \quad (2\text{-}4)$$

P(A|B) คือ ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้น ถ้าเหตุการณ์ B เกิดขึ้นแล้ว

P(B|A) คือ ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น ถ้าเหตุการณ์ A เกิดขึ้นแล้ว

P(A) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A

P(B) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B

วิธีการจำแนกหมวดหมู่โดยใช้หลักการความน่าจะเป็น เป็นการแก้ปัญหาแบบ Classification สามารถคาดการณ์ผลลัพธ์และสามารถอธิบายได้ ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เป็นวิธีการจำแนกประเภทข้อมูลที่มี ประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (Text Classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นดัง สมการที่ 2-5 [40]

$$P(A_1, A_2, \dots, A_n|C_i) = \prod_{i=1}^{n} P\,(A_i|C_i) \quad (2\text{-}5)$$

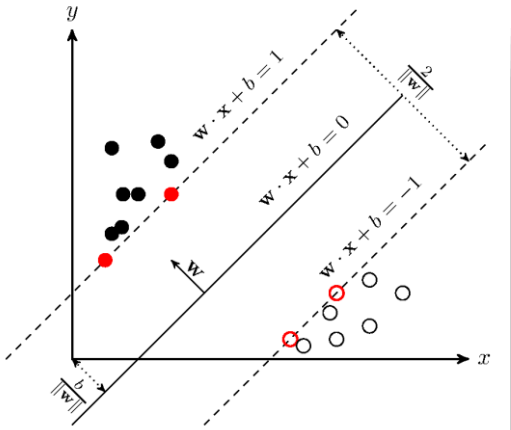
กลุ่ม C_i สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = (A_1, A_2, \dots, A_n)$ หรือใช้สัญลักษณ์ว่า $P(A_1, A_2, \dots, A_n|C_i)$ โดยที่ $\prod$ หมายถึงผลคูณของค่า $P(A_i|C_i)$ ทั้งหมด i = 1, 2, 3, ..., n และ j = 1, 2, 3,..., n ดังนั้นจะได้วิธีการจำแนกประเภทแบบเบย์ อย่างง่ายดังสมการที่ 2-6

$$V_{NB} = argmax P(C_i) \prod_{i=1}^{n} P\,(A_i|C_i) \quad (2\text{-}6)$$

จุฑาทิพย์ ทิพย์พูล [41] ได้นำเสนอเทคนิคการทำเหมืองข้อความ 3 เทคนิค ซึ่งประกอบด้วย เทคนิคต้นไม้ตัดสินใจ เทคนิค Naïve Bayes และเทคนิค k-NN โดยทำการเก็บรวบรวมข้อมูลจดหมายทั้งหมด 5,172 ข้อความ ซึ่งผลการทดลองพบว่าเทคนิควิธีของ Naïve Bayes ให้ค่าความถูกต้องมากที่สุด

2.1.7.2 Support Vector Machines (SVM)

ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) นำเสนอทฤษฎีโดย Cortes and Vapnik (1995) เพื่อลดความผิดพลาดจากการทำนาย (Minimize error) จัดเป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูลอาศัยหลักการจำแนกหมวดหมู่ข้อมูลด้วยการหาระนาบตัดสินใจและแบ่งข้อมูลออกเป็น 2 ส่วน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะระหว่างขอบเขตทั้ง 2 กลุ่มมากที่สุด (Optimal separating hyperplane) [42] เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูล โดยใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space ในลักษณะเชิงเส้น ดังแสดงในภาพที่ 2.4



ภาพที่ 2.4 การวางตัวของข้อมูลในลักษณะเชิงเส้น

และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) Kernal Function ในสื่อตีพิมพ์เกี่ยวกับ SVM [43] จะเรียกตัวแปรในการตัดสินใจว่าคุณสมบัติ และตัวแปรที่เปลี่ยนแปลงใช้ในการกำหนดระนาบหลายมิติเรียกว่า คุณลักษณะ (Feature) ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า การคัดเลือกคุณลักษณะ (Feature Selection) จำนวนเซตของคุณลักษณะที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่า เวกเตอร์ (Vector) ดังนั้น จุดมุ่งหมายของตัวแบบ SVM คือการได้ประโยชน์สูงสุดจากระนาบหลายมิติที่แบ่งแยกกลุ่มของเวกเตอร์ ในกรณีนี้ด้วย Feature Space เหมาะใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง กำหนดให้ $(x_i, y_i), \ldots, (x_n, y_n)$ เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือ จำนวนมิติข้อมูลเข้า และ y คือผลลัพธ์ มีค่า +1 หรือ -1 ดังสมการที่ 2-7

$$(xR\{+1,-1\}(5)_i, y_i), \ldots, (x_n, y_n) \quad \text{เมื่อ } x\,m\,,\,y \in\in \qquad (2\text{-}7)$$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูก แบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจ ซึ่งคำนวณได้ดัง สมการที่ 2-8

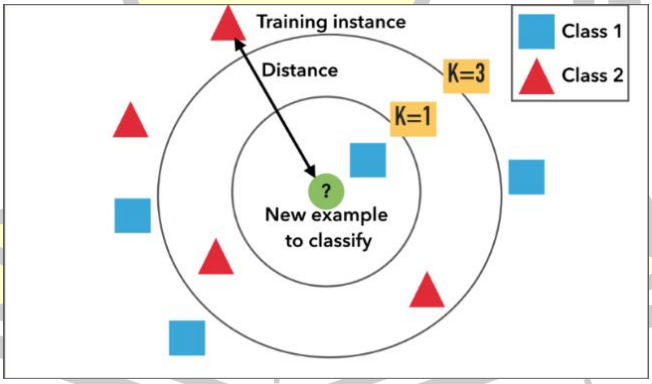
$$(w \times x) + b = 0 \quad (2\text{-}8)$$

เมื่อ w คือ ค่าน้ำหนักและ b คือค่า bias สมการ ใช้สำหรับจำแนกประเภทของข้อมูล ดังสมการที่ 2-9

$$(w \times x) + b > 0 \text{ ถ้า } y_i + 1 \text{ และ } (w \times x) + b < 0 \text{ ถ้า } y_i - 1 \quad (2\text{-}9)$$

2.1.6.3 K-Nearest Neighbor

เคเนียเรสเนเบอร์ (K-Nearest Neighbor) เป็นวิธีการในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง หลักการของวิธีการนี้ [44] จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงกันมากที่สุด k ตัวจากข้อมูลบนชุดข้อมูลตัวอย่าง ทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่หรือข้อมูลที่ป้อนถาม (Input Query Instance) กับข้อมูลตัวอย่างฝึกฝนจะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว หลังจากนั้นจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด k ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม k ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ดังแสดงในภาพที่ 2.5



ภาพที่ 2.5 การจัดกลุ่มของเทคนิค k-NN

ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง k ตัว ประกอบด้วยแอตทริบิวท์หลายตัวแปร $\mathbf{x_i}$ ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ k ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายหากรณีใหม่ อัลกอริทึมแบบ KNN ได้แก่ 1-NN , 2-NN , 3-NN , ... k-NN โดยค่า k ต้องระบุในการสร้างโมเดล [44] มาตรวัด

ความถูกต้อง (Distance Measure) การหาความยาวระหว่างจุดที่ต้องการโดยใช้เครื่องมือ และวิธีต่าง ๆ งานวิจัยได้เลือกวิธีการหามาตรวัดความแม่นยำ โดย Euclidean Distance ระยะทาง ระหว่าง 2 จุด จุดที่จะวัดนั้นมีเงื่อนไขมีหลายค่าจากหลายมิติหรือขนาดขึ้นกับรูปแบบ ซึ่งสามารถพิสูจน์หาค่าได้ด้วยทฤษฎีของ Pythagorean เมื่อมีการใช้สูตรเพื่อหาระยะทางขนาดของ Euclidean ระยะทางระหว่างจุด $P = (p_1, p_2, \ldots, p_3)$ และ $Q = (q_q, q_2, \ldots, q_n)$ ใน Euclidean หลายขนาดระบุได้เป็น ดังสมการที่ 2-10

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2\text{-}10)$$

สมมาตร อังคเศรณีกุล [45] ได้นำระบบตรวจข้อสอบแบบอัตนัยแบบอัตโนมัติโดยเลือกใช้เทคนิค k-NN มาทำการจำแนกคำตอบโดยการเปรียบเทียบผลของการตรวจเทียบกับการตรวจจากผู้สอน ซึ่งจากผลที่ได้นั้น ค่าความแม่นยำรวมอยู่ที่ระหว่าง 76.73% - 92.73% บรรหาร จันทะวงค์ และคณะ [46] ได้นำเสนอการจัดประเภทเอกสารงานวิจัยทางโลจิสติกส์ เพื่อสร้างระบบช่วยในการทำนายประเภทของเอกสารงานวิจัย โดยใช้เทคนิคการทำเหมืองข้อมูลด้านการจัดประเภทข้อมูล ซึ่งได้เลือกวิธี k-Nearest Neighbors โดยจากการทดลองได้ค่าความแม่นยำสูงสุดในการทำนายที่ 99.00% โดยใช้วิธีการมาตรวัดระยะห่างแบบ Overlap Similarity

2.1.7.4 ต้นไม้ตัดสินใจ (Decision tree) ID3

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) ID3 [47] คือ แบบจำลองทางคณิตศาสตร์เพื่อการหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างแบบจำลองการจัดหมวดหมู่(Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล่วงหน้า (Training set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น

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

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

ส่วนประกอบต้นไม้ตัดสินใจ ประกอบด้วย

1) โหนด (Node) คือ คุณสมบัติต่างๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node)

2) กิ่ง (Branch, Link) คือ ค่าคุณสมบัติของคุณสมบัติในโหนดที่แตกกิ่งนี้ออกมา ซึ่งโหนดจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดนั้น

3) โหนดใบ (Leaf Node) คือ กลุ่ม (Class) ต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกข้อมูล ซึ่งเป็นขั้นตอนวิธีหนึ่งที่ใช้ในการสร้างต้นไม้ตัดสินใจจะใช้ค่าที่วัดที่เรียกว่าค่าเกน (Gain) เป็นตัวตัดสินใจว่าจะใช้แอตทริบิวท์ใดในการแบ่งข้อมูลเพื่อใช้ในการตัดสินใจ โดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นตามลำดับของค่าตัวชี้วัดหรือค่าเกนของแอตทริบิวท์ที่มีค่าสูงที่สุด ซึ่งเริ่มหาค่าสารสนเทศ (Information) แต่ละแอตทริบิวท์ซึ่งบ่งบอกความสามารถของแอตทริบิวท์ในการแยกแต่ละคลาส (Class) ตามสมการที่ 2-11 ดังนี้

$$I(S_1, S_2, \ldots, S_n) = -\sum_{i=1}^{n} \frac{Si}{S} \log_2 \frac{Si}{S} \qquad (2\text{-}11)$$

โดย S คือ จำนวนข้อมูลทั้งหมด ซึ่งมีจำนวนข้อมูล $\mathbf{S_i}$ ของ $\mathbf{CLASS_i}$

n คือ จำนวน Class ทั้งหมด

จากนั้นหาค่าเอนโทปี (Entropy) เป็นค่าของผลรวมแต่ละแอตทริบิวท์ A ที่สามารถแยกประเภทข้อมูลสารสนเทศ (Information) แต่ละคลาสของแต่ละค่าของแอตทริบิวท์ m จำนวน ดังสมการที่ 2-12

$$E(A) = \sum_{j=1}^{m} \frac{S_{1j} + \ldots + S_{nj}}{S} I(S_{1f}, \ldots, S_{2j}) \qquad (2\text{-}12)$$

ดังนั้นค่าเกน (Gain) ของแอตทริบิวท์ที่ถูกเลือกหาได้จากการแยกประเภทข้อมูลสารสนเทศทั้งหมดของแอตทริบิวท์นั้นดังสมการที่ 2-13

$$Gain(A) = I(S_1, S_2, \ldots, S_n) - E(A) \qquad (2\text{-}13)$$

2.1.6.5 ต้นไม้ตัดสินใจ (Decision tree) C4.5

C4.5 [48] เป็นอัลกอริทึมการสร้างกฎจากต้นไม้ตัดสินใจ (Decision Tree) มีพื้นฐานเดียวกับอัลกอริทึม ID3 ถูกออกแบบโดย Quinlan (1992) ซึ่งได้พัฒนาเพิ่มเติมจาก ID3 ดังนี้

1) สามารถหลีกเลี่ยงการสร้างโครงสร้างต้นไม้ที่ใหญ่เกินไป เนื่องจากมีข้อมูลจำนวนมาก อย่างไรก็ตามขึ้นอยู่กับการกำหนดความลึก เมื่อมีการเจริญเติบโตของ Decision Tree

2) ความผิดพลาดลดลง เพราะมีการตัดทอนความผิดพลาดออกไป (Pruning node)

3) มีการสร้างกฎหลังการตัดทอนข้อมูลที่ผิดพลาดออก

4) สามารถใช้กับข้อมูลที่มีความต่อเนื่อง (Continuous Attributes) ที่เป็นตัวเลขได้ เช่น ค่าอุณหภูมิ จานวนเงิน เป็นต้น

5) การเลือก Attribute ที่วัดการเลือกให้เหมาะสม

6) สามารถใช้กับชุดข้อมูล (Training Data) ที่มีค่าผิดพลาด (Missing Attribute)

7) สามารถใช้สำหรับ Attribute กับ Costs ที่แตกต่างกันได้

2.1.8 การวัดประสิทธิภาพของแบบจำลอง

2.1.8.1 K-fold cross validation

K-fold cross validation คือ การเลือกสุ่มข้อมูลแบบความเที่ยงตรง ซึ่งวิธีนี้เป็นที่นิยมในการทำงานวิจัยเพื่อใช้ในการทดสอบประสิทธิภาพของโมเดล เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี K-fold cross-validation จะทำการเลือกสุ่มข้อมูลออกเป็น K ชุดเท่าๆ กัน จากนั้นจะทำการทดลองครั้งแรกด้วยข้อมูลชุดที่ 1 ซึ่งเป็นข้อมูลทดสอบและกำหนดให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั้งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ โดยวิธีการทดสอบประสิทธิภาพแบบ K-fold cross validation มีข้อเสียคือ จะต้องทำการเริ่มทดสอบใหม่โดยจะต้องทำทั้งหมด K รอบ

	set1	set2	set3	set4	set5	set6	set7	set8	set9	set10
Data training	data10	data1	data2	data3	data4	data5	data6	data7	data8	data9
	data9	data10	data1	data2	data3	data4	data5	data6	data7	data8
	data8	data9	data10	data1	data2	data3	data4	data5	data6	data7
	data7	data8	data9	data10	data1	data2	data3	data4	data5	data6
	data6	data7	data8	data9	data10	data1	data2	data3	data4	data5
	data5	data6	data7	data8	data9	data10	data1	data2	data3	data4
	data4	data5	data6	data7	data8	data9	data10	data1	data2	data3
	data3	data4	data5	data6	data7	data8	data9	data10	data1	data2
	data2	data3	data4	data5	data6	data7	data8	data9	data10	data1
Data test	**data1**	**data2**	**data3**	**data4**	**data5**	**data6**	**data7**	**data8**	**data9**	**data10**

ภาพที่ 2.6 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

จากภาพที่ 2.6 จะเป็นการทดสอบประสิทธิภาพแบบ 10-fold ซึ่งจะทำการแบ่งชุดข้อมูลออกเป็น 10 ชุด โดยในแต่ละรอบจะใช้ชุดข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ 1 ชุด และให้

ชุดข้อมูลอื่น ๆ เป็นข้อมูลชุดสอน โดยจะทำการทดสอบทั้งหมด 10 รอบ จากนั้นนำค่าความถูกต้อง (Accuracy) ที่ได้จากการทดสอบประสิทธิภาพในแต่ละรอบมาหาค่าเฉลี่ยเพื่อเป็นค่าความถูกต้องโดยรวมของการจำแนกข้อมูลของแบบจำลอง ซึ่งสามารถคำนวณค่าความถูกต้องได้ดังสมการที่ 2-14

$$overall\ accuracy = Average\ (\sum_{i=1}^{K} Accuracy_i) \quad (2\text{-}14)$$

เมื่อ K คือ จำนวนชุดข้อมูลที่ทดสอบทั้งหมด

2.1.8.2 การวิเคราะห์ประสิทธิภาพ

การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) โดย

	Class	Prediction Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

ภาพที่ 2.7 Confusion Matrix

ผลของการจำแนกกลุ่มจะสามารถใช้ค่าจากภาพที่ 2.7 ใช้ในการคำนวณหาคือค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) ได้ดังสมการที่ 2-15 ถึง 2-17

ซึ่งสามารถหาได้จากสมการดังต่อไปนี้ [42, 49]

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (2\text{-}15)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2\text{-}16)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2\text{-}17)$$

โดยที่ TP คือ จำนวนข้อมูลที่จำแนกถูกต้องจาก Class Positive

TN คือ จำนวนข้อมูลที่จำแนกถูกต้องจาก Class Negative

FP คือ จำนวนข้อมูลที่จำแนกไม่ถูกต้องใน Class Positive

FN คือ จำนวนข้อมูลที่จำแนกไม่ถูกต้องใน Class Negative

ค่าความถูกต้อง คือ การวัดผลจากผลลัพธ์ของการเรียนรู้ในการทำนายกลุ่มตัวอย่างชุดใหม่ได้อย่างถูกต้อง

ค่าความไว หรือ True-Positive Rate คือ ค่าความน่าจะเป็นหรืออัตราส่วนการจำแนกถูกต้องในกลุ่มที่สนใจ

ค่าความจำเพาะ หรือ True-Negative Rate คือ ค่าความน่าจะเป็นหรืออัตราส่วนการจำแนกถูกต้องในกลุ่มอื่น ๆ

2.2 งานวิจัยที่เกี่ยวข้อง

ประพัฒน์ พรมน้ำอ่าง [6] ได้นำเสนอการจำแนกกลุ่มข้อความรีวิวโดยการใช้เทคนิคเหมืองข้อมูล ซึ่งประกอบด้วยเทคนิค SVM เทคนิคต้นไม้ตัดสินใจ เทคนิค k-NN และ Naïve Bayes โดยทำการเก็บข้อมูลจากข้อความรีวิวทั้งหมด 12,500 ข้อความ และแบ่งออกมาเป็นคำต่าง ๆ ตามคุณลักษณะเชิงบวกและเชิงลบจำนวนทั้งหมด 1,433 คำ ผลการทดลองพบว่าเทคนิค SVM ได้ให้ค่าความถูกต้องมากที่สุดที่ 86.26%

สุวิมล วงศ์สิงทอง [50] ได้ใช้วิธีจัดกลุ่มของการแชร์ข้อมูลบนเฟซบุ๊คของนักศึกษาทั้งหมด 700 คน ซึ่งได้ใช้วิธีเปรียบเทียบการจัดกลุ่มโดยการใช้ K-NN การจำแนกแบบต้นไม้ การเรียนรู้แบบเบย์และ SVM ซึ่งผลการวิจัยพบว่าวิธีการของ SVM ให้ค่าความถูกต้องมากที่สุดที่ 87.95%

อดิเทพ ไชยสาร [51] ได้นำเสนอการเปรียบเทียบการประมาณอารมณ์จากความคิดเห็นภาษาไทย โดยใช้ 3 เทคนิค คือ เทคนิค Naïve Bayes เทคนิค SVM และเทคนิคต้นไม้ตัดสินใจ ซึ่งได้รวบรวมความคิดเห็นจากเว็บไซต์บริการข่าว วาไรตี้บันเทิงและบริการวิจารณ์สินค้า จำนวน 6,000 ความคิดเห็น โดยได้แบ่งกลุ่มความคิดเห็นออกเป็น 6 กลุ่มอารมณ์ โดยวิธี SVM นั้นสามารถประมาณอารมณ์ได้ถูกต้องที่สุด 69.15% เมื่อเทียบกับเทคนิค Naïve Bayes และเทคนิคต้นไม้ตัดสินใจ

วลัยลักษณ์ สุขสมบุญ [52] ได้ทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทปัญหาสำหรับระบบถามตอบ โดยใช้เทคนิค SVM เทคนิค Naïve Bayes และ เทคนิค k-NN ซึ่งจากการทดลองเปรียบเทียบวิธีการจำแนกประเภทของปัญหาสำหรับ Helpdesk แสดงให้เห็นว่าเทคนิคที่นำเสนอด้วย SVM นั้นให้ค่าความถูกต้องที่ดีสุดโดยเมื่อใช้ฟังก์ชันแกน Polynomial ให้ค่าความ

ถูกต้องคือ 86.19% ค่าความแม่นยำ 86.2% ค่าระลึก 86.2% และ F-Measure 86% ได้ค่าดีที่สุด เมื่อทำการปรับค่า Degree = 3 Gamma = 0.2 0.4 และ 0.8 และ C = 1

K. Gowtham Reddy และ Jagadeesh Gopal [53] ได้นำเสนอการวิเคราะห์ความคิดเห็นเกี่ยวกับบทวิจารณ์เกมที่อยู่บน Twitter ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยนำความคิดเห็นมาจำแนกว่าความคิดเห็นแต่ละความคิดเห็นเป็นความคิดเห็นเชิงบวกหรือเชิงลบ โดยเก็บข้อมูลบน Twitter จำนวน 1,200 ข้อความ มาจำแนกด้วยเทคนิค Support vector machine, Naïve Bayes และ Maximum Entropy ซึ่งจากผลการวิจัยพบว่าเทคนิค Maximum Entropy ให้ค่าความถูกต้องมากที่สุดที่ 90%

Ha-Na Kang, Hye-Ryeon Yong, และ Hyun-Seok Hwang [54] ได้ศึกษาการวิเคราะห์ข้อมูลรีวิวเกมออนไลน์บนชุมชม STEAM โดยใช้เทคนิคเหมืองมูลมาจำแนกว่ารีวิวใดมีประโยชน์บ้าง ซึ่งใช้ข้อมูลรีวิวเกมจำนวน 79,437 ข้อความ จากทั้งหมด 11 เกม โดยใช้เทคนิคต้นไม้ตัดสินใจ CART เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks) และใช้หลักการการวัดประสิทธิภาพแบบ 10-fold cross validation จากผลการทดลองพบว่าเทคนิคต้นไม้ตัดสินใจ CART มีประสิทธิภาพดีกว่า

Rohan Bais, Pasal Odek และ Seyla Ou [55] ได้ศึกษาและทดสอบการจำแนกความคิดเห็นของรีวิวเกมบน STEAM ซึ่งใช้ข้อมูลจำนวน 5,000 ข้อความ จากเกมทั้งหมด 100 เกม โดยแบ่งความคิดเห็นเป็น 2 กลุ่มคือความคิดเห็นเชิงบวกและเชิงลบ โดยใช้เทคนิค Naïve Bayes เทคนิค SVM และเทคนิค Logistic Regression และใช้หลักการการวัดประสิทธิภาพแบบ 10-fold cross validation โดยแบ่งเป็นชุดข้อมูลฝึก 70% ชุดข้อมูลทดสอบ 30% จากผลการทดลองพบว่าเทคนิค SVM ให้ความถูกต้องมากที่สุดที่ 93.5%

Peiman Barnaghi และ John G. Breslin [8] ได้ทำการศึกษาถึงการบวนการสร้างโมเดลในการวิเคราะห์ความคิดเห็นภาพสะท้อนในความเชื่อมั่นของประชาชนต่อเหตุการณ์ต่าง ๆ ผ่านการใช้งานบน Twitter ซึ่งสามารถนำมาใช้เป็นพื้นฐานสำหรับการคาดคะเนเหตุการณ์ในอนาคต ซึ่งงานวิจัยนี้ได้ทำการศึกษาถึงข้อคิดเห็นเกี่ยวกับการจัดงาน FIFA World Cup 2014 โดยทำการพิจารณาข้อคิดเห็นที่เป็นเชิงบวกและเชิงลบจากการใช้เทคนิค TF-IDF เพื่อทำการคัดกรองคุณลักษณะดังกล่าว จากนั้นได้ทำการเปรียบเทียบการสร้างโมเดลจากการใช้เทคนิค Bayesian Logistic Regression (BLR) และเทคนิค Naïve Bayes โดยผลการทดสอบพบว่าเทคนิค BLR ให้ผลที่ดีกว่า

Mondher Bouazizi และ Tomoaki Ohtsuki [56] ได้ศึกษาถึงกระบวนการวิเคราะห์ข้อความคิดเห็นร่วมกับการวิเคราะห์ความคิดเห็นของภาษาบนสื่อออนไลน์ โดยได้ทำการรวบรวมข้อความแสดงความคิดเห็นต่างโดยไม่สนในหัวข้อหรือกลุ่มบทความบน Twitter เพื่อทำการศึกษาถึงการจำแนกหัวข้อและทัศนคติที่ผู้ใช้งาน Twitter สนใจ ในการวิจัยดังกล่าวได้ทำการเก็บรวบรวมข้อมูลแสดงความคิดเห็นจำนวนทั้งสิ้น 20,000 ข้อความ จากนั้นใช้เทคนิคการจำแนกกลุ่ม 2 เทคนิคด้วยกัน คือ เทคนิค Naïve Bayes และเทคนิค SVM โดยได้ทำการเปรียบเทียบผลการทดลองของทั้งสองเทคนิควิธีด้วยข้อมูล 2 ลักษณะคือ ข้อมูลแสดงความคิดเห็นและข้อมูลแสดงความคิดที่ผ่านกระบวนการวิเคราะห์ความเชื่อมั่น ผลที่ได้นั้น เทคนิค SVM ให้ผลของการจำแนกกลุ่มดีที่สุด

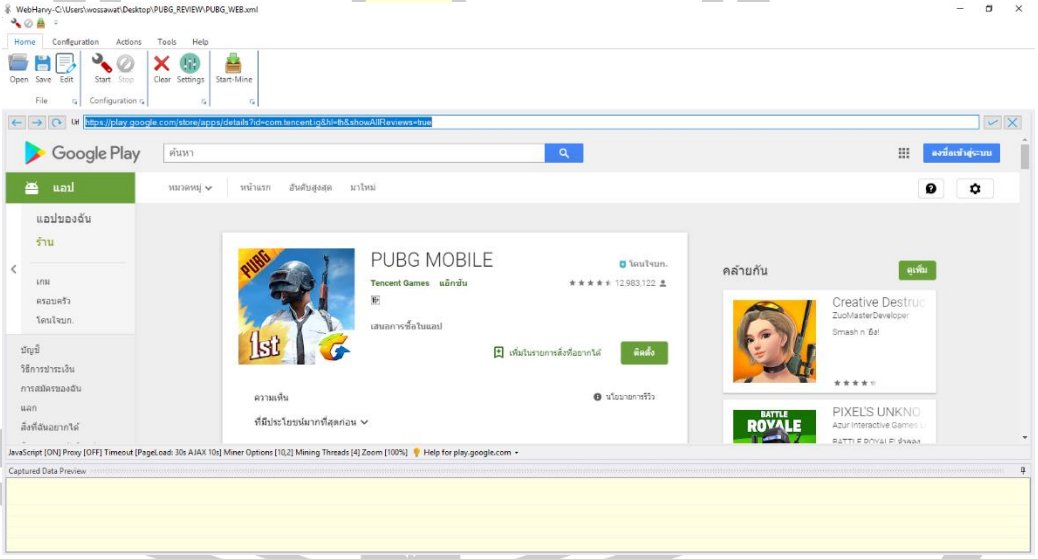
พหุน ปณ ฑิโต ชีเว

บทที่ 3
วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยเพื่อให้ได้ตามวัตถุประสงค์ที่ตั้งไว้ กระบวนการในการทำเหมืองข้อความซึ่งมี 4 ขั้นตอนได้ถูกนำมาใช้โดยเริ่มจากการเก็บรวบรวมข้อมูล ขั้นตอนก่อนการสร้างแบบจำลอง การสร้างแบบจำลองและการวัดประสิทธิภาพ

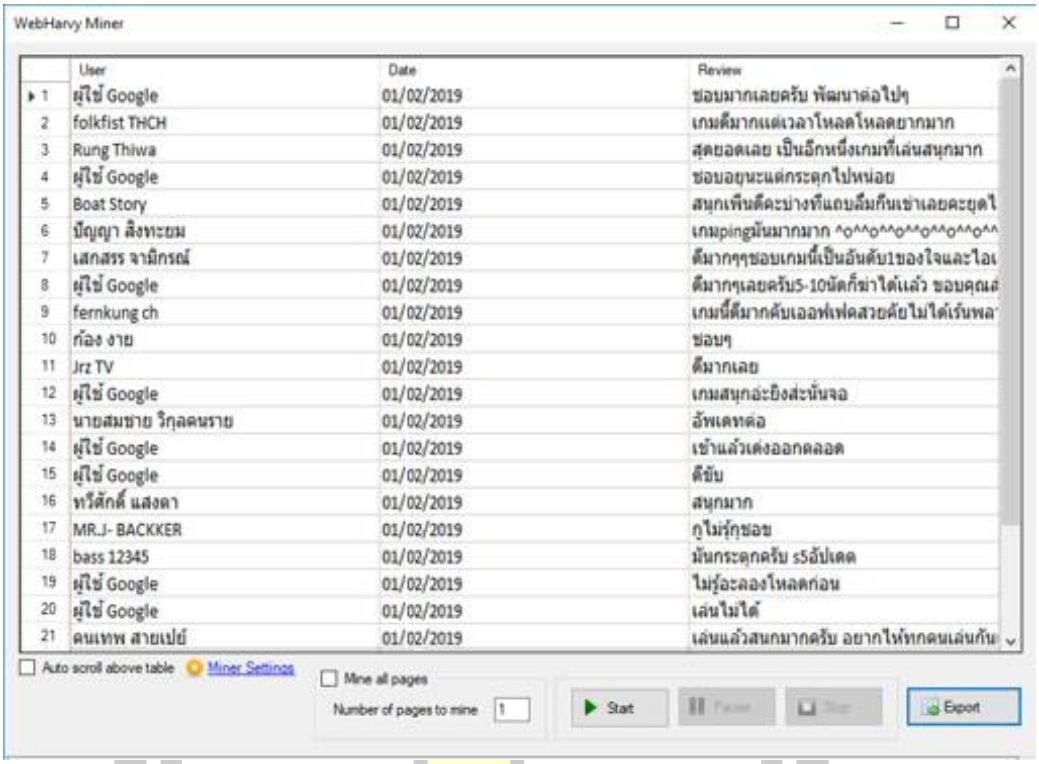
3.1 การเก็บรวบรวมข้อมูล

ในการเก็บรวบรวมข้อมูล ผู้วิจัยจะทำการเก็บรวบรวมข้อมูลความคิดเห็นต่อเกมมือถือผับจี ที่ตั้งแต่วันที่ 1 กุมภาพันธ์ 2562 ถึงวันที่ 15 กุมภาพันธ์ 2562 ทั้งนี้ข้อความที่ได้ทำการเก็บรวบรวมนั้นจะเป็นข้อความที่มีลักษณะเป็นภาษาไทยที่ได้มีการแสดงความคิดเห็นบน Google Play เท่านั้น ซึ่งจำนวนของความคิดเห็นที่ได้เก็บรวบรวมนั้นมีจำนวนทั้งสิ้น 3,798 ความคิดเห็น โดยในการเก็บรวบรวมข้อมูลนั้นจะเป็นการดึงข้อมูลความคิดเห็นผ่านการใช้งานโปรแกรม WebHarvy ดังแสดงในภาพที่ 3.1



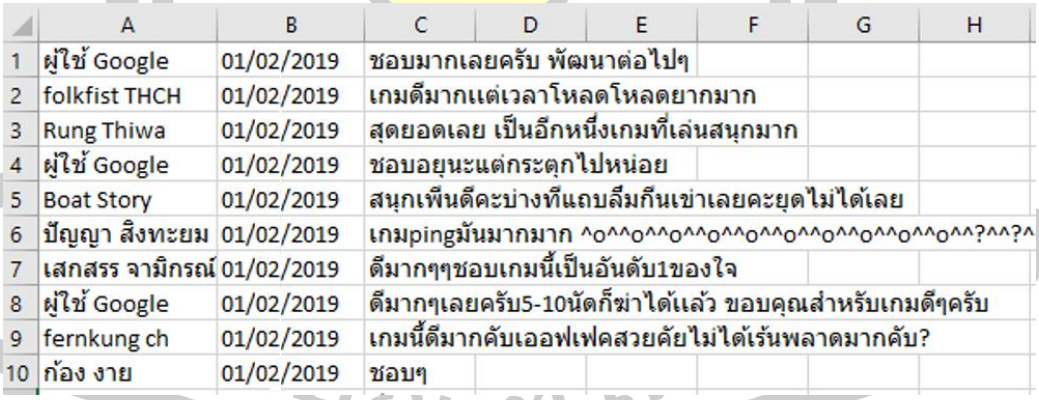
ภาพที่ 3.1 การเก็บข้อมูลผ่านโปรแกรม WebHarvy

จากนั้นทำการดึงข้อมูลความคิดเห็น โดยในแต่ละเรคคอร์ดจะประกอบไปด้วยชื่อผู้ใช้ วันที่ และข้อความความคิดเห็นของผู้ใช้งาน ดังแสดงในภาพที่ 3.2



ภาพที่ 3.2 ตัวอย่างการดึงข้อความความคิดเห็น

เมื่อทำการดึงข้อมูลออกมาแล้ว จะได้ไฟล์ข้อมูลที่อยู่ในรูปแบบของไฟล์ CSV (Comma-Separated Values) ซึ่งสามารถใช้โปรแกรม Microsoft Excel ในการเปิดดูข้อมูล หรือแก้ไขข้อมูลต่าง ๆ ได้ ดังแสดงในภาพที่ 3.3



	A	B	C	D	E	F	G	H
1	ผู้ใช้ Google	01/02/2019	ชอบมากเลยครับ พัฒนาต่อไปๆ					
2	folkfist THCH	01/02/2019	เกมดีมากแต่เวลาโหลดโหลดยากมาก					
3	Rung Thiwa	01/02/2019	สุดยอดเลย เป็นอีกหนึ่งเกมที่เล่นสนุกมาก					
4	ผู้ใช้ Google	01/02/2019	ชอบอยุนะแต่กระตุกไปหน่อย					
5	Boat Story	01/02/2019	สนุกเพื่นดีคะบ่างทีแถบลืมกินเข่าเลยคะยุดไม่ได้เลย					
6	ปัญญา สิงทะยม	01/02/2019	เกมpingมันมากมาก ^o^^o^^o^^o^^o^^o^^o^^o^^o^^?^^?^					
7	เสกสรร จามิกรณ์	01/02/2019	ดีมากๆๆชอบเกมนี้เป็นอันดับ1ของใจ					
8	ผู้ใช้ Google	01/02/2019	ดีมากๆเลยครับ5-10นัดก็ฆ่าได้แล้ว ขอบคุณสำหรับเกมดีๆครับ					
9	fernkung ch	01/02/2019	เกมนี้ดีมากคับเออฟเฟคสวยคัยไม่ได้เร้นพลาดมากคับ?					
10	ก้อง งาย	01/02/2019	ชอบๆ					

ภาพที่ 3.3 ข้อมูลที่อยู่ในรูปแบบไฟล์ CSV ที่ถูกเปิดด้วย Microsoft Excel

3.2 ขั้นตอนก่อนการสร้างแบบจำลอง

ขั้นตอนก่อนการสร้างแบบจำลอง เป็นขั้นตอนที่ใช้ในการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลเพื่อเข้าสู่กระบวนการวิเคราะห์ความคิดเห็นที่มีต่อเกมมือถือผับจี โดยจะนำความคิดเห็นที่ผ่านกระบวนการ

คัดแยกข้อความที่เกี่ยวข้อง ผ่านกระบวนการตัดคำ กำจัดคำหยุด มาทำการคัดเลือกคำเพื่อแยกประเภทของความคิดเห็นซึ่งจะทำการแบ่งคุณลักษณะที่สนใจออกเป็น 2 ลักษณะ คือ คุณลักษณะเชิงบวกและคุณลักษณะเชิงลบ จากนั้นจึงนำข้อมูลเข้าสู่กระบวนการถัดไป ซึ่งในกระบวนการนี้มีขั้นตอนดังต่อไปนี้

3.2.1 กระบวนการกลั่นกรองข้อมูล

เนื่องจากข้อคิดเห็นที่ได้จากผู้ใช้งานในแต่ละคนอาจมีความถูกต้องต่างกัน ซึ่งอาจเกิดจากการพิมพ์ผิด เช่น “กระโดด” เป็น “กะโดด” หรือใช้คำที่สั้นเพื่อประหยัดเวลาในการพิมพ์ เช่น “ไม่รู้” เป็น “มะรุ” รวมถึงคำที่ไม่สุภาพ เป็นต้น ในกระบวนการนี้จึงเป็นการนำเอาข้อความความคิดเห็นที่ได้เก็บรวบรวมมาทำการแก้ไขคำที่ผิดให้ถูกต้อง จากนั้นกำจัดตัวเลข ตัวอักษรภาษาอังกฤษ สัญลักษณ์ และช่องว่าง รวมถึงคำที่ไม่สุภาพออกไป โดยในการกำจัดดังกล่าวจะใช้การเขียนโปรแกรมด้วยภาษา Python เพื่อตัดสิ่งเหล่านั้นออกไป จากนั้นทำการคัดแยกข้อคิดเห็นที่เกี่ยวข้องกับหัวข้อที่กำหนดและจะต้องเป็นข้อความภาษาไทยเท่านั้น โดยหลังจากผ่านกระบวนการนี้จะเหลือข้อความความคิดเห็นจำนวน 3,123 ความคิดเห็น

3.2.2 กระบวนการตัดคำและกำจัดคำหยุด

หลังจากผ่านกระบวนการกลั่นกรองข้อมูลแล้วจะนำข้อมูลที่ได้มาทำการตัดคำ ดังภาพที่ 3.4 โดยการตัดคำนั้นจะใช้การตัดคำด้วยโปรแกรม LexTo [57] จากนั้นทำการกำจัดคำหยุดต่าง ๆ ออกไป เช่นคำว่า ครับ ค่ะ เป็นต้น โดยจะไม่ทำให้ความหมายของข้อความนั้นเปลี่ยนไป ซึ่งในการกำจัดคำหยุดจะใช้การเขียนโปรแกรมด้วยภาษา Python โดยใช้ไลบลารีของ PyThaiNLP

No	Comment	Num
1	ชอบ,มาก,เลย,ครับ,พัฒนา,ต่อไป	6
2	เกม,ดีมาก,แต่,เวลา,โหลด,โหลด,ยาก,มาก	8
3	สุดยอด,เลย,ลูกพี่,เป็น,อีก,หนึ่ง,เกม,ที่,เล่น,สนุก,มาก	11
4	ชอบ,อยู่,นะ,แต่,กระตุก,ไป,หน่อย	7
5	สนุก,เพลิน,ดี,ค่ะ,บางที,แทบ,ลืม,กินข้าว,เลย,ค่ะ,อยู่,ไม่ได้,เลย	13
6	เกม,มันส์,มาก,มาก,ชอบ,มาก	6
7	ดีมาก,ชอบ,เกม,นี้,เป็น,อันดับ,1,ของ,ใจ	9
8	ดีมาก,เลย,ครับ,นัด,ก็,ฆ่า,ได้,แล้ว,ขอบคุณ,สำหรับ,เกม,ดี,ครับ	13
9	เกม,นี้,ดีมาก,ครับ,เอฟเฟกต์,สวย,ครับ,ไม่ได้,เล่น,พลาด,มาก,ครับ	12
10	ชอบ	1

ภาพที่ 3.4 ตัวอย่างข้อความที่ถูกตัดคำ

3.2.3 กระบวนการเลือกคำบ่งชี้

หลังจากผ่านกระบวนการตัดคำและกำจัดคำหยุดแล้วจะเหลือคำทั้งหมดจำนวน 1,212 คำ จากนั้นจะนำคำแต่ละคำมาให้ประเภทของคำตามพจนานุกรม ซึ่งแบ่งเป็นทั้งหมด 11 ประเภท ได้แก่ คำกริยา คำนาม คำบุพบท คำวิเศษณ์ คำสรรพนาม คำสันธาน คำอุทาน คำหยุด วลี คำสแลง และคำไม่ตรงพจนานุกรม ดังแสดงในภาพที่ 3.5 โดยกระบวนการนี้ใช้พจนานุกรมอิเล็กทรอนิกส์ ฉบับราชบัณฑิตยสถาน ซึ่งพัฒนาโดย NECTEC

Rows	word	in documents	total	type	notice
1	กระตุก	1	1	v	คำกริยา
2	กินข้าว	1	1	v	คำกริยา
3	ขอบคุณ	1	1	v	คำกริยา
4	ฆ่า	1	1	v	คำกริยา
5	ชอบ	5	5	v	คำกริยา
6	ดี	2	2	adv	คำวิเศษณ์
7	ดีมาก	4	4	adv	คำวิเศษณ์
8	ต่อไป	1	1	adv	คำวิเศษณ์
9	นัด	1	1	n	คำนาม
10	พลาด	1	1	v	คำกริยา
11	พัฒนา	1	1	v	คำกริยา
12	มันส์	1	1	adv	สแลง
13	ลืม	1	1	v	คำกริยา
14	สนุก	2	2	adv	คำวิเศษณ์
15	สวย	1	1	adv	คำวิเศษณ์
16	สำหรับ	1	1	prep	คำบุพบท
17	สุดยอด	1	1	adv	คำวิเศษณ์
18	อันตับ	1	1	n	คำนาม
19	เกม	6	6	n	คำนาม

ภาพที่ 3.5 ตัวอย่างประเภทของคำตามความหมายในพจนานุกรม

จากนั้นเลือกคำวิเศษณ์ ซึ่งคำวิเศษณ์นี้จะสามารถสื่อถึงการแสดงอารมณ์เชิงบวกและเชิงลบได้ดี [58] และเลือกคำสแลง 2 คำ คือคำว่า “มันส์” มีความหมายว่า “สนุก” และคำว่า “เด็ด” มีความหมายว่า “เยี่ยม” ซึ่งความหมายของทั้งสองคำนี้เป็นคำวิเศษณ์ รวมทั้งหมด 208 คำ ดังภาพที่ 3.6 แล้วทำการแบ่งคำบ่งชี้คุณลักษณะออกมาโดยผู้เชี่ยวชาญจะได้คำบ่งชี้คุณลักษณะเชิงบวกจำนวน 70 คำ คำบ่งชี้คุณลักษณะเชิงลบจำนวน 57 คำ และคำที่ไม่บ่งชี้คุณลักษณะใดเลยจำนวน 81 คำ

Rows	word	in documents	total	type	notice
1	กระจอก	7	7	adv	คำวิเศษณ์
2	กรุณา	7	8	adv	คำวิเศษณ์
3	ขึ้นไป	1	1	adv	คำวิเศษณ์
4	คนเดียว	6	6	adv	คำวิเศษณ์
5	ครบเครื่อง	1	1	adv	คำวิเศษณ์
6	ครึ่ง	2	2	adv	คำวิเศษณ์
7	งอมแงม	1	1	adv	คำวิเศษณ์
8	จริง	1	1	adv	คำวิเศษณ์
9	เด็ด	2	2	adv	สแลง
10	ฉิบ	5	5	adv	คำวิเศษณ์
11	ฉิบหาย	30	32	adv	คำวิเศษณ์
12	ชัด	3	3	adv	คำวิเศษณ์
13	ชัดเจน	1	1	adv	คำวิเศษณ์
14	ช่าง	1	1	adv	คำวิเศษณ์
15	ซ้ำ	3	3	adv	คำวิเศษณ์
16	มันส์	159	160	adv	สแลง
17	ดังนี้	1	1	adv	คำวิเศษณ์
18	ดี	533	559	adv	คำวิเศษณ์
19	ดีมาก	239	241	adv	คำวิเศษณ์
20	ดีเยี่ยม	3	3	adv	คำวิเศษณ์

ภาพที่ 3.6 ตัวอย่างของคำวิเศษณ์และคำสแลงที่ใช้ในการบ่งชี้คุณลักษณะ

จากนั้นแทนข้อความให้อยู่ในรูปแบบโครงสร้างฟีเจอร์เวกเตอร์ ด้วยค่า TF-IDF ที่มีสมการการคำนวณค่าน้ำหนักของคำ เช่น ตัวอย่าง ประโยค S_1, S_2 และ S_3 ดังนี้

S_1 = ชอบ|พัฒนา|ต่อไป

S_2 = สุดยอด|เกม|เล่น|สนุก

S_3 = สนุก|เพลิน|ดี|แทบ|ลืม|กินข้าว|ไม่ได้

จากประโยคตัวอย่างจะได้คำที่จะนำมาวิเคราะห์ คือคำ (w) ในเซตของประโยคทั้งหมด (D) ที่นำมาวิเคราะห์ โดยที่ D=$\{S_1, S_2, S_3\}$ และ w={กินข้าว, ชอบ, ดี, ต่อไป, พัฒนา, ลืม, สนุก, สุดยอด, เกม, เพลิน, เล่น, แทบ, ไม่ได้}

แต่เนื่องจากขนาดของแต่ข้อความ (S) มีความยาว (คำ) ไม่เท่ากัน จึงต้องทำ Vector normalization ให้แต่ละข้อความมีความยาวเท่ากันก่อนคือ เวกเตอร์ 1 หน่วย (Unit vector) เพื่อให้ข้อความที่ถูกแปลงอยู่ในรูปแบบฟีเจอร์เวกเตอร์สามารถนำไปวิเคราะห์ได้อย่างถูกต้องและเหมาะสม ตัวอย่างที่ได้จากการแปลงข้อความให้เป็นฟีเจอร์เวกเตอร์ดังแสดงภาพที่ 3.7

TF-IDF	กินข้าว	ชอบ	ดี	ต่อไป	พัฒนา	ลืม	สนุก	สุดยอด	เกม	เพลิน	เล่น	แทบ	ไม่ได้
S_1	0	0.577	0	0.577	0.577	0	0	0	0	0	0	0	0
S_2	0	0	0	0	0	0	0.208	0.565	0.565	0	0.565	0	0
S_3	0.404	0	0.404	0	0	0.404	0.149	0	0	0.404	0	0.404	0.404

ภาพที่ 3.7 ตัวอย่างการแปลงข้อความให้เป็นฟีเจอร์เวกเตอร์ ด้วยค่า TF-IDF

จากนั้นแบ่งคำหรือวลีนั้นออกเป็น 2 กลุ่ม คือ คำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบ ถ้าหากน้ำหนักของคำคุณลักษณะเชิงบวกและเชิงลบเท่ากันข้อความนั้นจะถูกตัดออก ดังที่แสดงในภาพที่ 3.8

ROW	กระจอก	ดี	ดีมาก	มันส์	สนุก	สุดยอด	...	แย่	ALL	P	N	CLASS
1	1	0	0	0	0	0	0	0	1	0	1	N
2	0	0	0	0	0	0	0	0	1	0	1	N
3	0	0.707	0	0.707	0	0	0	0	1.414	1.414	0	P
4	0	0	0	0	0	0	0	1	1	0	1	N
5	0	0	0	0	1	0	0	0	1	1	0	P
6	0	0	0.894	0	0	0	0	0.447	1.342	0.894	0.447	P
7	0	0	0	0	1	0	0	0	1	1	0	P
8	0	0	0	0	0	0	0	1	1	0	1	N
9	0	0	0	0.707	0.707	0	0	0	1.414	1.414	0	P
10	0	0	0	0	0	1	0	0	1	1	0	P

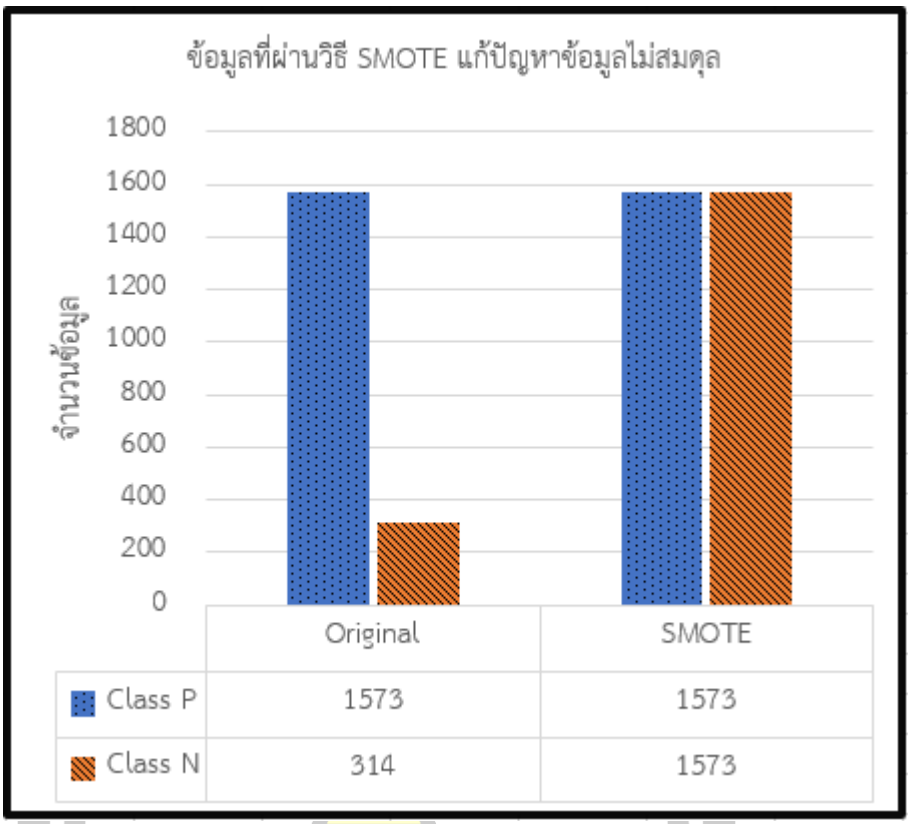
ภาพที่ 3.8 ตัวอย่างการกำหนดคุณลักษณะ

ซึ่งจะเหลือเพียงชุดข้อมูลที่เป็นคลาส P และ N เท่านั้น มีจำนวนทั้งหมด 1,887 ข้อมูล แบ่งเป็นคลาส P จำนวน 1,573 ข้อมูล และคลาส N จำนวน 314 ข้อมูล ดังภาพที่ 3.9

ROW	กระจอก	ดี	ดีมาก	มันส์	สนุก	สุดยอด	...	แย่	CLASS
1	1	0	0	0	0	0	0	0	N
2	0	0	0	0	0	0	0	0	N
3	0	0.707	0	0.707	0	0	0	0	P
4	0	0	0	0	0	0	0	1	N
5	0	0	0	0	1	0	0	0	P
6	0	0	0.894	0	0	0	0	0.447	P
7	0	0	0	0	1	0	0	0	P
8	0	0	0	0	0	0	0	1	N
9	0	0	0	0.707	0.707	0	0	0	P
10	0	0	0	0	0	1	0	0	P

ภาพที่ 3.9 ตัวอย่างของชุดข้อมูล

เมื่อพิจารณาข้อมูลแล้วจะพบว่าข้อมูลที่ได้มีปัญหาของความไม่สมดุลกันของข้อมูล คือ คลาส P มีข้อมูลจำนวน 1,573 ข้อมูล ส่วนคลาส N มีจำนวน 314 ข้อมูล ดังนั้นผู้วิจัยจึงนำวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) ซึ่งเป็นวิธีการเพิ่มจำนวนข้อมูลกลุ่มที่มีข้อมูลน้อยให้เพิ่มมากขึ้นเพื่อให้ใกล้เคียงหรือเท่ากันกับอีกคลาส ดังแสดงในภาพที่ 3.10



ภาพที่ 3.10 ข้อมูลที่ผ่านวิธี SMOTE แก้ปัญหาข้อมูลไม่สมดุล

หลังจากการปรับความสมดุลแล้วมีจำนวนข้อมูลเพิ่มขึ้น 1,259 ชุดข้อมูล ดังนั้นจะมีจำนวนข้อมูลทั้งสิ้นก่อนการสร้างแบบจำลองจำนวน 3,146 ข้อมูล โดยแบ่งเป็นคลาส P จำนวน 1,573 ชุดข้อมูลและคลาส N จำนวน 1,573 ชุดข้อมูล

3.3 การสร้างแบบจำลอง

ในกระบวนการสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการวิเคราะห์ความคิดเห็นที่มีต่อเกมมือถือผับจี ใช้ข้อมูลความคิดเห็นที่เป็นภาษาไทยบน Google Play โดยได้นำเทคนิคการจัดกลุ่มมาทำการจัดกลุ่มข้อมูลและใช้เทคนิคการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 4 เทคนิค ยิ่งไปกว่านั้นผู้วิจัยได้ใช้ค่าพารามิเตอร์ของแต่ละแบบจำลองที่ทำให้แต่ละเทคนิคมีประสิทธิภาพสูงสุดดังต่อไปนี้

1. เทคนิค Naïve Bayes
   - ไม่มีพารามิเตอร์
2. เทคนิค Support Vector Machine
   - kernel type: polynomial

- kernel degree: 1.0
- C: 0.0

3. เทคนิค K-Nearest Neighbor
   - k: 3
4. เทคนิคต้นไม้ตัดสินใจ C4.5
   - criterion: gain ratio
   - maximal depth: 40
   - confidence: 0.1
   - minimal gain: 0.01
   - minimal leaf size: 2

ในงานวิจัยนี้ผู้วิจัยได้ใช้แบบจำลองในการสร้างแบบจำลองเพื่อวัดประสิทธิภาพของการทำงานแบบจำลองเพื่อทำการเปรียบเทียบเทคนิคที่ใช้ในการวิเคราะห์ความคิดเห็นที่มีต่อเกมมือถือผับจี ในงานวิจัยนี้ผู้วิจัยได้ใช้โปรแกรม RapidMiner Studio ในการสร้างแบบจำลองเพื่อวัดประสิทธิภาพของเทคนิคทั้ง 4

3.4 การวัดประสิทธิภาพของแบบจำลอง

ในการประเมินผลของแบบจำลอง ได้มีการใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold cross validation โดยทำการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่าๆ กัน และทดสอบจำนวน 10 รอบซึ่งในแต่ละรอบของการทดสอบจะใช้ข้อมูล 1 ชุดเป็นชุดทดสอบและให้ชุดที่เหลือเป็นชุดฝึกสอน โดยผู้วิจัยจะทำการทดสอบแบบจำลองในการจัดกลุ่มความคิดเห็นที่มีต่อเกมมือถือผับจี โดยใช้แบบจำลองทีละ 1 แบบจำลอง และนำข้อมูลทั้งหมดที่ได้จากการเก็บข้อมูลการแสดงความคิดเห็นซึ่งผ่านกระบวนการเตรียมข้อมูลแล้ว มาทำการแบ่งออกเป็นทั้งหมด 10 กลุ่ม ทั้งนี้จะแบ่งกลุ่มข้อมูลเพื่อใช้เป็นข้อมูลทดสอบ (Test data) 1 ชุด และที่เหลือจะเป็นข้อมูลฝึก (Training data) ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อปริมาณข้อมูลฝึก คิดเป็นอัตราร้อยละ 10:90 โดยค่าที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity)

บทที่ 4
ผลการวิจัยและการอภิปราย

งานวิจัยนี้ได้ทำการวิเคราะห์ความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจี โดยนำเทคนิคของการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 4 เทคนิค ได้แก่ เทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 ผ่านการใช้งานโปรแกรม RapidMiner Studio โดยข้อมูลที่ใช้ในงานวิจัยนี้เก็บรวบรวมจาก Google Play ตั้งแต่วันที่ 1 - 15 กุมภาพันธ์ 2562 จำนวน 3,798 ชุดข้อมูล ซึ่งหลังจากผ่านกระบวนการต่าง ๆ แล้วจะเหลือชุดข้อมูลที่เตรียมเข้าสู่โมเดลทั้ง 4 จำนวน 1,887 ชุดข้อมูล แบ่งเป็นคลาส P จำนวน 1,573 ข้อมูล และคลาส N จำนวน 314 ข้อมูล ซึ่งทั้งสองคลาสมีความไม่สมดุลกันของข้อมูล จากปัญหาดังกล่าว ผู้วิจัยจึงได้นำวิธีการ SMOTE [9] ซึ่งเป็นวิธีการในการแก้ปัญหาข้อมูลไม่สมดุล (Imbalanced data) วิธีการนี้เป็นการเพิ่มคลาสที่มีกลุ่มน้อยให้เพิ่มมากขึ้นเพื่อให้ใกล้เคียงกันกับอีกคลาส โดยวิธีนี้จะสุ่มเพิ่มชุดข้อมูลของคลาสจำนวนน้อย ซึ่งหลังการปรับสมดุลของข้อมูลแล้วจะมีข้อมูลเข้าสู่แบบจำลองทั้งสิ้น 3,146 ชุดข้อมูล โดยในงานวิจัยได้ใช้ชุดข้อมูลที่มีการใช้การแทนข้อความต่างกันคือชุดข้อมูล Original ที่ใช้การ Weight โดยทั่วไป และชุดข้อมูลที่ใช้ TF-IDF จากนั้นสร้างแบบจำลองด้วยเทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 โดยใช้ 10-Fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ

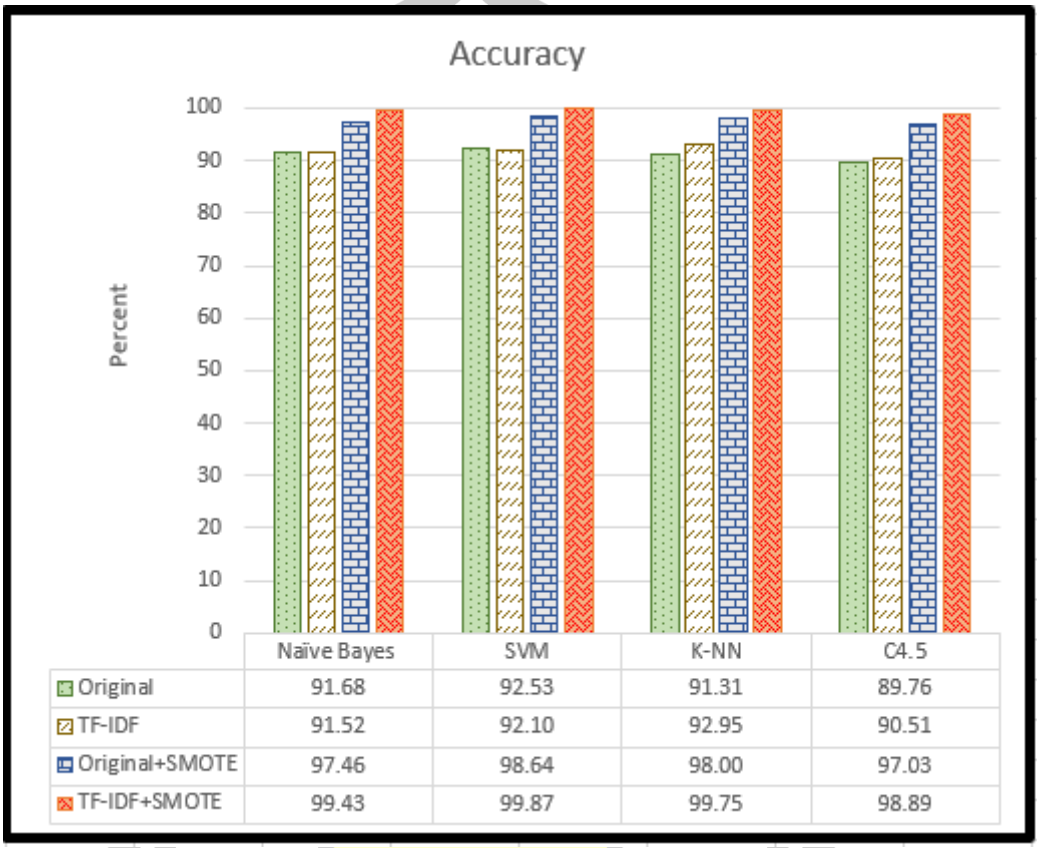
4.1 วิธีการวัดผลการวิจัย

การประเมินความสามารถของแบบจำลองการวิเคราะห์ความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจีส่วนใหญ่จะใช้วิธีการวัดประสิทธิผล โดยเน้นความสามารถของการตัดสินใจหรือทำนายผลของการจำแนกข้อมูลได้อย่างถูกต้อง ทั้งนี้จะเป็นการวัดค่าค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ

4.2 ผลการวิจัย

ในกระบวนการสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการวิเคราะห์ความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจีที่เป็นภาษาไทยด้วยเทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 โดยใช้ 10-Fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของ

แบบจำลองด้วยค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ ซึ่งจากการเปรียบเทียบค่าความถูกต้องของผลการนำข้อมูลเข้าสู่กระบวนการสร้างแบบจำลองของทั้ง 4 เทคนิค สามารถแสดงได้ดังนี้

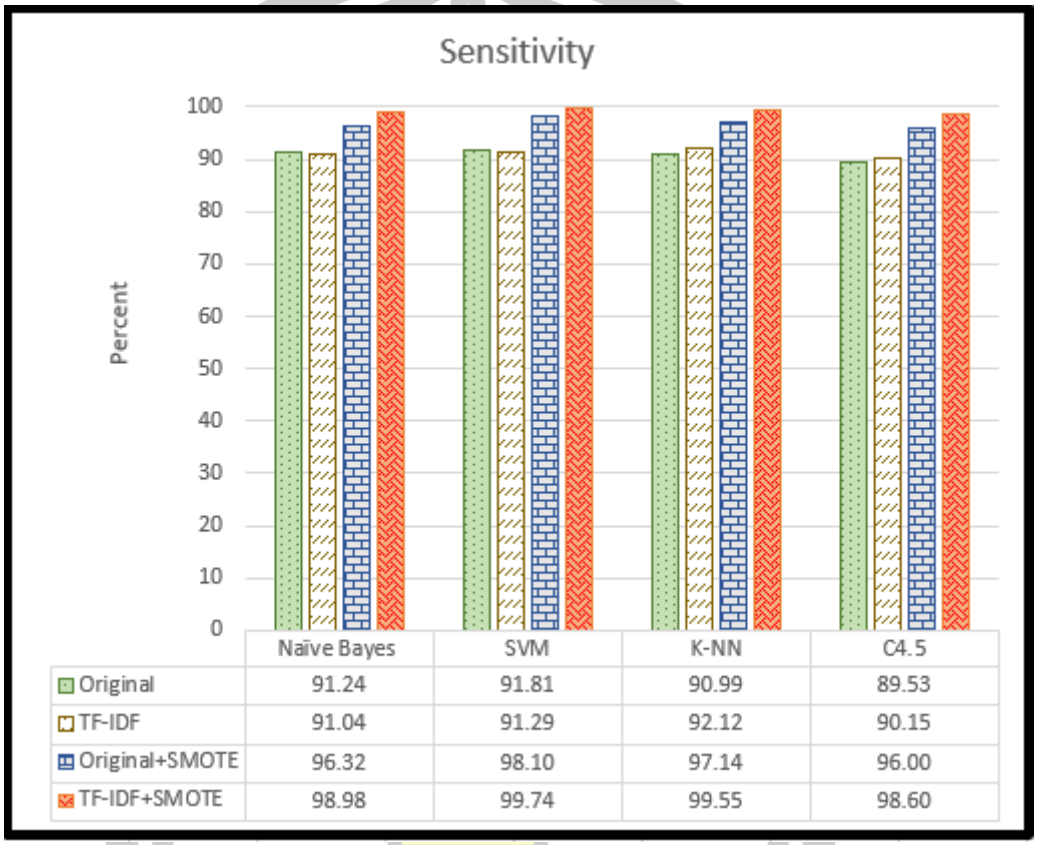


	Naïve Bayes	SVM	K-NN	C4.5
Original	91.68	92.53	91.31	89.76
TF-IDF	91.52	92.10	92.95	90.51
Original+SMOTE	97.46	98.64	98.00	97.03
TF-IDF+SMOTE	99.43	99.87	99.75	98.89

ภาพที่ 4.1 ค่าความถูกต้อง

จากภาพที่ 4.1 แสดงการเปรียบเทียบค่าความถูกต้องซึ่งถือว่าเป็นประสิทธิภาพโดยรวมของแบบจำลอง จากชุดข้อมูล Original ก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความถูกต้องสูงที่สุดคือ เทคนิค Support Vector Machine 92.53% รองลงมาคือเทคนิค Naïve Bayes 91.68% ตามด้วยเทคนิค K-Nearest Neighbor 91.31% และเทคนิคต้นไม้ตัดสินใจ C4.5 89.76% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความถูกต้องของแบบจำลองสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความถูกต้องสูงสุด 98.64% รองลงมาคือเทคนิค K-Nearest Neighbor 98.00% ตามด้วยเทคนิค Naïve Bayes 97.46% และเทคนิคต้นไม้ตัดสินใจ C4.5 ให้ค่าความถูกต้องน้อยที่สุด 97.03% ตามลำดับ และจากชุดข้อมูล TF-IDF ซึ่งก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความถูกต้องสูงที่สุดคือ เทคนิค K-Nearest Neighbor 92.95% รองลงมาคือเทคนิค Support Vector Machine 92.10% ตามด้วยเทคนิค Naïve Bayes 91.52% และเทคนิคต้นไม้ตัดสินใจ C4.5 90.51% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความถูกต้องของแบบจำลองสูงขึ้นทุกเทคนิค

โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความถูกต้องสูงสุด 99.87% รองลงมาคือเทคนิค K-Nearest Neighbor 99.75% ตามด้วยเทคนิค Naïve Bayes 99.43% และเทคนิคต้นไม้ตัดสินใจ C4.5 ให้ค่าความถูกต้องน้อยที่สุด 98.89% ตามลำดับ

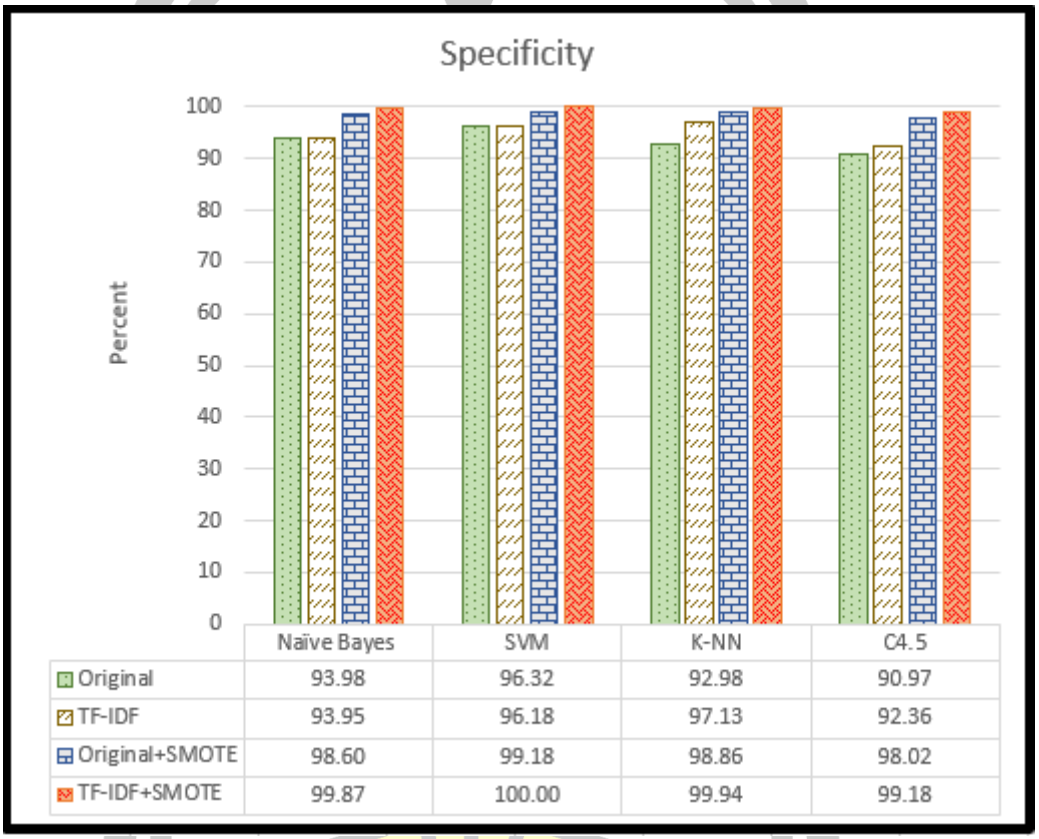


	Naïve Bayes	SVM	K-NN	C4.5
Original	91.24	91.81	90.99	89.53
TF-IDF	91.04	91.29	92.12	90.15
Original+SMOTE	96.32	98.10	97.14	96.00
TF-IDF+SMOTE	98.98	99.74	99.55	98.60

ภาพที่ 4.2 ค่าความไว

จากภาพที่ 4.2 แสดงการเปรียบเทียบค่าความไว ซึ่งการวัดค่าความไวเป็นการบอกถึงอัตราส่วนของการจำแนกถูกต้องในกลุ่มที่สนใจ เช่น กลุ่มที่สนใจ คือ กลุ่มความคิดเห็นเชิงบวก หมายความว่าโอกาสที่ความคิดเห็นที่เป็นเชิงบวกจะได้รับผลการจำแนกว่าเป็นความคิดเห็นเชิงบวก โดยจากชุดข้อมูล Original ก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความไวสูงสุดคือ เทคนิค Support Vector Machine 91.81% รองลงมาคือเทคนิค Naïve Bayes 91.24% ตามด้วยเทคนิค K-Nearest Neighbor 90.99% และเทคนิคต้นไม้ตัดสินใจ C4.5 89.53% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความไวของแบบจำลองสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความไวสูงสุด 98.10% รองลงมาคือเทคนิค K-Nearest Neighbor 97.14% ตามด้วยเทคนิค Naïve Bayes 96.32% และเทคนิคต้นไม้ตัดสินใจ C4.5 ให้ค่าความไวน้อยที่สุด 96.00% ตามลำดับ และจากชุดข้อมูล TF-IDF ก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความไวสูงสุดคือ เทคนิค K-Nearest Neighbor 92.12% รองลงมาคือเทคนิค

Support Vector Machine 91.29% ตามด้วยเทคนิค Naïve Bayes 91.04% และเทคนิคต้นไม้ตัดสินใจ C4.5 90.15% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความไวของแบบจำลองสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความไวสูงสุด 99.74% รองลงมาคือเทคนิค K-Nearest Neighbor 99.55% ตามด้วยเทคนิค Naïve Bayes 98.98% และเทคนิคต้นไม้ตัดสินใจ C4.5 ให้ค่าความไวน้อยที่สุด 98.60% ตามลำดับ



	Naïve Bayes	SVM	K-NN	C4.5
Original	93.98	96.32	92.98	90.97
TF-IDF	93.95	96.18	97.13	92.36
Original+SMOTE	98.60	99.18	98.86	98.02
TF-IDF+SMOTE	99.87	100.00	99.94	99.18

ภาพที่ 4.3 ค่าความจำเพาะ

จากภาพที่ 4.3 แสดงการเปรียบเทียบค่าความจำเพาะ ซึ่งการวัดค่าความจำเพาะเป็นการบอกถึงอัตราส่วนการจำแนกถูกต้องในกลุ่มอื่น ๆ เช่น กลุ่มความคิดเห็นเชิงลบ หมายความว่า โอกาสที่ความคิดเห็นที่เป็นเชิงลบจะได้รับผลการจำแนกว่าเป็นความคิดเห็นเชิงลบ โดยจากชุดข้อมูล Original ก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความจำเพาะสูงสุดคือ เทคนิค Support Vector Machine 96.32% รองลงมาคือเทคนิค Naïve Bayes 93.98% ตามด้วยเทคนิค K-Nearest Neighbor 92.98% และเทคนิคต้นไม้ตัดสินใจ C4.5 90.97% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความจำเพาะของแบบจำลองสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความจำเพาะสูงสุด 99.18% รองลงมาคือเทคนิค K-Nearest Neighbor 98.86% ตามด้วยเทคนิค Naïve Bayes 98.60% และเทคนิคต้นไม้ตัดสินใจ

C4.5 ให้ค่าความจำเพาะน้อยที่สุด 98.02% ตามลำดับ และจากชุดข้อมูล TF-IDF ก่อนการปรับความสมดุลของข้อมูลเทคนิคที่ให้ค่าความจำเพาะสูงสุดคือ เทคนิค K-Nearest Neighbor 97.13% รองลงมาคือเทคนิค Support Vector Machine 96.18% ตามด้วยเทคนิค Naïve Bayes 93.95% และเทคนิคต้นไม้ตัดสินใจ C4.5 92.36% ตามลำดับ และหลังจากปรับความสมดุลของข้อมูลด้วย SMOTE แล้ว ทำให้ค่าความจำเพาะของแบบจำลองสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค Support Vector Machine ให้ค่าความจำเพาะสูงสุด 100% รองลงมาคือเทคนิค K-Nearest Neighbor 99.94% ตามด้วยเทคนิค Naïve Bayes 99.87% และเทคนิคต้นไม้ตัดสินใจ C4.5 ให้ค่าความจำเพาะน้อยที่สุด 99.18% ตามลำดับ

พหุน ปณฺฑิโต ชีเว

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

ในการศึกษาเพื่อค้นหาเทคนิคการทำเหมืองข้อความที่มีประสิทธิภาพในการวิเคราะห์ความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจีซึ่งถูกเขียนขึ้นด้วยภาษาไทย ซึ่งเก็บรวบรวมข้อความความคิดเห็นจาก Google Play ตั้งแต่วันที่ 1 กุมภาพันธ์ – 15 กุมภาพันธ์ 2562 จำนวนทั้งหมด 3,798 ข้อความ โดยงานวิจัยนี้ได้ทำการแบ่งคุณลักษณะออกเป็น 2 กลุ่ม คือ คุณลักษณะเชิงบวกและคุณลักษณะเชิงลบ ซึ่งได้นำเอาเทคนิควิธีการวิเคราะห์เหมืองข้อความมาทำการวิเคราะห์ข้อความความคิดเห็นโดยทำการเปรียบเทียบทั้งหมด 4 เทคนิควิธีด้วยกันคือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor และเทคนิคต้นไม้ตัดสินใจ C4.5 และทำการวัดประสิทธิภาพด้วยวิธีการ 10-fold- cross-validation

5.1 สรุปผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคการทำเหมืองข้อความที่มีประสิทธิภาพในการวิเคราะห์ความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจีซึ่งถูกเขียนขึ้นด้วยภาษาไทย ซึ่งเก็บรวบรวมข้อความความคิดเห็นจาก Google Play ตั้งแต่วันที่ 1 กุมภาพันธ์ – 15 กุมภาพันธ์ 2562 จำนวนทั้งหมด 3,798 ข้อความ โดยงานวิจัยนี้ได้ทำการแบ่งคุณลักษณะออกเป็น 2 กลุ่ม คือ คุณลักษณะเชิงบวกและคุณลักษณะเชิงลบ โดยนำเอาคำวิเศษณ์และคำสแลงจากความคิดเห็นของผู้เล่นที่มีต่อเกมมือถือผับจี ซึ่งคำวิเศษณ์นี้สามารถแสดงถึงอารมณ์เชิงบวกและเชิงลบได้ดี [58]

โดยในกระบวนการเตรียมข้อมูลเพื่อหาคำบ่งชี้ในแต่ละคุณลักษณะได้เลือกวิธีในการตัดคำแบบอิงพจนานุกรม [57] โดยข้อดีของการตัดคำด้วยวิธีนี้คือสามารถเพิ่มคำศัพท์อื่น ๆ ได้ และประเภทของคำที่ใช้ในการกำหนดคุณลักษณะทั้งสองคุณลักษณะนั้นใช้เพียงคำวิเศษณ์ และคำสแลงที่มีความหมายของคำเป็นคำวิเศษณ์ที่สามารถระบุความหมายที่เป็นคุณลักษณะเชิงบวกและเชิงลบได้เท่านั้น ดังนั้นผลของการตัดคำอาจมีผลทำให้คำบางคำไม่พบในพจนานุกรม และไม่สามารถระบุประเภทของคำ ๆ นั้นได้ อีกทั้งลักษณะของคำที่ปรากฏยังมีลักษณะเป็นภาษาธรรมชาติซึ่งเป็นลักษณะภาษาที่ไม่ได้มีการกำหนดรูปแบบที่แน่นอนเอาไว้ ซึ่งในการแยกความหมายและประเภทของคำในภาษาธรรมชาตินั้นทำได้ยาก ถึงแม้ว่าคำนั้นอาจสามารถเป็นคำที่บ่งชี้คุณลักษณะได้ ดังนั้นคำที่ใช้ในการวิเคราะห์อาจไม่ได้มีเฉพาะคำที่เป็นคำวิเศษณ์หรือคำสแลงที่มีความหมายของคำเป็นคำวิเศษณ์อย่างเดียวเท่านั้น ซึ่งในงานวิจัยได้ใช้ชุดข้อมูลที่มีการใช้การแทนข้อความต่างกันคือชุดข้อมูล Original ที่ใช้การ Weight โดยทั่วไป และชุดข้อมูลที่ใช้ TF-IDF เพื่อเปรียบเทียบว่าการแทนข้อความ

ด้วยวิธีใดมีประสิทธิภาพ และจากข้อมูลที่ใช้ในการสร้างแบบจำลองที่มีความไม่สมดุลกันของข้อมูล ผู้วิจัยได้แก้ปัญหาโดยการนำวิธีการ SMOTE มาใช้เพื่อปรับให้ข้อมูลมีความสมดุลกัน

จากผลการวิจัยพบว่าวิธีการ SMOTE สามารถทำให้ประสิทธิภาพของแบบจำลองเพิ่มขึ้นทุกเทคนิค และจากการเปรียบเทียบค่าความถูกต้องของชุดข้อมูล Original ที่มีการปรับสมดุลของข้อมูลด้วย SMOTE กับชุดข้อมูล TF-IDF ที่มีการปรับสมดุลของข้อมูลด้วย SMOTE พบว่าชุดข้อมูลที่ใช้การแทนข้อความด้วย TF-IDF ร่วมกับการปรับสมดุลของข้อมูลด้วย SMOTE มีประสิทธิภาพดีกว่า โดยจากการเปรียบเทียบแบบจำลองทั้ง 4 เมื่อพิจารณาเป็นด้านสามารถสรุปได้ว่า ค่าความถูกต้องเพิ่มขึ้นเฉลี่ย 1.70% ค่าความไวเพิ่มขึ้นเฉลี่ย 2.33% และค่าความจำเพาะเพิ่มขึ้นเฉลี่ย 1.08% ซึ่งเทคนิค Support Vector Machine ให้ผลการทดสอบดีที่สุด โดยมีค่าความถูกต้อง 99.87% ค่าความไว 99.74% และค่าความจำเพาะ 100%

5.2 อภิปรายผล

จากการทดลองแม้ว่าแบบจำลองของชุดข้อมูลที่มีการแทนข้อความด้วยค่า TF-IDF ร่วมกับการปรับสมดุลของข้อมูลด้วย SMOTE ที่ใช้ในการทดสอบนี้จะได้ผลลัพธ์ที่ดี แต่อย่างไรก็ตามข้อผิดพลาดที่เกิดขึ้นจากการจำแนกกลุ่มของแบบจำลองทั้ง 4 นี้ ส่วนใหญ่จะเกิดข้อผิดพลาดขึ้นในกระบวนการของการตัดคำ เนื่องจากข้อความภาษาไทยนั้นมีโครงสร้างของประโยคที่ไม่แน่นอนอีกทั้งความหมายของคำที่มีต่อประโยคนั้น ๆ มีหลายความหมายด้วยกัน รวมไปถึงคำบางคำเป็นที่เป็นคำศัพท์เฉพาะที่ใช้เกี่ยวกับเกม ซึ่งคำบ่งชี้คุณลักษณะที่ใช้ในงานวิจัยนี้นำมาจากข้อความที่ใช้ในการวิจัยนี้เท่านั้น

ดังจะเห็นได้จากผลการทดลองในการเปรียบเทียบของแบบจำลองทั้ง 4 นี้ เทคนิค Support Vector Machine เป็นเทคนิคที่มีประสิทธิภาพดีที่สุดเมื่อเปรียบเทียบกับเทคนิค K-Nearest Neighbor เทคนิค Naïve Bayes และเทคนิคต้นไม้ตัดสินใจ C4.5 นอกจากนั้นเทคนิค Support Vector Machine ยังสามารถเพิ่มประสิทธิภาพได้ด้วยวิธีการเลือกใช้ฟังก์ชั่นเคอร์เนล และการปรับพารามิเตอร์ที่เหมาะสมเพื่อประสิทธิภาพที่ดีที่สุด โดยใช้ฟังก์ชั่นเคอร์เนล Polynomial และค่าพารามิเตอร์ ดังนี้

พารามิเตอร์ดีกรี เท่ากับ 1 หมายถึง การปรับค่าความโค้งของเส้นไฮเปอร์เพลนที่เหมาะสมกับชุดข้อมูล เนื่องจากชุดข้อมูลที่นำมาเรียนรู้และทดสอบไม่สามารถแบ่งได้ด้วยเส้นตรงหรือได้แต่ประสิทธิภาพไม่ดีเท่าที่ควร การปรับเส้นไฮเปอร์เพลนให้มีความโค้งที่สอดคล้องกับชุดข้อมูลจะทำให้ได้เส้นแบ่งชุดข้อมูล 2 กลุ่มออกจากกันได้อย่างเหมาะสม ซึ่งได้แก่ค่าพารามิเตอร์ดีกรี เท่ากับ 1 และหากปรับเพิ่มค่าดีกรีจะได้เส้นแบ่งที่โค้งมากขึ้น ซึ่งทำให้ได้เส้นแบ่งชุดข้อมูลที่ไม่เหมาะสม นอกจากนั้นแล้วหากเพิ่มค่าดีกรีให้มีจำนวนมากขึ้นจะทำให้ต้องใช้ระยะเวลาการคำนวณที่มากขึ้นตามไปด้วย ซึ่งหากปรับค่าดีกรีต่ำที่สุดก็คือการแบ่งด้วยเส้นตรงนั่นเอง

พารามิเตอร์ C เท่ากับ 0 หมายถึงค่า C ที่เหมาะสมสำหรับการสร้างแบบจำลองการแบ่งกลุ่มข้อมูลที่มีการกำหนดให้ยอมรับความผิดพลาดที่เกิดขึ้นของชุดข้อมูลบางชุดได้ และมีระยะห่างระหว่างกลุ่ม 2 กลุ่มข้อมูลที่เหมาะสม กล่าวคือไม่เกิดปัญหาเรื่องการเข้ากันเกินไป (Overfitting) หรือการสร้างแบบจำลองที่สามารถนำมาใช้ในการเรียนรู้ได้ดี แต่ไม่สามารถใช้แบ่งกลุ่มข้อมูลอื่น ๆ ได้หรือแบ่งกลุ่มได้ไม่ดี

ดังนั้นการปรับค่าพารามิเตอร์ควรพิจารณาถึงความเหมาะสมของชุดข้อมูล การกำหนดค่าพารามิเตอร์ที่สูงเกินไปอาจเกิดปัญหาเรื่องการเข้ากันมากเกินไปได้ และหากปรับค่าพารามิเตอร์น้อยเกินไปจนไม่เหมาะสมกับข้อมูล จะทำให้ได้ผลลัพธ์ที่ไม่สามารถวิเคราะห์ข้อมูลหรือนำไปใช้ประโยชน์ได้ให้

5.3 ข้อเสนอแนะ

สำหรับงานวิจัยครั้งต่อไป อาจมีการนำคำบางคำที่เป็นศัพท์เฉพาะมาใช้เป็นคำบ่งชี้คุณลักษณะและพิจารณาเลือกชุดข้อมูลที่มีความสมดุลกันของข้อมูล เพื่อเพิ่มประสิทธิภาพให้กับการวิเคราะห์ประเภทของคำและประโยคได้ดีมากขึ้น

พหูน ปณฺฑิโต ชีเว

บรรณานุกรม

[1] Mthai. (2019, May 5). PUBG Mobile. Available: https://game.mthai.com/mobile-games/112880.html

[2] J. Zagal, A. Ladd, and T. Johnson, Characterizing and understanding game reviews. 2009, pp. 215-222.

[3] ยืน ภู่วรวรรณ, การประมวลผลภาษาธรรมชาติ. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี, 1992, p. 241.

[4] N. R, N. Tikoo, S. Kurle, H. S. Pisupati, and D. P. G. R, "A Survey on Text Mining and Sentiment Analysis for Unstructured Web Data," International Journal of Emerging Technologies and Innovative Research vol. 2, no. 4, pp. 1292-1296, 2015.

[5] กานดา แผ่วัฒนากุล and ปราโมทย์ ลือนาม, "การวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายสังคมออนไลน์," บทความ วิชาการวารสารการจัดการสมัยใหม่ปีที่ 11, vol. 11, no. 2, กรกฎาคม - ธันวาคม 2013.

[6] ประพัฒน์ พรมน้ำอ่าง, วสุวรรธน์ พงศ์ขจร, and นิเวศ จิระวิชิตชัย, "การจำแนกกลุ่มข้อความรีวิวโดยใช้เทคนิคเหมืองข้อมูล," วารสารวิทยาศาสตร์และเทคโนโลยี มทร. ธัญบุรี ปีที่ : 6 vol. 1, pp. 94-102, 2016.

[7] พัชรนิกานต์ พงษ์ธนู, "วิเคราะห์ความพึงพอใจของลูกค้าจากข้อความคำแนะนำโดยการทำเหมืองความคิดเห็น," โครงการประชุมวิชาการนานาชาติ Knowledge and Smart Technologies, no. 1, pp. 53-60 2012.

[8] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment," in 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), 2016, pp. 52-57.

[9] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res. (JAIR), vol. 16, pp. 321-357, 01/01 2002.

[10] เชาวนันท์ โสโท, "แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม," KKU Res J (GS), vol. 13, pp. 39-49, 2013.

[11] สิริวรรณ ปัญญากาศ, "ความสัมพันธ์ระหว่างความรุนแรงในสื่อเกมออนไลน์กับพฤติกรรมก้าวร้าว ของเด็กวัยรุ่นตอนปลายในจังหวัดเชียงใหม่," มหาวิทยาลัยเชียงใหม่2008, Available: http://cmuir.cmu.ac.th/handle/6653943832/15240.

[12] ศิริวรรณ จันทร์เชื้อ, "พฤติกรรมการเล่นเกมออนไลน์ของนักเรียนช่วงชั้นที่ 2 ในกรุงเทพมหานคร และจังหวัดนครสวรรค์," มหาวิทยาลัยเกษตรศาสตร์, 2008.

[13] ธนพัทธ์ เอมะบุตร, "ปัจจัยที่มีอิทธิพลต่อการตัดสินใจเลือกเล่นเกมออนไลน์ ของประชาชนในกรุงเทพมหานครปี 2558," มหาวิทยาลัยกรุงเทพ, 2015.

[14] dotlife. (2018, Oct 4). แนวเกมในปัจจุบัน แนวไหนที่เหมาะกับคุณ. Available: https://www.dotlife.store/blog/game/

[15] J. Forward. (2019, Mar 27). PUBG new map – all the latest details on the upcoming snow map, Vikendi. Available: https://www.pcgamesn.com/playerunknowns-battlegrounds/playerunknowns-battlegrounds-new-maps

[16] C. Carter. (2019, Mar 27). Understanding Playerunknown's Battlegrounds. Available: https://www.polygon.com/playerunknowns-battlegrounds-guide/2017/6/9/15721366/pubg-how-to-play-blue-wall-white-red-circle-map-weapon-vehicle-inventory-air-drop

[17] A. Bradley. (2019, Mar 27). Designing the giant battle royale maps of Playerunknown's Battlegrounds. Available: http://www.gamasutra.com/view/news/296881/Designing_the_giant_battle_royale_maps_of_Playerunknowns_Battlegrounds.php

[18] T. Marks. (2019, Mar 27). Battle Royale modder Brendan Greene on his official H1 Z1 mode. Available: http://www.pcgamer.com/battle-royale-modder-brendan-greene-on-his-official-h1z1-mode/

[19] J. Batchelor. (2019, Mar 27). PlayerUnknown Battlegrounds dev: "I want to find the next me". Available: http://www.gamesindustry.biz/articles/2017-04-26-playerunknown-battlegrounds-dev-i-want-to-find-the-next-me

[20] C. Kerr. (2019, Mar 27). Skin economy is a good thing,' says Playerunknown's Battlegrounds creator. Available: http://www.gamasutra.com/view/news/297434/Skin_economy_is_a_good_thing_says_Playerunknowns_Battlegrounds_creator.php

[21] A. McAloon. (2019, Mar 27). Tencent working to bring Battlegrounds to mobile in China. Available: https://www.gamasutra.com/view/news/310471/Tencent_working_to_bring_Battlegrounds_to_mobile_in_China.php

[22] J. Brightman. (2019, Mar 27). PUBG going mobile with Tencent's help. Available: http://www.gamesindustry.biz/articles/2017-11-27-pubg-going-mobile-with-tencents-help

[23] I. Higton. (2019, Mar 27). Here's how PUBG on mobile phones compares to the original game. Available: http://www.eurogamer.net/articles/2018-02-16-heres-how-pubg-on-mobile-phones-compares-to-the-original-game

[24] S. Arif. (2019, Mar 27). PUBG's Chinese Mobile Games Are Now in Early Access. Available: http://www.ign.com/articles/2018/02/09/pubgs-chinese-mobile-games-are-now-in-early-access

[25] M. Madnani. (2019, Mar 27). Both of Tencent's 'PUBG' Mobile Games Are Now Available on the App Store in China for Free. Available: http://toucharcade.com/2018/02/09/pubg-mobile-china-free-released/

[26] S. Saed. (2019, Mar 27). PUBG Mobile gets surprise release in the West. Available: https://www.vg247.com/2018/03/15/pubg-mobile-gets-surprise-release-in-the-west/

[27] A. Chalk. (2019, Mar 27). PUBG Mobile comes to the West with a 'soft launch' in Canada. Available: https://www.pcgamer.com/pubg-mobile-comes-to-the-west-with-a-soft-launch-in-canada/

[28] K. Knezevic. (2019, Mar 27). PUBG Mobile Out Now For Free In The US And Other Regions. Available: https://www.gamespot.com/articles/pubg-mobile-out-now-for-free-in-the-us-and-other-r/1100-6457509/

[29] บุญเสริม กิจศิริกุล, ปัญญาประดิษฐ์ (ARTIFICIAL INTELLIGENCE), 2003, p. 217. [Online]. Available.

[30] อัศนีย์ ก่อตระกูล, "การประมวลผลภาษามนุษย์ด้วยคอมพิเตอร์," มนุษยศาสตร์วิชาการ, vol. 6, no. 1, 1998.

[31] มาลินี นาคใหญ่, "การวิเคราห์ประโยคข้อมูล โดยอิงฐานความรู้กรณีศึกษาการครองราชย์สมัยกรุงศรีอยุธยา," Master, สาขาวิชาวิทยาการคอมพิวเตอร์, มหาวิทยาลัยศิลปากร, 2006.

[32] สายัณห์ เทพแดง, "การศึกษาเปรียบเทียบการใช้และการไม่ใช้คำนำหน้าชื่อ คำต่อท้ายในการจดจำนิพจน์ระบุนาม," in การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ, 2015, pp. 358-362.

[33] สรวัตร ประภานิติเสถียร and ไกรศักดิ์ เกษร, "การตรวจการโจรกรรมทางวิชาการด้วยใช้เทคนิค N-gram ร่วมกับเทคนิคการตรวจสอบเชิงความหมายสำหรับเอกสารภาษาไทย," วารสารวิทยาการและเทคโนโลยีสารสนเทศ, vol. 5, no. 1, pp. 42-50, 2015.

[34] กรมวุฒิ นงนุช, อนุชา ซาเฮาะ, and สุวุฒิ ตุ้มทอง, "การวิเคราะห์บทความอัตโนมัติ โดยใช้กระบวนการภาษาธรรมชาติ," in การประชุมวิชาการระดับชาติมหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ ครั้งที่ 1, 2016, pp. 472-479.

[35] Notebookspec. (2019, Mar 28). Google Play คืออะไร? Available: https://notebookspec.com/google-play-คืออะไร/195449/

[36] H. Cherfi, A. Napoli, and Y. Toussaint, Towards a Text Mining Methodology Using Association Rules Extraction. 2006, pp. 431-441.

[37] สมศักดิ์ วิชัยกิจ and มาลีรัตน์ โสดานิล, "การจัดกลุ่มลูกค้าสินเชื่อจากการปฏิเสธด้วยวิธีการทำเหมืองข้อความ," การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 11, pp. 359-363, 2015.

[38] N. Zhong, Y. Li, and S. Wu, "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30-44, 2012.

[39] นพมาศ อัครจันทโชต and ดิเรก พนิตสุภากมล, "การเปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุล สำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1," presented at the การประชุมเสนอผลงานวิจัยระดับชาติ มหาวิทยาลัยสุโขทัยธรรมาธิราช ครั้งที่ 9 อาคารเฉลิมพระเกียรติ 80 พรรษา 5 ธันวาคม 2550 มหาวิทยาลัยสุโขทัยธรรมาธิราช, 29/11/2019, 2019.

[40] G. Backfried et al., "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project," in 2013 European Intelligence and Security Informatics Conference, 2013, pp. 143-146.

[41] จุฑาทิพย์ ทิพย์พูล and นิเวศ จิระวิชิตชัย, "การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยการใช้เทคนิคเหมืองข้อมูล," Science and Technology RMUTT Journal, vol. 6, no. 8, pp. 102-109, 2016.

[42] รวิสุดา เทศเมือง and นิเวศ จิระวิชิตชัย, "การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน," วารสารวิศวกรรมศาสตร์

มหาวิทยาลัยสยาม, vol. 18, no. 1, pp. 1-11, 2017.

[43] อรทิพย์ เลื่อยงาม and ชัยพร เขมะภาตะพันธ์, "การจัดประเภทเอกสารด้วยวิธีเอสวีเอ็ม เพื่อการป้องกันเอกสารรั่วไหล," การประชุมวิชาการ "นเรศวรวิจัย" ครั้งที่ 7, pp. 3-12, 2011.

[44] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in Machine Learning: ECML-98, Berlin, Heidelberg, 1998, pp. 137-142: Springer Berlin Heidelberg.

[45] สมมาตร อังคเศรณีกุล and ชุลีรัตน์ จรัสกุลชัย, "การตรวจข้อสอบอัตนัยภาษาไทยแบบอัตโนมัติโดยใช้ K-NN," in การประชุมทางวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 48, 2012, pp. 35-42.

[46] บรรหาร จันทะวงค์ et al., "การจัดประเภทเอกสารงานวิจัยทางโลจิสติกส์ กรณีศึกษามหาวิทยาลัยราชภัฏเชียงราย," in การประชุมวิชาการ งานวิจัยและพัฒนาเชิงประยุกต์ ครั้งที่ 7, 2015, pp. 491-494.

[47] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of Twitter during a natural disaster: Case study of 2011 Thai Flood," in 2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies, 2012, pp. 2227-2232.

[48] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," Machine Learning, vol. 16, no. 3, pp. 235-240, 1994/09/01 1994.

[49] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space," Expert Systems with Applications, vol. 36, no. 10, pp. 12200-12209, 2009/12/01/ 2009.

[50] S. Vongsingthong and N. Wisitpongphan, "Classification of university students' behaviors in sharing information on Facebook," in 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2014, pp. 134-139.

[51] อดิเทพ ไชยสาร and รัฐสิทธิ์ สุขะหุต, "การประมาณอารมณ์จากความคิดเห็นภาษาไทยโดยใช้เทคนิคการเรียนรู้ของเครื่อง," in การประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 9, 2013, pp. 260-266.

[52] วลัยลักษณ์ สุขสมบุญ and สมชาย ปราการเจริญ, "การเปรียบเทียบประสิทธิภาพการจำแนกประเภทปัญหาสำหรับระบบถามตอบโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และเคเนียร

เรสต์เนเบอร์.," in การประชุมวิชาการ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ครั้งที่ 7, 2010, pp. 747-756.

[53] K. G. Reddy and J. Gopal, "Twitter sentiment analysis of game reviews using machine learning techniques," Journal of Chemical and Pharmaceutical Sciences, vol. 10, no. 1, 2017.

[54] H.-N. Kang, H.-R. Yong, and H.-S. Hwang, "A Study of Analyzing on Online Game Reviews using a Data Mining Approach: STEAM Community Data," International Journal of Innovation, Management and Technology, vol. 8, no. 2, pp. 90-94, 2017.

[55] R. Bais, P. Odek, and S. Ou, "Sentiment Classification on Steam Reviews," 2017.

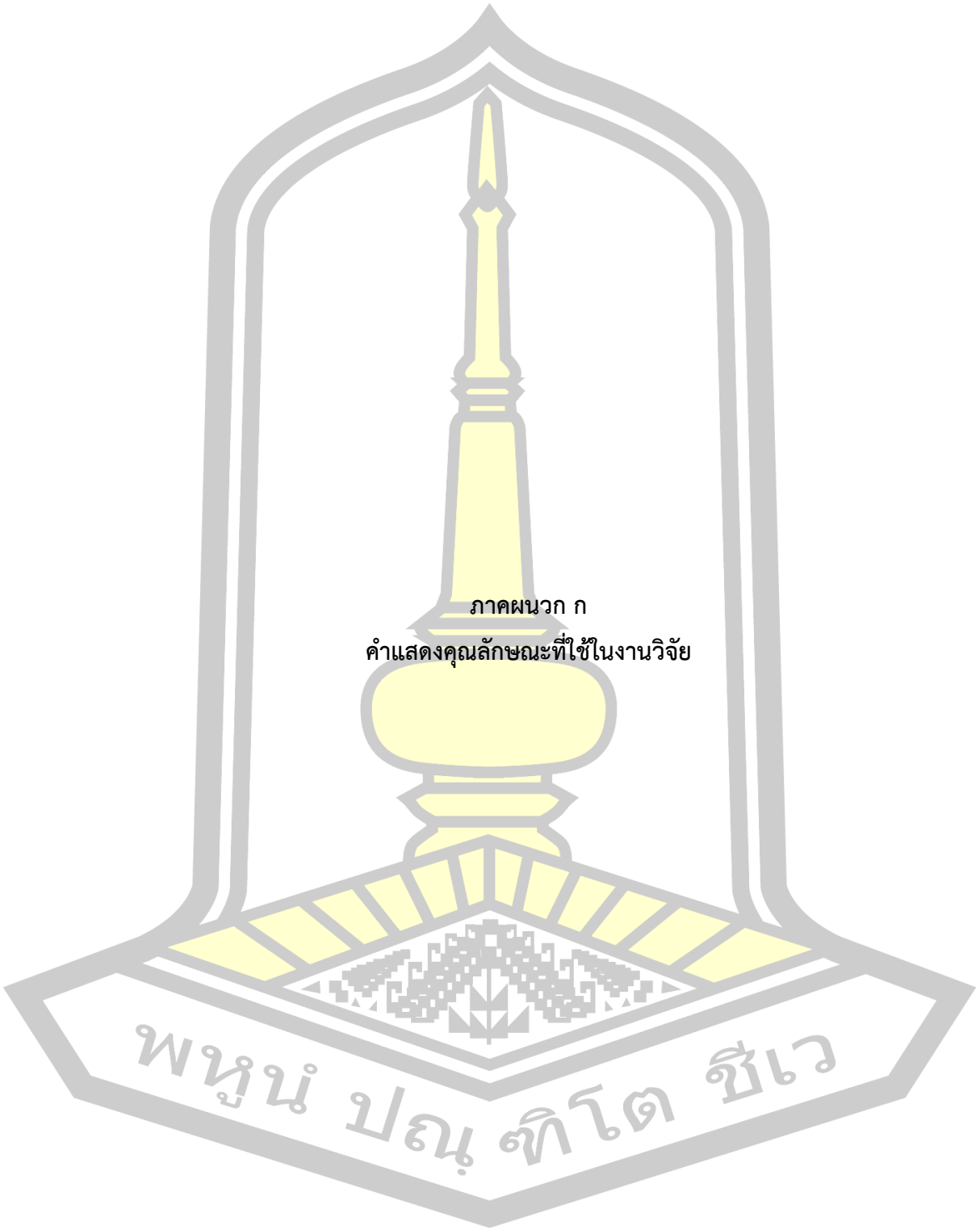
[56] M. Bouazizi and T. Ohtsuki, "Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis," in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2015, pp. 1594-1597.

[57] ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. LexTo: เล็กซ์โตโปรแกรมตัดคำสำหรับข้อความภาษาไทย. Available: http://www.sansarn.com/lexto/

[58] Y. Sun, C. Quan, X. Kang, Z. Zhang, and F. Ren, "Customer emotion detection by emotion expression analysis on adverbs," Information Technology and Management, journal article vol. 16, no. 4, pp. 303-311, December 01 2015.

พหุน ปณ ฑิโต ชีเว

ภาคผนวก



ภาคผนวก ก
คำแสดงคุณลักษณะที่ใช้ในงานวิจัย

คำแสดงคุณลักษณะที่ใช้ในงานวิจัย

คำแสดงคุณลักษณะเชิงบวก จำนวน 70 คำ (ก่อนนำคำที่มีค่าน้ำหนักน้อยว่า 2 ออก)

กรุณา	พอได้	แม่น	สุดท้าย
เก่ง	เต็ม	ยอด	สุดยอด
ขึ้นไป	เต็มที่	ยอดเยี่ยม	สุ่ม
ครบเครื่อง	ถนัด	เยอะแยะ	เสถียร
จริง	ที่หนึ่ง	เยี่ยม	เสมอ
เจ๋ง	นิดเดียว	เรียบร้อย	หนัก
แจ่ม	ปานกลาง	ลื่น	หนา
แจ๋ว	เปล่า	ไว	หน้าเดิม
ชัด	พอใช้	สนุก	หน้าใหม่
ชัดเจน	พอใช้ได้	สนุกสนาน	หนึบ
ใช้ได้	พอได้	สมจริง	หล่อ
ด้วยดี	เพียบ	สมดุล	หลากหลาย
ดี	เพียว	ส่วนตัว	หลายครั้ง
ดีมาก	มหัศจรรย์	สวย	อมตะ
ดีเยี่ยม	มันส์	สักพัก	อย่างเร็ว
เด็ด	มาก	สักหน่อย	อย่างแรง
ต่อไป	มากขึ้น	สาธุ	อุ่น
ตั้งแต่แรก	มีระบบ	สารพัด	

พหุน ปณฺ ฑิโต ชีเว

คำแสดงคุณลักษณะเชิงลบ จำนวน 57 คำ (ก่อนนำคำที่มีค่าน้ำหนักน้อยว่า 2 ออก)

กระจอก	เดิม	เฟื้อย	สะดุด
เก่า	ตลก	มากเกินไป	สักที
เกินควร	ตอนแรก	โม้	หงุดหงิด
เกินไป	ต่ำ	ไม่ดี	หมดอารมณ์
เข้าด้ายเข้าเข็ม	ต่ำสุด	ไม่ได้ความ	ห่วย
เข้าไป	ติดกัน	ไม่ได้เรื่อง	ห่วยแตก
งอมแงม	ทำไม	ไม่ทัน	โหด
โง่	ที่ไหนได้	ไม่เท่าไร	อย่า
ใจดำ	ธรรมดา	ไม่มีวัน	อยู่ดีๆ
ฉิบ	น้อยลง	ยังกับ	ออกจะ
ฉิบหาย	นักหนา	แย่	อ่อน
ช่าง	นิดหนึ่ง	ล่าช้า	เอาเปรียบ
เซ็ง	บ้า	ลำบาก	
ดุเดือด	บ้าบอ	สมควร	
เด็ดขาด	เป็นบ้า	สะใจ	

พหุน ปณฺฑิโต ชีเว

คำแสดงคุณลักษณะเชิงบวก จำนวน 36 คำ (หลังนำคำที่มีค่าน้ำหนักน้อยว่า 2 ออก)

กรุณา	เต็ม	เยี่ยม	สุดท้าย
เก่ง	เต็มที่	เรียบร้อย	สุดยอด
เจ๋ง	ที่หนึ่ง	ลื่น	สุ่ม
ชัด	เปล่า	ไว	เสถียร
ใช้ได้	พอใช้	สนุก	หนัก
ดี	พอได้	สมจริง	หลายครั้ง
ดีมาก	มันส์	ส่วนตัว	
ดีเยี่ยม	มากขึ้น	สวย	
เด็ด	ยอด	สักพัก	
ต่อไป	ยอดเยี่ยม	สักหน่อย	

คำแสดงคุณลักษณะเชิงลบ จำนวน 33 คำ (หลังนำคำที่มีค่าน้ำหนักน้อยว่า 2 ออก)

กระจอก	ตอนแรก	ไม่ดี	โหด
เก่า	ต่ำ	ไม่ทัน	อย่า
เกินไป	ต่ำสุด	แย่	อยู่ดีๆ
เข้าไป	ทำไม	สะดุด	ออกจะ
โง่	ธรรมดา	สักที	อ่อน
ฉิบ	น้อยลง	หงุดหงิด	เอาเปรียบ
ฉิบหาย	นักหนา	หมดอารมณ์	
เซ็ง	นิดหนึ่ง	ห่วย	
เดิม	บ้า	ห่วยแตก	

พหุน ปณฺฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นายวสวัตติ์ อินทร์แปลง
วันเกิด	29 กุมภาพันธ์ 2535
สถานที่เกิด	โรงพยาบาลกาฬสินธุ์
สถานที่อยู่ปัจจุบัน	99 หมู่ 10 ต.สมเด็จ อ.สมเด็จ จ.กาฬสินธุ์ 46150
ตำแหน่งหน้าที่การงาน	นักวิชาการคอมพิวเตอร์
สถานที่ทำงานปัจจุบัน	องค์การบริหารส่วนจังหวัดมหาสารคาม 123 ถ.ถีนานนท์ ต.ตลาด อ.เมือง จ.มหาสารคาม
ประวัติการศึกษา	2557 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม 2563 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

พหูน ปณฺฑิโต ชีเว