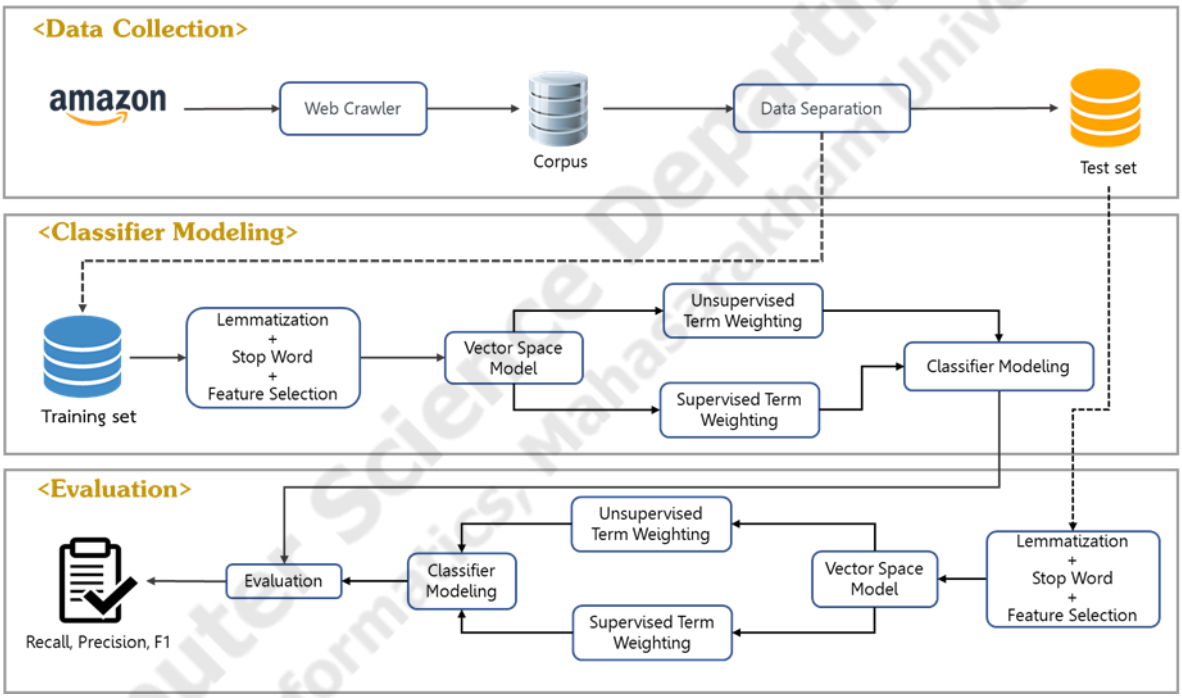
บทที่ 3

วิธีดำเนินงานวิจัย

ในบทนี้จะอธิบายถึงชุดข้อมูลข้อความแสดงความคิดเห็นที่เกี่ยวกับอุปกรณ์อิเล็กทรอนิกส์ ซึ่งรวบรวมมาจากเว็บไซต์ Amazon ที่ใช้ในโครงงานนี้ และวิธีการดำเนินงาน ดังนี้

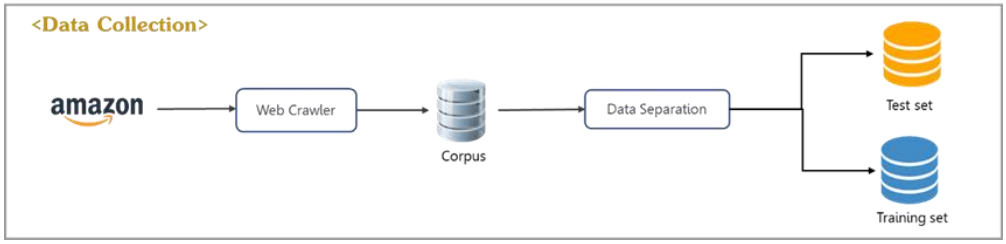
3.1 กรอบการดำเนินงาน



ภาพประกอบที่ 3.1 กรอบการดำเนินงานของระบบ

ภาพรวมของระบบการควบคุมข้อมูลไม่สมดุลในการจำแนกความรู้สึก จะแบ่งการทำงานออกเป็น 3 ส่วนหลัก คือ

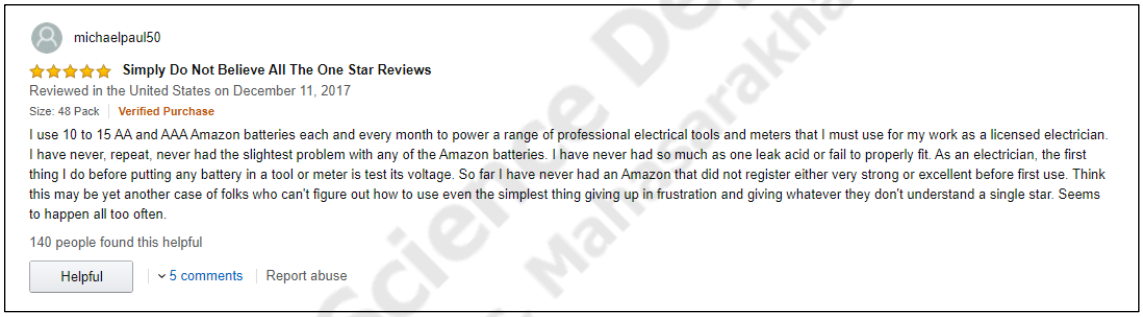
3.2 ชุดข้อมูล (Data set)



ภาพประกอบที่ 3.2 Data Collection

ในส่วนนี้ เป็นส่วนของการเก็บรวบรวมข้อมูล ในโครงงานปริญญานิพนธ์นี้เป็นข้อความแสดงความคิดเห็นที่เกี่ยวกับข้อความแสดงความคิดเห็นที่เกี่ยวกับอุปกรณ์อิเล็กทรอนิกส์ ซึ่งรวบรวมจากเว็บไซต์ Amazon โดยแบ่งเป็น 2 รูปแบบ คือ ข้อความแสดงความคิดเห็นที่เป็นเชิงบวก (Positive) และข้อความแสดงความคิดเห็นที่เป็นเชิงลบ (Negative) และในการเตรียมข้อมูล จะใช้ข้อมูล Train อย่างน้อย 1000 บทวิจารณ์ต่อกลุ่มความคิดเห็น และใช้ข้อมูลชุดทดสอบ (Test) อย่างน้อย 200 บทวิจารณ์ต่อกลุ่มความคิดเห็น

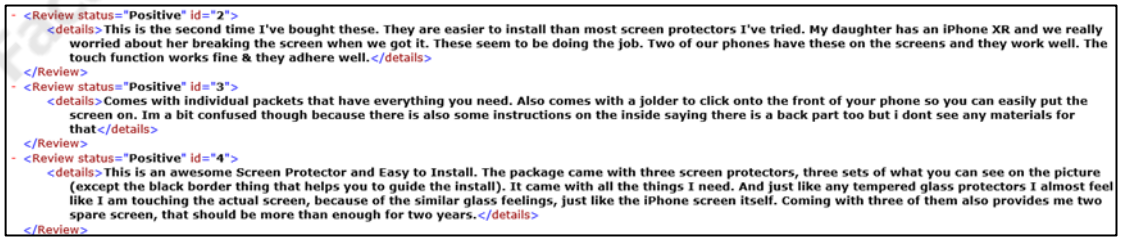
โดยในโครงงานปริญญานิพนธ์นี้ ได้ใช้ชุดข้อความแสดงความคิดเห็นที่เกี่ยวกับอุปกรณ์อิเล็กทรอนิกส์ ซึ่งรวบรวมมาจากเว็บไซต์ Amazon โดยจะมีการแบ่งเอกสารออกเป็น 2 ชุด คือ ชุดข้อมูลสอน (Training set) และ ชุดข้อมูลทดสอบ (Test set) ซึ่งเอกสารจะอยู่ในรูปแบบ XML ข้อมูลที่ใช้ทั้งหมด 50,000 ความคิดเห็นและมีคำระหว่าง 30 ถึง 300 คำต่อหนึ่งเอกสารข้อความแสดงความคิดเห็น



ภาพประกอบที่ 3.3 ตัวอย่างเอกสารข้อความแสดงความคิดเห็น

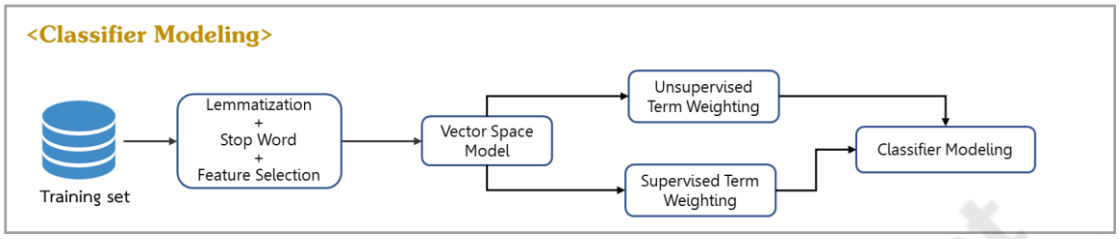
ที่มา : https://www.amazon.com/AmazonBasics-Performance-Alkaline-Batteries-Count/product-reviews/B00MNV8E0C/

จากภาพประกอบที่ 3.3 เป็นตัวอย่างเอกสารข้อความแสดงความคิดเห็นจากเว็บ Amazon ที่ใช้ในงานวิจัยนี้โดยจะทำการดาวน์โหลดออกมาในรูปแบบ XML ซึ่งจะประกอบไปด้วย รหัส (ID), สถานะ (Status) และเนื้อหาของเอกสาร (details) ดังภาพประกอบที่ 3.4



ภาพประกอบที่ 3.4 ตัวอย่างเอกสารที่อยู่ในรูปแบบ XML

3.3 การสร้างโมเดลเพื่อการจำแนกความรู้สึกของบทวิจารณ์ (Classifier Modeling)



ภาพประกอบที่ 3.5 Classifier Modeling

ในการสร้างโมเดลเพื่อจำแนกความรู้สึกของบทวิจารณ์ จะมีขั้นตอนหลักในการประมวลผลดังนี้

3.3.1 การเตรียมข้อมูลก่อนการประมวลผล

ในขั้นตอนก่อนการประมวลผล จะเป็นการเตรียมเอกสารหรือบทความให้อยู่ในรูปแบบที่พร้อมจะนำเข้าไปประมวลผลในขั้นตอนถัดไปได้ ซึ่งจะมีขั้นตอนดังนี้

สมมุติให้มีเอกสารบทวิจารณ์เกี่ยวกับอุปกรณ์อิเล็กทรอนิกส์ 6 เอกสาร ได้แก่

D_1 : One of worst electronic items.

D_2 : That's the worst electronic device ever used.

D_3 : Bad HDMI.

D_4 : So good

D_5 : This's a Good electric device!

D_6 : Best device

ขั้นตอนที่ 1 : การตัดคำและการตัดคำหยุด (Stop-word Removal) เป็นกระบวนการตัดคำหรือสัญลักษณ์ที่พบบ่อยมากในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อการจัดกลุ่มเอกสาร

ตัวอย่างเอกสารหลังจากทำการตัดคำ

D_1 : one / of / worst / electrical / items

D_2 : that / 's / the / worst / electric / device / ever / used

D_3 : bad / electric

D_4 : so / good

D_5 : good / this / 's / a / electronic / device

D_6 : best / device

ตัวอย่างเอกสารหลังจากทำการตัดคำหยุด

D_1 : worst / electrical / items

D_2 : worst / electric / device

D_3 : bad / electric

D_4 : so / good

D_5 : good / electronic / device

D_6 : best / device

ขั้นตอนที่ 2 : การทำ Lemmatization Tagging จะเป็นการเปลี่ยนคำให้อยู่ในรูปแบบดั้งเดิม โดยมีขั้นตอนดังนี้

1. TokenizerAnnotator เป็นกระบวนการตัดคำโดยใช้หลักการเดียวกันกับ Penn Treebank

D_1 : worst / electrical / items

D_2 : worst / electric / device

D_3 : bad / electric

D_4 : so / good

D_5 : good / electronic / device

D_6 : best / device

2. ssplit เป็นการนำคำที่ผ่านกระบวนการตัดคำมาเรียงลำดับตามประโยคเดิม

D_1 : worst / electrical / items

D_2 : worst / electric / device

D_3 : bad / electric

D_4 : so / good

D_5 : good / electronic / device

D_6 : best / device

3. POS (Part-Of-Speech Tagging) เป็นการติด tag ให้กับคำแต่ละคำ โดยใช้ Penn Treebank Tagset

D_1 : worst (JJS) | electronic (JJ) | items (NNS)

D_2 : worst (JJS) | electronic (JJ) | device (NN)

D_3 : bad (JJ) | electric (JJ)

D_4 : good (JJ)

D_5 : electronic (JJ) | device (NN)

D_6 : Best (RB) | device (NN)

4. Lemma เป็นการนำคำที่ได้ภายหลังการติด tag มาทำ lemma โดยใช้ Wordnet

D_1 : worst / electrical / items

D_2 : worst / electric / device

D_3 : bad / electric

D_4 : good

D_5 : good / electronic / device

D_6 : best / device

ขั้นตอนที่ 3 : การนำคำที่ได้จากขั้นตอนที่ 2 ไปเปรียบเทียบกับคำใน Dictionary หาก คำนั้นไม่มีใน Dictionary จะทำการตัดคำนั้นทิ้ง เช่น คำว่า “hdmi” ซึ่งเป็นชื่อของอุปกรณ์อิเล็กทรอนิกส์และไม่ใช่คำหยุด เป็นต้น

ขั้นตอนที่ 4 : การสร้างตัวแทนเอกสาร จะเป็นการนำเสนอความสัมพันธ์ระหว่างคำและเอกสารในรูปแบบเวกเตอร์ จากขั้นตอนที่ 3 สามารถแสดงในรูปของ BOW ได้ดังนี้

ตารางที่ 3.1 แสดงการนำเสนอความสัมพันธ์ระหว่างคำและเอกสาร

W_i	worst	electric	bad	good	best	items	device
D_1	1	1	0	0	0	1	0
D_2	1	1	0	0	0	0	1
D_3	0	1	1	0	0	0	0
D_4	0	0	0	1	0	0	0
D_5	0	1	0	1	0	0	1
D_6	0	0	0	0	1	0	1

จากตารางที่ 3.1 จะเห็นว่า BOW นอกจากจะแสดงความสัมพันธ์ระหว่างคำและเอกสารแล้ว ยังสามารถแสดงให้เห็นความถี่ของคำที่ปรากฏในเอกสารนั้นๆ อีกด้วย

ขั้นตอนที่ 5 : การเลือกคุณลักษณะด้วย Information Gain เพื่อตัดคำที่ไม่มีนัยสำคัญออก เพื่อให้โมเดลมีประสิทธิภาพและลดระยะเวลาในการประมวลผลลง

คำนวณค่า Info(D) หรือค่าเอนโทรปี (entropy) ของชุดข้อมูล (dataset: D) ที่กำลังศึกษาตามสมการที่ 13

$$Info(D) = -\sum_{i=1}^{n} P(c_i) * log_2 P(c_i) \quad (13)$$

โดย D คือ ชุดข้อมูล

$P(c_i)$ คือ ความน่าจะเป็นของแต่ละคลาสในชุดข้อมูลนั้นๆ

log คือ log ฐาน 2

$Info(D) = - [(0.5) \log_2(0.5)] - [(0.5) \log_2(0.5)]$

$= 1$

จากนั้นคำนวณค่า *Info* ของแต่ละ *sub-class* ในแต่ละ *Attribute* นั้นๆ ด้วย $info(attribute, a_i)$ ซึ่งเป็นฟังก์ชันที่ระบุปริมาณข้อมูลที่ต้องการเพื่อการจำแนก *class* ของข้อมูลโดยใช้ *attribute A* เป็นตัวตรวจสอบเพื่อแยกข้อมูลตามสมการที่ 14

$$Info(attribute, a_i) = \sum_{i=1}^{n} \frac{|a_i|}{|A|} * Info(a_i) \quad (14)$$

โดย A คือ จำนวนข้อมูลทั้งหมดใน *Attribute* ที่กำลังพิจารณา

a_i คือ *sub-class* ใน *Attribute* ที่กำลังพิจารณา

$|a_i|$ คือ จำนวนข้อมูลใน *sub-class* a_i

Info(*Allword*, *worst*) $= 2/7 \times [-(2/7 \times \log_2(2/7)) - (0/7 \times \log_2(0/7))]$

$= 0.1475$

Info(*Allword*, *electric*) $= 5/7 \times [-(3/7 \times \log_2(3/7))) - (1/7 \times \log_2(1/7))]$

$= 0.7004$

Info(*Allword*, *bad*) $= 1/7 \times [-(1/7 \times \log_2(1/7)) - (0/7 \times \log_2(0/7))]$

= 0.0572

Info(*Allword*, *good*) = 2/7 x [-(0/7 x $\log_2(0/7)$) - (2/7 x $\log_2(2/7)$)]

= 0.1475

Info(*Allword*, *best*) = 1/7 x [-(0/7 x $\log_2(0/7)$) – (1/7 x $\log_2(1/7)$)]

= 0.0572

Info(*Allword*, *items*) = 1/7 x [-(1/7 x $\log_2(1/7)$) – (0/7 x $\log_2(0/7)$)]

= 0.0572

Info(*Allword*, *device*) = 3/7 x [-(1/7 x $\log_2(1/7)$) - (2/7 x $\log_2(2/7)$)]

= 0.3931

เมื่อได้ค่า *Info* ของแต่ละคำหรือแอตทริบิวต์เรียบร้อยแล้ว ต่อไปจะเป็นการหาค่า Information Gain (*IG*) ของแต่ละคำนั้นๆ ด้วยสมการที่

$$Gain(A) = Info(D) - Info(Attribute, a_i) \quad (15)$$

โดย *Gain*(*A*) คือ ค่าความน่าเชื่อถือของคำนั้นๆ

$Gain_{worst}$ = 1.0 - 0.1475 = 0.8525

$Gain_{electric}$ = 1.0 - 0.7004 = 0.2996

$Gain_{bad}$ = 1.0 - 0.0572 = 0.9428

$Gain_{good}$ = 1.0 - 0.1475 = 0.8525

$Gain_{best}$ = 1.0 - 0.0572 = 0.9428

$Gain_{items}$ = 1.0 - 0.0572 = 0.9428

$Gain_{device}$ = 1.0 - 0.0572 = 0.9428

เมื่อคำนวณค่า *Gain* ของแต่ละคำเสร็จเรียบร้อยแล้ว จะทำการเรียงค่า *Gain* จากมากไปหาน้อยเพื่อลดจำนวนคำลง โดยจะตัดคำที่ไม่มีนัยสำคัญต่อเอกสารออกด้วยการวัดค่า *Gain* หากคำใดมีค่า *Gain* เป็น 0 จะถูกตัดทิ้งทั้งหมด จากตัวอย่างข้างต้นจะเห็นว่าไม่มีคำที่มีค่า *Gain* เป็น 0 นั่นหมายความว่า คำทุกคำในตัวอย่างมีความสำคัญต่อเอกสารทั้งหมด

ขั้นตอนที่ 6 : การให้น้ำหนักคำ (Term weighting) จะมี 2 รูปแบบหลัก ดังนี้

รูปแบบที่ 1 : Unsupervised Term Weighting (UTW) โดยในปริญญานิพนธ์ฉบับนี้ จะใช้รูปแบบที่ได้รับความนิยมมากที่สุดของ UTW คือ *tf-idf* การให้น้ำหนักแบบ *tf-idf* เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์เพื่อใช้ในการจัดกลุ่มของเอกสารให้ตรงกับหมวดหมู่ที่ถูกกำหนดไว้ โดย *tf* เป็นการหาความถี่ของคำหนึ่งๆ ที่พบในแต่ละเอกสาร และ *idf* ก็คือ global weight ที่เป็นการหาส่วนกลับของความถี่ของคำในเอกสาร หรือที่เรียกว่าระบบน้ำหนักความถี่เอกสารผกผัน โดยจะสามารถแสดงขั้นตอนได้ดังนี้

ขั้นตอนที่ 1 : หาค่า *tf* ที่เป็นความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ ว่าพบกี่ครั้ง

ขั้นตอนที่ 2 : หาค่า *idf* คือการหาค่าส่วนกลับของแต่ละคำในเอกสารนั้นๆ

การคำนวณหา *idf* ทำได้โดยใช้สมการ $idf = log(N/df)$ โดย N คือจำนวนเอกสารทั้งหมดในคลังเอกสาร และ *df* คือจำนวนเอกสารที่มีคำนั้นๆ ปรากฏอยู่ และสามารถคำนวณหาค่า *idf* ได้ดังนี้ในที่นี้จะให้ $N = 5$

$$idf_{worst} = \log(6/2) = 0.477$$
$$Idf_{electric} = \log(6/4) = 0.176$$
$$Idf_{bad} = \log(6/1) = 0.778$$
$$Idf_{good} = \log(6/2) = 0.477$$
$$Idf_{best} = \log(6/1) = 0.778$$
$$Idf_{items} = \log(6/1) = 0.778$$
$$Idf_{device} = \log(6/3) = 0.301$$

ขั้นตอนที่ 3 : การคำนวณหาค่า *tf-idf*

ในขั้นตอนนี้จะเป็นการนำเอาค่า *tf* ที่ได้คูณเข้ากับค่า *idf* เช่น ในเอกสารที่ 1 จะพบคำ 3 คำ คือ “worst” ,“electric” และ“items” โดยคำเหล่านี้ ที่ปรากฏในเอกสารที่ 1 มีค่า *tf* เป็น 1, 1 และ 1 ตามลำดับ เมื่อนำมาหาค่า *tf-idf* จะได้ผลลัพธ์ ดังต่อไปนี้

$$tf\text{-}idf_{worst} \text{ ในเอกสารที่ 1} = 1 \times 0.477 = 0.477$$
$$tf\text{-}idf_{worst} \text{ ในเอกสารที่ 2} = 1 \times 0.477 = 0.477$$
$$tf\text{-}idf_{electric} \text{ ในเอกสารที่ 1} = 1 \times 0.176 = 0.176$$

$tf\text{-}idf_{electric}$ ในเอกสารที่ 2 = 1 x 0.176 = 0.176

$tf\text{-}idf_{electric}$ ในเอกสารที่ 3 = 1 x 0.176 = 0.176

$tf\text{-}idf_{electric}$ ในเอกสารที่ 5 = 1 x 0.176 = 0.176

$tf\text{-}idf_{bad}$ ในเอกสารที่ 3 = 1 x 0.778 = 0.778

$tf\text{-}idf_{best}$ ในเอกสารที่ 6 = 1 x 0.778 = 0.778

$tf\text{-}idf_{good}$ ในเอกสารที่ 4 = 1 x 0.477 = 0.477

$tf\text{-}idf_{good}$ ในเอกสารที่ 5 = 1 x 0.477 = 0.477

$tf\text{-}idf_{items}$ ในเอกสารที่ 1 = 1 x 0.778 = 0.778

$tf\text{-}idf_{device}$ ในเอกสารที่ 2 = 1 x 0.301 = 0.301

$tf\text{-}idf_{device}$ ในเอกสารที่ 5 = 1 x 0.301 = 0.301

$tf\text{-}idf_{device}$ ในเอกสารที่ 6 = 1 x 0.301 = 0.301

ตารางที่ 3.2 *BOW* แสดงค่าและน้ำหนักค่าในแต่ละเอกสารด้วยการให้น้ำหนักแบบ *tf-idf*

W_i	worst	electric	bad	good	best	items	device
D_1	0.477	0.176	0	0	0	0.788	0
D_2	0.477	0.176	0	0	0	0	0.301
D_3	0	0.176	0.778	0	0	0	0
D_4	0	0	0	0.477	0	0	0
D_5	0	0.176	0	0.477	0	0	0.301
D_6	0	0	0	0	0.788	0	0.301

รูปแบบที่ 2 : Supervised Term Weighting (STW) ในรูปแบบนี้จะมีทั้งหมด 4 รูปแบบ

1) *Delta TF-IDF*

Delta TF-IDF ช่วยเพิ่มความสำคัญของคำที่กระจายอย่างไม่สม่ำเสมอระหว่างคลาสบวกและคลาสลบ โดยที่นี้ N_p และ N_n คือจำนวนของเอกสารในคลาสบวกและลบ ในตัวอย่างของเรามีจำนวนเอกสารที่อยู่ในคลาสบวก 1 เอกสารและคลาสลบ 4 เอกสาร ส่วน *A* และ *C* แสดงความถี่เอกสารของคำว่า t_i ในคลาสบวกและลบตามลำดับ จากตารางที่ 3.1 สามารถนำมาคำนวณน้ำหนักคำของ *Delta TF-IDF* ดังนี้

ขั้นตอนที่ 1 : หาค่า *tf* ที่เป็นความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ ว่าพบกี่ครั้ง

ขั้นตอนที่ 2 : หาค่า *Delta TF-IDF* ของแต่ละคำในเอกสารนั้นๆ ซึ่งสามารถคำนวณหาค่า *Delta TF-IDF* ของแต่ละเอกสาร ได้ดังนี้

$W_{\&.TF.IDF}$(worst) ในเอกสารที่ 1 $= 1 * log_2\left(\frac{3*0+1.5}{2*3+1.5}\right)$ = -2.321

$W_{\&.TF.IDF}$(electric) ในเอกสารที่ 1 $= 1 * log_2\left(\frac{3*1+1.5}{3*3+1.5}\right)$ = -1.222

$W_{\&.TF.IDF}$(items) ในเอกสารที่ 1 $= 1 * log_2\left(\frac{3*0+1.5}{1*3+1.5}\right)$ = -1.584

$W_{\&.TF.IDF}$(worst) ในเอกสารที่ 2 $= 1 * log_2\left(\frac{3*0+1.5}{2*3+1.5}\right)$ = -2.321

$W_{\&.TF.IDF}$(electric) ในเอกสารที่ 2 $= 1 * log_2\left(\frac{3*1+1.5}{3*3+1.5}\right)$ = -1.222

$W_{\&.TF.IDF}$(device) ในเอกสารที่ 2 $= 1 * log_2\left(\frac{3*2+1.5}{1*3+1.5}\right)$ = 0.736

$W_{\&.TF.IDF}$(electric) ในเอกสารที่ 3 $= 1 * log_2\left(\frac{3*1+1.5}{3*3+1.5}\right)$ = -1.222

$W_{\&.TF.IDF}$(bad) ในเอกสารที่ 3 $= 1 * log_2\left(\frac{3*0+1.5}{1*3+1.5}\right)$ = -1.584

$W_{\&.TF.IDF}$(good) ในเอกสารที่ 4 $= 1 * log_2\left(\frac{3*0+1.5}{2*3+1.5}\right)$ = -2.321

$W_{\&.TF.IDF}$(electric) ในเอกสารที่ 5 $= 1 * log_2\left(\frac{3*3+1.5}{1*3+1.5}\right)$ = 1.222

$W_{\&.TF.IDF}$(good) ในเอกสารที่ 5 $= 1 * log_2\left(\frac{3*0+1.5}{2*3+1.5}\right)$ = -2.321

$W_{\&.TF.IDF}$(device) ในเอกสารที่ 5 $= 1 * log_2\left(\frac{3*1+1.5}{2*3+1.5}\right)$ = -0.893

$W_{\&.TF.IDF}$(best) ในเอกสารที่ 6 $= 1 * log_2\left(\frac{3*0+1.5}{1*3+1.5}\right)$ = -1.584

$W_{\&.TF.IDF}$(device) ในเอกสารที่ 6 $= 1 * log_2\left(\frac{3*1+1.5}{2*3+1.5}\right)$ = -0.736

ตารางที่ 3.3 *BOW* แสดงค่าและน้ำหนักค่าในแต่ละเอกสารด้วยการให้น้ำหนักแบบ *Delta TF-IDF*

W_i	worst	electric	bad	good	best	items	device
D_1	-2.321	-1.222	0	0	0	-1.584	0
D_2	-2.321	-1.222	0	0	0	0	0.736
D_3	0	-1.222	-1.584	0	0	0	0
D_4	0	0	0	-2.321	0	0	0
D_5	0	1.222	0	-2.321	0	0	-0.736
D_6	0	0	0	0	-1.584	0	-0.736

2) TF-IDF-ICF

TF-IDF-ICF เป็นรูปแบบการควบคุมน้ำหนักตามแบบ TF-IDF แบบดั้งเดิม โดยเพิ่มปัจจัยความถี่ผกผันในคลาส (Inverse Class Frequency: ICF) เพื่อให้ค่าน้ำหนักคำสูงขึ้นสำหรับคำหายากที่เกิดขึ้นน้อยในเอกสารและคลาส โดย M คือจำนวนคลาสในที่นี้เท่ากับ 2 จากตารางที่ 3.1 สามารถนำมาคำนวณน้ำหนักคำของ TF-IDF-ICF ดังนี้

ขั้นตอนที่ 1 : หาค่า *tf* ที่เป็นความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ

ขั้นตอนที่ 2 : หาค่า *idf* คือการหาค่าส่วนกลับของแต่ละคำในเอกสารนั้นๆ

ขั้นตอนที่ 3 : หาค่า *icf* คือปัจจัยความถี่ผกผันในคลาสของแต่ละคำในเอกสารนั้นๆ

$ICF(worst) = 1 + \log(2/1) = 1.301$

$ICF(electric) = 1 + \log(2/2) = 1.000$

$ICF(bad) = 1 + \log(2/1) = 1.301$

$ICF(good) = 1 + \log(2/1) = 1.301$

$ICF(best) = 1 + \log(2/1) = 1.301$

$ICF(items) = 1 + \log(2/1) = 1.301$

$ICF(device) = 1 + \log(2/2) = 1.000$

ขั้นตอนที่ 4 : หาค่า TF-IDF-ICF ของแต่ละคำในเอกสารนั้นๆ

$W_{TF.ICF}(worst)$ ในเอกสารที่ 1 $= 1 \times 0.477 \times 1.301 = 0.620$

$W_{TF.ICF}(electric)$ ในเอกสารที่ 1 $= 1 \times 0.176 \times 1.000 = 0.176$

$W_{TF.ICF}(items)$ ในเอกสารที่ 1 $= 1 \times 0.778 \times 1.301 = 1.012$

$W_{TF.ICF}(worst)$ ในเอกสารที่ 2 $= 1 \times 0.477 \times 1.301 = 0.620$

$W_{TF.ICF}(electric)$ ในเอกสารที่ 2 $= 1 \times 0.176 \times 1.000 = 0.176$

$W_{TF.ICF}(device)$ ในเอกสารที่ 2 $= 1 \times 0.301 \times 1.000 = 0.301$

$W_{TF.ICF}(electric)$ ในเอกสารที่ 3 $= 1 \times 0.176 \times 1.000 = 0.176$

$W_{TF.ICF}(bad)$ ในเอกสารที่ 3 $= 1 \times 0.778 \times 1.301 = 1.012$

$W_{TF.ICF}(good)$ ในเอกสารที่ 4 $= 1 \times 0.477 \times 1.301 = 0.620$

$W_{TF.ICF}(electric)$ ในเอกสารที่ 5 $= 1 \times 0.176 \times 1.000 = 0.176$

$W_{TF.ICF}(good)$ ในเอกสารที่ 5 $= 1 \times 0.477 \times 1.301 = 0.620$

$W_{TF.ICF}(device)$ ในเอกสารที่ 5 $= 1 \times 0.301 \times 1.000 = 0.301$

$W_{TF.ICF}(best)$ ในเอกสารที่ 6 $= 1 \times 0.778 \times 1.301 = 1.012$

$W_{TF.ICF}(device)$ ในเอกสารที่ 6 $= 1 \times 0.301 \times 1.000 = 0.301$

ตารางที่ 3.4 *BOW* แสดงค่าและน้ำหนักค่าในแต่ละเอกสารด้วยการให้น้ำหนักแบบ *TF-IDF-ICF*

W_i	worst	electric	bad	good	best	items	device
D_1	0.620	0.176	0	0	0	1.012	0
D_2	0.620	0.176	0	0	0	0	0.301
D_3	0	0.176	1.012	0	0	0	0
D_4	0	0	0	0.620	0	0	0
D_5	0	0.176	0	0.620	0	0	0.301
D_6	0	0	0	0	1.012	0	0.301

3) TF-RF

TF-RF มีความเกี่ยวข้องของความถี่ (RF) ของข้อกำหนด TF-RF โดยที่ตัวหารน้อยที่สุดคือ 1 เพื่อหลีกเลี่ยงการหารด้วยศูนย์ จากตารางที่ 3.1 สามารถนำมาคำนวณน้ำหนักคำของ TF-IDF-ICF ดังนี้

ขั้นตอนที่ 1 : หาค่า *tf* ที่เป็นความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ

ขั้นตอนที่ 2 : หาค่า *TF-RF* คือการหาค่าส่วนกลับของแต่ละคำในเอกสารนั้นๆ

$W_{TF.RF}(worst)$ ในเอกสารที่ 1 $= 1 * log_2\left(2 + \frac{2}{max\,(1,0)}\right) = 2.000$

$W_{TF.RF}(electric)$ ในเอกสารที่ 1 $= 1 * log_2\left(2 + \frac{3}{max\,(1,1)}\right) = 2.321$

$W_{TF.RF}(items)$ ในเอกสารที่ 1 $= 1 * log_2\left(2 + \frac{1}{max\,(1,0)}\right) = 1.584$

$W_{TF.RF}(worst)$ ในเอกสารที่ 2 $= 1 * log_2\left(2 + \frac{2}{max\,(1,0)}\right) = 2.000$

$W_{TF.RF}(electric)$ ในเอกสารที่ 2 $= 1 * log_2\left(2 + \frac{3}{max\,(1,1)}\right) = 2.321$

$W_{TF.RF}(device)$ ในเอกสารที่ 2 $= 1 * log_2\left(2 + \frac{1}{max\,(1,2)}\right) = 1.584$

$W_{TF.RF}(electric)$ ในเอกสารที่ 3 $= 1 * log_2\left(2 + \frac{3}{max\,(1,1)}\right) = 2.321$

$W_{TF.RF}(bad)$ ในเอกสารที่ 3 $= 1 * log_2\left(2 + \frac{1}{max\,(1,0)}\right) = 1.584$

$W_{TF.RF}(good)$ ในเอกสารที่ 4 $= 1 * log_2\left(2 + \frac{2}{max\,(1,0)}\right) = 2.000$

$W_{TF.RF}(electric)$ ในเอกสารที่ 5 $= 1 * log_2\left(2 + \frac{1}{max\,(1,3)}\right) = 1.736$

$W_{TF.RF}(good)$ ในเอกสารที่ 5 $= 1 * log_2\left(2 + \frac{2}{max\,(1,0)}\right) = 2.000$

$W_{TF.RF}(device)$ ในเอกสารที่ 5 $= 1 * log_2\left(2 + \frac{2}{max\,(1,1)}\right) = 2.000$

$W_{TF.RF}(best)$ ในเอกสารที่ 6 $= 1 * log_2\left(2 + \frac{1}{max\,(1,0)}\right) = 1.584$

$W_{TF.RF}$*(device)* ในเอกสารที่ 6 $= 1 * log_2\left(2 + \frac{2}{max\ (1,1)}\right)$ = 2.000

ตารางที่ 3.5 *BOW* แสดงค่าและน้ำหนักค่าในแต่ละเอกสารด้วยการให้น้ำหนักแบบ *TF-RF*

W_i	worst	electric	bad	good	best	items	device
D_1	2.000	2.321	0	0	0	1.584	0
D_2	2.000	2.321	0	0	0	0	1.584
D_3	0	2.321	1.584	0	0	0	0
D_4	0	0	0	2.000	0	0	0
D_5	0	1.736	0	2.000	0	0	2.000
D_6	0	0	0	0	1.584	0	2.000

4) TF-IGM

ระยะความถี่-ช่วงเวลาแรงโน้มถ่วงผกผัน (Term Frequency - Inverse Gravity Moment : *TF-IGM*) [20] ถูกนำเสนอให้วัดความไม่สม่ำเสมอหรือความเข้มข้นของการแจกแจงคำศัพท์ระหว่างคลาส โดยสามารถนำมาคำนวณได้ดังนี้

ขั้นตอนที่ 1 : หาค่า *tf* ที่เป็นความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ

ขั้นตอนที่ 2 : หาค่า *igm* ที่เป็นความถี่ของคำแต่ละคำที่กระจายอยู่ในแต่ละคลาส

igm(worst) $= \frac{2}{(1\ x\ 2)+(2\ x\ 0)}$ = 1.0

igm(electric) $= \frac{3}{(1\ x\ 3)+(2\ x\ 1)}$ = 0.6

igm(bad) $= \frac{1}{(1\ x\ 1)+(2\ x\ 0)}$ = 1.0

igm(good) $= \frac{2}{(1\ x\ 2)+(2\ x\ 0)}$ = 1.0

igm(best) $= \frac{1}{(1\ x\ 1)+(2\ x\ 0)}$ = 1.0

igm(items) $= \frac{1}{(1\ x\ 1)+(2\ x\ 0)}$ = 1.0

igm(device) $= \frac{2}{(1\ x\ 2)+(2\ x\ 1)}$ = 0.5

ขั้นตอนที่ 4 : หาค่า *TF-IGM* ของแต่ละคำในเอกสารนั้นๆ

IGM(worst) ในเอกสารที่ 1 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(electric) ในเอกสารที่ 1 = 1 x (1 + 7.0 x 0.6) = 5.2

IGM(items) ในเอกสารที่ 1 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(worst) ในเอกสารที่ 2 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(electric) ในเอกสารที่ 2 = 1 x (1 + 7.0 x 0.6) = 5.2

IGM(device) ในเอกสารที่ 2 = 1 x (1 + 7.0 x 0.5) = 3.5

IGM(electric) ในเอกสารที่ 3 = 1 x (1 + 7.0 x 0.6) = 5.2

IGM(bad) ในเอกสารที่ 3 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(good) ในเอกสารที่ 4 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(electric) ในเอกสารที่ 5 = 1 x (1 + 7.0 x 0.6) = 5.2

IGM(good) ในเอกสารที่ 5 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(device) ในเอกสารที่ 5 = 1 x (1 + 7.0 x 0.5) = 3.5

IGM(best) ในเอกสารที่ 6 = 1 x (1 + 7.0 x 1.0) = 8.0

IGM(device) ในเอกสารที่ 6 = 1 x (1 + 7.0 x 0.5) = 3.5

ตารางที่ 3.6 *BOW* แสดงค่าและน้ำหนักค่าในแต่ละเอกสารด้วยการให้น้ำหนักแบบ *TF-IGM*

W_i	worst	electric	bad	good	best	items	device
D_1	8.0	5.2	0	0	0	8.0	0
D_2	8.0	5.2	0	0	0	0	3.5
D_3	0	5.2	8.0	0	0	0	0
D_4	0	0	0	8.0	0	0	0
D_5	0	5.2	0	8.0	0	0	3.5
D_6	0	0	0	0	8.0	0	3.5

3.3.2 การสร้างโมเดลการจำแนกความรู้สึกของบทวิจารณ์

โครงงานปริญญานิพนธ์นี้ ได้นำเสนออัลกอริทึมสำหรับการจำแนกความรู้สึกของบทวิจารณ์ทั้งหมด 2 อัลกอริทึม นั่นคือ อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด (K-nearest Neighbor) เนื่องจากอัลกอริทึมที่เลือกใช้นี้เป็นอัลกอริทึมที่มีประสิทธิภาพดีในการจัดกลุ่มเอกสารข้อความ โดยในการสร้างโมเดลด้วย Naïve Bayes และ K-nearest Neighbor จะมีการนำเอา BOW ที่ได้จากการคำนวณน้ำหนักคำของแต่ละรูปแบบ มาทำการสร้างโมเดล ดังตัวอย่างต่อไปนี้

1) การจำแนกระดับบทวิจารณ์ด้วยอัลกอริทึม Naïve Bayes

การจำแนกบทวิจารณ์ด้วยอัลกอริทึม *Naïve Bayes* เป็นการนำเทคนิคการวิเคราะห์ความรู้สึกจากข้อความ และการจำแนกหมวดหมู่เอกสารมาประยุกต์ใช้ในการจำแนกบทวิจารณ์สินค้าอิเล็กทรอนิกส์แบบ 2 กลุ่ม

$$P(v_j \mid a_1, a_2, \dots, a_n) = \prod_{i=1}^{n} P(a_i \mid v_j) \quad (16)$$

จากเอกสารทั้งหมด 6 เอกสาร จะมีเอกสารที่เป็นเอกสารที่มีความรู้สึกเป็นบวก (Positive) จำนวน 3 เอกสาร และเอกสารที่มีความรู้สึกเป็นลบ (Negative) จำนวน 3 เอกสาร ดังนั้นจะหาความน่าจะเป็นของคำสำคัญที่อยู่ในแต่ละเอกสารที่แยกคลาสออกจากกันจะได้ความน่าจะเป็น ดังสมการที่ 17

$$P(a_i \mid v_j) = \frac{count(a_i, v_j)}{count(v_i)} \quad (17)$$

โดยที่ $count\ (a_i, v_j)$ คือค่าความถี่ของคำที่ *i* ที่อยู่ในกลุ่มที่ *j*

$count\ (v_i)$ คือค่าความถี่รวมในกลุ่มที่ *j*

แต่ในบางครั้งที่หาความน่าจะเป็นโดยใช้ *Naïve Bayes* นั้นอาจจะมีกรณีที่ค่าความถี่ของคำที่เกิดขึ้นเป็น 0 หรือก็คือคำที่อยู่ในถุงคำ ไม่ปรากฏอยู่ในเอกสารทำให้ค่าความน่าจะเป็นของคำนั้นเป็น 0 ตามไปด้วย ซึ่งไม่เป็นที่ยอมรับในทางสถิติที่โอกาสในการพยากรณ์จะมีค่าเป็นศูนย์ และเพื่อหลีกเลี่ยงการเกิดกรณีนี้จึงมีการปรับสมการด้วย *Laplace Smoothing* ที่มีการเพิ่มค่าความถี่ของข้อมูลเข้าไปอีกครั้งละ 1 และบวกเพิ่มค่าความถี่รวมด้วยค่าคงที่ *k* (อาจใช้ค่าขนาดของ *BOW*) จากคำทั้งหมด *n* คำ และกลุ่มทั้งหมด *m* กลุ่ม ดังนั้นจึงได้สมการ *Naïve Bayes* ที่ปรับแล้วดังนี้

$$P(a_i \mid v_j) = \frac{1 + count(a_i, v_j)}{k + count(v_i)} \quad (18)$$

โดยที่ $count\ (a_i, v_j)$ คือ ค่าความถี่ของคำที่ i ที่อยู่ในกลุ่มที่ j

$count\ (v_i)$ คือ ค่าความถี่รวมในกลุ่มที่ *j*

k คือ ค่าคงที่ที่มีการนำมาบวกเข้า

i มีค่าเท่ากับ *1, 2, 3..., n*

j มีค่าเท่ากับ *1, 2, 3..., m*

ในโครงงานปริญญานิพนธ์นี้จะใช้สมการ *Naïve Bayes* ที่มีการปรับสมการ มาใช้ในการประมาณค่าความน่าจะเป็น โดยจะใช้ในการประมาณค่าความน่าจะเป็นของกลุ่ม และประมาณค่าความน่าจะเป็นของคำที่อยู่ในกลุ่ม โดยใช้ค่าน้ำหนักของคำดังที่ได้แสดงไว้ในขั้นตอนการนำเสนอเอกสาร ดังตัวอย่างต่อไปนี้

$$P(Class_i) = \frac{1 + count(doc, Class_j)}{NumClass + count(Class_i)} \quad (19)$$

โดยที่ *count* (*doc, Class*$_j$) คือ จำนวนของเอกสารที่อยู่ในคลาส *j*

count (*Class*$_i$) คือ จำนวนของเอกสารทั้งหมด

NumClass คือ จำนวน *class* ที่ใช้ในการสร้างโมเดล

$$P(w_i|Class_j) = \frac{1 + count(w_i \mid Class_j)}{TotalWord + count(Class_j)} \quad (20)$$

โดยที่ *count* (*w*$_i$ | *Class*$_j$) คือ ความถี่ของคำ *i* ที่อยู่ในกลุ่มที่ *j*

count (*Class*$_j$) คือ ความถี่รวมของคำทุกคำที่อยู่ในกลุ่มที่ *j*

TotalWord คือ จำนวนคำทั้งหมด

โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย *Naïve Bayes* โดยใช้การให้น้ำหนักคำแบบ *tf-idf*

Class = “*Positive*”

P(Positive)	= (1+3)/(2+6)	= 0.5
P(worst | Positive)	= (1+0.0)/(7+2.52)	= 0.1050
P(electric | Positive)	= (1+0.176)/(7+2.52)	= 0.1235
P(bad | Positive)	= (1+0.0)/(7+2.52)	= 0.1050
P(good | Positive)	= (1+0.954)/(7+2.52)	= 0.2052
P(best | Positive)	= (1+0.788)/(7+2.52)	= 0.1878

P(items | Positive) = (1+0.0)/(7+2.52) = 0.1050

P(device | Positive) = (1+0.602)/(7+2.52) = 0.1682

Class = “Negative”

P(Negative) = (1+3)/(2+6) = 0.5

P(worst | Negative) = (1+0.954)/(7+2.395) = 0.2079

P(electric | Negative) = (1+0.352)/(7+2.395) = 0.1439

P(bad | Negative) = (1+0.788)/(7+2.395) = 0.1903

P(best | Negative) = (1+0.0)/(7+2.395) = 0.1064

P(good | Negative) = (1+0.0)/(7+2.395) = 0.1064

P(items | Negative) = (1+0.788)/(7+2.395) = 0.1903

P(device | Negative) = (1+0.301)/(7+2.395) = 0.1384

ตารางที่ 3.7 โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย Naïve Bayes โดยใช้การให้น้ำหนักคำแบบ *tf-idf*

W_i	worst	electric	bad	good	best	items	device	
D_1	0.2079	0.1439	0.1903	0.1064	0.1064	0.1903	0.1384	Negative
D_2	0.2079	0.1439	0.1903	0.1064	0.1064	0.1903	0.1384	
D_3	0.2079	0.1439	0.1903	0.1064	0.1064	0.1903	0.1384	
D_4	0.1050	0.1235	0.1050	0.2052	0.1878	0.1050	0.1682	Positive
D_5	0.1050	0.1235	0.1050	0.2052	0.1878	0.1050	0.1682	
D_6	0.1050	0.1235	0.1050	0.2052	0.1878	0.1050	0.1682	

โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย *Naïve Bayes* โดยใช้การให้น้ำหนักคำแบบ *Delta TF-IDF*

Class = “*Positive*”

P(worst | Positive) = (1+0.0)/(7+-10.553) = -0.2814

P(electric | Positive) = (1+1.440)/(7+-10.553) = -0.6867

P(bad | Positive) = (1+0.0)/(7+-10.553) = -0.2814

P(good | Positive) = (1+-7.4)/(7+-10.553) = 1.8012

P(best | Positive) = (1+-2.807)/(7+-10.553) = 0.5085

P(items | Positive) = (1+0.0)/(7+-10.553) = -0.2814

P(device | Positive) = (1+-1.786)/(7+-10.553) = 0.2212

Class = “Negative”

P(worst | Negative) = (1+-7.4)/(7+-16.441) = 0.6778

P(electric | Negative) = (1+-4.32)/(7+-16.441) = 0.3516

P(bad | Negative) = (1+-2.807)/(7+-16.441) = 0.1913

P(best | Negative) = (1+0.0)/(7+-16.441) = -0.1059

P(good | Negative) = (1+0.0)/(7+-16.441) = -0.1059

P(items | Negative) = (1+-2.807)/(7+-16.441) = 0.1913

P(device | Negative) = (1+0.893)/(7+-16.441) = -0.2005

ตารางที่ 3.8 โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย Naïve Bayes โดยใช้การให้น้ำหนักคำแบบ *Delta TF-IDF*

W_i	worst	electric	bad	good	best	items	device	
D_1	0.6778	0.3516	0.1913	-0.1059	-0.1059	0.1913	-0.2005	Negative
D_2	0.6778	0.3516	0.1913	-0.1059	-0.1059	0.1913	-0.2005	
D_3	0.6778	0.3516	0.1913	-0.1059	-0.1059	0.1913	-0.2005	
D_4	-0.2814	-0.6867	-0.2814	1.8012	0.5085	-0.2814	0.2212	Positive
D_5	-0.2814	-0.6867	-0.2814	1.8012	0.5085	-0.2814	0.2212	
D_6	-0.2814	-0.6867	-0.2814	1.8012	0.5085	-0.2814	0.2212	

โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย *Naïve Bayes* โดยใช้การให้น้ำหนักคำแบบ *TF-IDF-ICF*

Class = “*Positive*”

P(worst | Positive) = (1+0.0) / (7+ 3.029) = 0.0997

P(electric | Positive) = (1+0.176) / (7+ 3.029) = 0.1172

P(bad | Positive) = (1+0.0) / (7+ 3.029) = 0.0997

P(good | Positive) = (1+1.240) / (7+ 3.029) = 0.2233

P(best | Positive) = (1+1.012) / (7+ 3.029) = 0.2006

P(items | Positive) = (1+0.0) / (7+ 3.029) = 0.0997

P(device | Positive) = (1+0.602) / (7+ 3.029) = 0.1597

Class = “Negative”

P(worst | Negative) = (1+1.240) / (7+4.093) = 0.2019

P(electric | Negative) = (1+0.528) / (7+4.093) = 0.1377

P(bad | Negative) = (1+1.012) / (7+4.093) = 0.1813

P(best | Negative) = (1+0.0) / (7+4.093) = 0.0901

P(good | Negative) = (1+0.0) / (7+4.093) = 0.0901

P(items | Negative) = (1+1.012) / (7+4.093) = 0.1813

P(device | Negative) = (1+0.301) / (7+4.093) = 0.1172

ตารางที่ 3.9 โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย Naïve Bayes โดยใช้การให้น้ำหนักคำแบบ *TF-IDF-ICF*

W_i	worst	electric	bad	good	best	items	device	
D_1	0.2019	0.1377	0.1813	0.0901	0.0901	0.1813	0.1172	Negative
D_2	0.2019	0.1377	0.1813	0.0901	0.0901	0.1813	0.1172	
D_3	0.2019	0.1377	0.1813	0.0901	0.0901	0.1813	0.1172	
D_4	0.0997	0.1172	0.0997	0.2233	0.2006	0.0997	0.1597	Positive
D_5	0.0997	0.1172	0.0997	0.2233	0.2006	0.0997	0.1597	
D_6	0.0997	0.1172	0.0997	0.2233	0.2006	0.0997	0.1597	

โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย *Naïve Bayes* โดยใช้การให้น้ำหนักคำแบบ *TF-RF*

Class = “*Positive*”

P(worst | Positive) = (1+0.0) / (7+11.32) = 0.0545

P(electric | Positive) = (1+1.736) / (7+11.32) = 0.1493

P(bad | Positive) = (1+0.0) / (7+11.32) = 0.0545

P(good | Positive) = (1+4.0) / (7+11.32) = 0.2729

P(best | Positive) = (1+1.584) / (7+11.32) = 0.1410

P(items | Positive) = (1+0.0) / (7+11.32) = 0.0545

P(device | Positive) = (1+4.0) / (7+11.32) = 0.2729

Class = “Negative”

P(worst | Negative) = (1+4.0) / (7+15.715) = 0.2201

P(electric | Negative) = (1+6.963) / (7+15.715) = 0.3505

P(bad | Negative) = (1+1.584) / (7+15.715) = 0.1137

P(best | Negative) = (1+0.0) / (7+15.715) = 0.0440

P(good | Negative) = (1+0.0) / (7+15.715) = 0.0440

P(items | Negative) = (1+1.584) / (7+15.715) = 0.1137

P(device | Negative) = (1+1.584) / (7+15.715) = 0.1137

ตารางที่ 3.10 โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย Naïve Bayes โดยใช้การให้น้ำหนักคำแบบ *TF-RF*

W_i	worst	electric	bad	good	best	items	device	
D_1	0.2201	0.3505	0.1137	0.0440	0.0440	0.1137	0.1137	Negative
D_2	0.2201	0.3505	0.1137	0.0440	0.0440	0.1137	0.1137	
D_3	0.2201	0.3505	0.1137	0.0440	0.0440	0.1137	0.1137	
D_4	0.0545	0.1493	0.0545	0.2729	0.1410	0.0545	0.2729	Positive
D_5	0.0545	0.1493	0.0545	0.2729	0.1410	0.0545	0.2729	
D_6	0.0545	0.1493	0.0545	0.2729	0.1410	0.0545	0.2729	

โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย *Naïve Bayes* โดยใช้การให้น้ำหนักคำแบบ *TF-IGM*

Class = “*Positive*”

P(worst | Positive) = (1+0.0)/(7+36.2) = 0.0231

P(electric | Positive) = (1+5.2)/(7+36.2) = 0.1435

P(bad | Positive) = (1+0.0)/(7+36.2) = 0.0231

P(good | Positive) = (1+16)/(7+36.2) = 0.3935

P(best | Positive) = (1+8.0)/(7+36.2) = 0.2083

P(items | Positive) = (1+0.0)/(7+36.2) = 0.0231

P(device | Positive) = (1+7.0)/(7+36.2) = 0.1851

Class = “Negative”

P(worst | Negative) = (1+16.0)/(7+51.1) = 0.2925

P(electric | Negative) = (1+15.6)/(7+51.1) = 0.2857

P(bad | Negative) = (1+8.0)/(7+51.1) = 0.1549

P(best | Negative) = (1+0.0)/(7+51.1) = 0.0172

P(good | Negative) = (1+0.0)/(7+51.1) = 0.0172

P(items | Negative) = (1+8.0)/(7+51.1) = 0.1549

P(device | Negative) = (1+3.5)/(7+51.1) = 0.0774

ตารางที่ 3.11 โมเดลการจำแนกความรู้สึกของบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย Naïve Bayes โดยใช้การให้น้ำหนักคำแบบ *TF-IGM*

W_i	worst	electric	bad	good	best	items	device	
D_1	0.2925	0.2857	0.1549	0.0172	0.0172	0.1549	0.0774	Negative
D_2	0.2925	0.2857	0.1549	0.0172	0.0172	0.1549	0.0774	
D_3	0.2925	0.2857	0.1549	0.0172	0.0172	0.1549	0.0774	
D_4	0.0231	0.1435	0.0231	0.3935	0.2083	0.0231	0.1851	Positive
D_5	0.0231	0.1435	0.0231	0.3935	0.2083	0.0231	0.1851	
D_6	0.0231	0.1435	0.0231	0.3935	0.2083	0.0231	0.1851	

2) การจำแนกบทวิจารณ์ด้วย K-nearest Neighbor (KNN)

KNN เป็นอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลที่ไม่ซับซ้อนเข้าใจง่าย ซึ่งวิธีนี้จะสามารถสร้างโมเดลที่มีประสิทธิภาพได้แม้เงื่อนไขที่ใช้ในการตัดสินใจจะมีความซับซ้อนก็ตาม โดยจะใช้หลักการเปรียบเทียบข้อมูลที่สนใจ (x) กับข้อมูลที่ถูกจัดกลุ่มไว้ก่อนล่วงหน้าในคลังข้อมูล เพื่อตรวจสอบว่าข้อมูล x นั้นคล้ายคลึงกับกลุ่มใด และถ้าหากข้อมูล x คล้ายคลึงกับกลุ่มใดมากที่สุด ระบบก็จะจัดข้อมูลให้ข้อมูล x เข้าไปอยู่ในกลุ่มนั้น แต่ในการตัดสินใจว่า x จะคล้ายกับข้อมูลในกลุ่มใดในคลังข้อมูล

จะขึ้นอยู่กับการกำหนดค่า k (ค่า k คือการเอาข้อมูลจำนวน k ตัวที่อยู่ใกล้ x มากที่สุดมาพิจารณา) เช่น ในการจำแนกระดับคะแนนบทวิจารณ์มีข้อมูลอยู่ 5 กลุ่ม และกำหนด k=5 ภายหลังจากการประมวลผล พบว่า ข้อมูล 5 อันดับแรกที่อยู่ใกล้ x มากที่สุดนั้น มาจากกลุ่มที่ 2 จำนวน 3 ตัว และมาจากกลุ่มที่ 1 จำนวน 2 ตัว ระบบก็จะพิจารณาข้อมูล x ให้อยู่กลุ่มที่ 2

สมมุติให้มีเอกสารบทวิจารณ์เกี่ยวกับสินค้าอิเล็กทรอนิกส์ทั้งหมด 5 เอกสาร คือ

D_1 : One of worst electrical items.

D_2 : That's the worst electric device ever used.

D_3 : Bad HDMI.

D_4 : So Bad.

D_5 : This's a Good electronic device!

ตารางที่ 3.12 โมเดลวิเคราะห์ระดับคะแนนบทวิจารณ์ด้วย *KNN* โดยการให้น้ำหนักคำด้วย *tf-idf*

W_i	worst	electric	bad	good	best	items	device
D_1	0.477	0.176	0	0	0	0.788	0
D_2	0.477	0.176	0	0	0	0	0.301
D_3	0	0.176	0.778	0	0	0	0
D_4	0	0	0	0.477	0	0	0
D_5	0	0.176	0	0.477	0	0	0.301
D_6	0	0	0	0	0.788	0	0.301

ตารางที่ 3.13 โมเดลวิเคราะห์ระดับคะแนนบทวิจารณ์ด้วย *KNN* การให้น้ำหนักคำด้วย *Delta TF-IDF*

W_i	worst	electric	bad	good	best	items	device
D_1	-2.321	-1.222	0	0	0	-1.584	0
D_2	-2.321	-1.222	0	0	0	0	0.736
D_3	0	-1.222	-1.584	0	0	0	0
D_4	0	0	0	-2.321	0	0	0
D_5	0	1.222	0	-2.321	0	0	-0.736
D_6	0	0	0	0	-1.584	0	-0.736

ตารางที่ 3.14 โมเดลวิเคราะห์ระดับคะแนนบทวิจารณ์ด้วย *KNN* โดยการให้น้ำหนักคำด้วย *TF-IDF-ICF*

W_i	worst	electric	bad	good	best	items	device
D_1	0.620	0.176	0	0	0	1.012	0
D_2	0.620	0.176	0	0	0	0	0.301
D_3	0	0.176	1.012	0	0	0	0
D_4	0	0	0	0.620	0	0	0
D_5	0	0.176	0	0.620	0	0	0.301
D_6	0	0	0	0	1.012	0	0.301

ตารางที่ 3.15 โมเดลวิเคราะห์ระดับคะแนนบทวิจารณ์ด้วย *KNN* โดยการให้น้ำหนักคำด้วย *TF-RF*

W_i	worst	electric	bad	good	best	items	device
D_1	2.000	2.321	0	0	0	1.584	0
D_2	2.000	2.321	0	0	0	0	1.584
D_3	0	2.321	1.584	0	0	0	0
D_4	0	0	0	2.000	0	0	0
D_5	0	1.736	0	2.000	0	0	2.000
D_6	0	0	0	0	1.584	0	2.000

ตารางที่ 3.16 โมเดลวิเคราะห์ระดับคะแนนบทวิจารณ์ด้วย *KNN* โดยการให้น้ำหนักคำด้วย *TF-IGM*

W_i	worst	electric	bad	good	best	items	device
D_1	8.0	5.2	0	0	0	8.0	0
D_2	8.0	5.2	0	0	0	0	3.5
D_3	0	5.2	8.0	0	0	0	0
D_4	0	0	0	8.0	0	0	0
D_5	0	5.2	0	8.0	0	0	3.5
D_6	0	0	0	0	8.0	0	3.5

จากตารางที่ 3.12 ถึง ตารางที่ 3.16 จะเห็นได้ว่าเอกสารตัวอย่าง เมื่อผ่านกระบวนการ pre-processing ที่ได้น่าเสนอไปนั้น ก็จะได้เอกสารซึ่งเป็นข้อมูลที่ถูกคัดเลือกไว้ให้เป็น

ตัวแทนของแต่ละกลุ่ม และจะถูกน่าไปใช้เปรียบเทียบกับข้อมูลที่เข้ามาใหม่ต่อไป โดยขั้นตอนของ KNN มีดังนี้

ขั้นตอนที่ 1 : การก่าหนดค่า k

ขั้นตอนการกำหนดค่า k เป็นการกำหนดค่าเพื่อใช้เป็นเป้าหมายในการเลือกค่าที่ใกล้เคียงกับข้อมูลที่สนใจ โดยค่า k ที่ก่าหนดต้องเป็นเลขคี่ เพื่อให้โปรแกรมสามารถใช้ตัดสินใจได้ว่า x ควรจะถูกจัดอยู่ในกลุ่มใด ในโครงงานปริญญานิพนธ์นี้กำหนดให้ค่า k=3, k=5 และ k=7

ขั้นตอนที่ 2 : คำนวณหาระยะทางระหว่าง x กับข้อมูลทุกตัวในคลังข้อมูล

การคำนวณค่าระยะทางระหว่างข้อมูลที่สนใจ กับข้อมูลทุกตัวในคลังข้อมูลจะใช้การคำนวณระยะทางด้วย Euclidian distance เนื่องจากง่ายต่อความเข้าใจ และลักษณะการค่านวณที่คล้ายกับทฤษฎีบทพีทาโกรัส ซึ่งค่านวณได้ตามสมการดังต่อไปนี้

$$\sqrt{\sum_{i=0}^{r}[x_i - y_i]} \qquad (21)$$

โดยที่ E คือ ระยะทางระหว่างข้อมูลที่สนใจ x กับข้อมูลในคลัง y

x_i คือ คุณลักษณะที่ *i* ของข้อมูลที่สนใจ *x*

y_i คือ คุณลักษณะที่ *i* ของข้อมูลที่ถูกเลือกไว้ในคลังข้อมูล *y*

ซึ่งข้อมูลที่สนใจ *x* จะถูกเปรียบเทียบกับข้อมูลในคลังข้อมูล *y* ทั้งหมด

ขั้นตอนที่ 3 : จัดเรียงล่าดับของระยะทาง

เมื่อวัดระระทางระหว่างข้อมูลที่สนใจ x กับข้อมูลในคลังข้อมูลเสร็จเรียบร้อย จะมีการน่าระยะทางที่วัดได้มาเรียงลำดับจากระยะทางที่น้อยที่สุดไปหามากที่สุด

ขั้นตอนที่ 4 : พิจารณาข้อมูลที่ใกล้ที่สุด k ตัว

เมื่อท่าการจัดเรียงลำดับของระยะทางแล้วจะเลือกค่าระยะทางที่น้อยที่สุดจำนวน k ตัวมาพิจารณาหาค่าตอบ เช่น ถ้าหากค่า k=5 ก็จะเลือกข้อมูลจากล่าดับที่ 1 ถึง 5 มาพิจารณา

ขั้นตอนที่ 5 : ก่าหนด Class ให้กับข้อมูล x

การก่าหนด Class ให้กับข้อมูล x จะท่าโดยการพิจารณาว่าข้อมูลจ่านวน 5 ตัวที่อยู่ใกล้ x มากที่สุดอยู่กลุ่มใดบ้าง เช่น ถ้าข้อมูลในกลุ่มที่ 4 มีจ่านวน 3 ตัว อยู่ในกลุ่ม 5 จ่านวน 1 ตัว และกลุ่มที่ 2 จ่านวน 1 ตัว ระบบจึงตัดสินใจให้ข้อมูล x อยู่ในกลุ่มที่ 4

อย่างไรก็ตาม มีข้อสังเกตว่าถ้าเลือกค่า k น้อยเกินไปอาจจะท่าให้ไวต่อสัญญาณรบกวนได้ และถ้าหากเลือกค่า k มากเกินไปอาจจะท่าให้มีกลุ่มข้อมูลอื่นๆ มาปะปนกับข้อมูลที่ก่าลังสนใจได้เช่นกัน ดังนั้นวิธีการนี้จึงมีทั้งข้อดีและข้อเสียซึ่ง ข้อดีคือเป็นวิธีการที่ง่ายและให้ประสิทธิภาพความถูกต้องสูง แต่ข้อเสียคือเวลาที่ใช้ในการประมวลผลค่อนข้างนาน เพราะการท่านายข้อมูลที่เข้ามาใหม่จะอาศัยการเปรียบเทียบข้อมูลใหม่กับข้อมูลเรียนรู้จ่านวน k ตัวที่อยู่ใกล้ที่สุด

3.4 การวัดประสิทธิภาพของตัวจัดกลุ่มเอกสาร (Evaluation)

เป็นส่วนของการประเมินประสิทธิภาพแบบจ่าลองเพื่อการจ่าแนกความรู้สึกที่สร้างขึ้นภายใต้การให้น้่าหนักที่แตกต่างกัน โดยจะประเมินผลลัพธ์ของการจ่าแนกความรู้สึกด้วยเทคนิคการวัดค่าความระลึก (Recall), การวัดความแม่นย่า (Precision) และ การวัดค่าเอฟ (F-measure หรือ F1) โดยจะมีขั้นตอนดังนี้

3.4.1 การนำโมเดลเพื่อการจำแนกกลุ่มของบทวิจารณ์ไปใช้

เป็นขั้นตอนของการน่าเอาโมเดลการจ่าแนกบทวิจารณ์สินค้าอิเล็กทรอนิกส์มาใช้ในการวิเคราะห์ว่าบทวิจารณ์ที่ผู้ซื้อได้เขียนเกี่ยวกับสินค้าอิเล็กทรอนิกส์นั้นๆ ควรจัดอยู่ในกลุ่มใด โดยจะมีการรับ “ข้อความ” เข้ามา แล้วโมเดลจะวิเคราะห์ว่าข้อความที่เข้ามาถูกจัดอยู่ในกลุ่มใด

เมื่อได้โมเดลเพื่อการจ่าแนกบทวิจารณ์สินค้าอิเล็กทรอนิกส์แล้ว สามารถน่ามาใช้จัดกลุ่มข้อความแสดงความคิดเห็นที่ผู้ซื้อได้ไปแสดงความคิดเห็นไว้ในเว็บไซต์ Amazon ที่มีการแสดงความคิดเห็นเกี่ยวกับสินค้าอิเล็กทรอนิกส์ เพื่อจ่าแนกระดับตามที่ต้องการ ส่าหรับขั้นตอนในการจ่าแนกระดับคะแนนข้อความบทวิจารณ์ มีดังนี้

ตัวอย่างข้อความแสดงความคิดเห็น

D_{new} : impressive and good device.

ขั้นตอนแรกจะเป็นการตัดค่าและการตัดค่าหยุด เพื่อก่าจัดค่าที่ไม่มีนัยส่าคัญกับเอกสารออก

ตารางที่ 3.17 แสดงคำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ NV

Document	ข้อความที่ผ่านกระบวนการ pre-processing
New	impressive / good / device

เมื่อได้คำสำคัญจากข้อความแสดงความคิดเห็นแล้ว เราจะใช้โมเดลที่สร้างขึ้นด้วยอัลกอริทึมข้างต้นในการวิเคราะห์ข้อความแสดงความคิดเห็น โดยจะมี 2 โมเดล ดังนี้

(1) การนำโมเดลไปใช้ในจำแนกบทวิจารณ์ด้วย Naïve Bayes

ในการจัดกลุ่มเอกสารข้อความที่เข้ามาใหม่ จะประเมินจากผลรวมความน่าจะเป็นของแต่ละคำในเอกสาร โดยใช้ความน่าจะเป็นของแต่ละคำที่ถูกคำนวณไว้ก่อนหน้า ในที่นี้จะประเมินจากทุกคลาส ถ้าหากค่าประเมินในคลาสใดสูงสุด จะสรุปได้ว่าเอกสารที่นำมาประเมินอยู่ในกลุ่มนั้น

$$v_{NB} = argmax\ P(v_j) \times \prod_{i=1}^{n} P\ (a_i \mid v_j) \qquad : v_j \in V \tag{22}$$

จากสมการที่ 3.7 ซึ่งกำหนดให้ V_{NB} คือเอกสารที่ผ่านการจัดกลุ่ม ซึ่งสามารถคำนวณความน่าจะเป็นเพื่อประเมินคลาส ทีละคลาสตามลำดับ โดย *P*(*Class*) = 0.5 ในทุกคลาส จะได้ว่า

โมเดลการจำแนกบทวิจารณ์ที่มีการให้น้ำหนักคำแบบ *tf-idf*

พิจารณาใน *Class* = “*Positive*”

V_{NEW} = *P*(*Positive*) x *P*(*good*|*Positive*) x *P*(*device*|*Positive*)

= (0.5) x (0.2052) x (0.1682)

= 0.01725732

พิจารณาใน *Class* = “*Negative*”

V_{NEW} = *P*(*Negative*) x *P*(*good*| *Negative*) x *P*(*device*| *Negative*)

= (0.5) x (0.1064) x (0.1384)

= 0.00736288

จากผลลัพธ์ข้างต้นจะเห็นได้ว่า D_{NEW} นั้นมีค่าความน่าจะเป็นอยู่ที่ 0.01725732 ใน *Class*= “*Positive*” มากกว่า *Class* = “*Negative*” ดังนั้นจึงสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ *Positive*

โมเดลการจำแนกบทวิจารณ์ที่มีการให้น้ำหนักคำแบบ *Delta TF-IDF*

พิจารณาใน *Class = "Positive"*

V_{NEW} = *P*(*Positive*) x *P*(*good*|*Positive*) x *P*(*device*|*Positive*)

= (0.5) x (1.8012) x (0.2212)

= 0.19921272

พิจารณาใน *Class = "Negative"*

V_{NEW} = *P*(*Negative*) x *P*(*good*| *Negative*) x *P*(*device*| *Negative*)

= (0.5) x (-0.1059) x (-0.2005)

= 0.010616475

จากผลลัพธ์ข้างต้นจะเห็นได้ว่า D_{NEW} นั้นมีค่าความน่าจะเป็นอยู่ที่ 0.19921272 ใน *Class= "Positive"* มากกว่า *Class = "Negative"* ดังนั้นจึงสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ *Positive*

โมเดลการจำแนกบทวิจารณ์ที่มีการให้น้ำหนักคำแบบ *TF-IDF-ICF*

พิจารณาใน *Class = "Positive"*

V_{NEW} = *P*(*Positive*) x *P*(*good*|*Positive*) x *P*(*device*|*Positive*)

= (0.5) x (0.2233) x (0.1597)

= 0.017830505

พิจารณาใน *Class = "Negative"*

V_{NEW} = *P*(*Negative*) x *P*(*good*| *Negative*) x *P*(*device*| *Negative*)

= (0.5) x (0.0901) x (0.1172)

= 0.00527986

จากผลลัพธ์ข้างต้นจะเห็นได้ว่า D_{NEW} นั้นมีค่าความน่าจะเป็นอยู่ที่ 0.017830505 ใน *Class= "Positive"* มากกว่า *Class = "Negative"* ดังนั้นจึงสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ *Positive*

โมเดลการจำแนกบทวิจารณ์ที่มีการให้น้ำหนักคำแบบ *TF-RF*

พิจารณาใน *Class = "Positive"*

V_{NEW} = *P*(*Positive*) x *P*(*good*|*Positive*) x *P*(*device*|*Positive*)

= (0.5) x (0.2729) x (0.2729)

= 0.037237205

พิจารณาใน *Class* = "*Negative*"

V_{NEW} = *P*(*Negative*) x *P*(*good*| *Negative*) x *P*(*device*| *Negative*)

= (0.5) x (0.0440) x (0.1137)

= 0.0025014

จากผลลัพธ์ข้างต้นจะเห็นได้ว่า D_{NEW} นั้นมีค่าความน่าจะเป็นอยู่ที่ 0.037237205 ใน *Class*= "*Positive*" มากกว่า *Class* = "*Negative*" ดังนั้นจึงสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ *Positive*

โมเดลการจำแนกบทวิจารณ์ที่มีการให้น้ำหนักคำแบบ *TF-IGM*

พิจารณาใน *Class* = "*Positive*"

V_{NEW} = *P*(*Positive*) x *P*(*good*|*Positive*) x *P*(*device*|*Positive*)

= (0.5) x (0.3935) x (0.1851)

= 0.036418425

พิจารณาใน *Class* = "*Negative*"

V_{NEW} = *P*(*Negative*) x *P*(*good*| *Negative*) x *P*(*device*| *Negative*)

= (0.5) x (0.0172) x (0.0774)

= 0.00066564

จากผลลัพธ์ข้างต้นจะเห็นได้ว่า D_{NEW} นั้นมีค่าความน่าจะเป็นอยู่ที่ 0.036418425 ใน *Class*= "*Positive*" มากกว่า *Class* = "*Negative*" ดังนั้นจึงสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่มของ *Positive*

(2) การนำโมเดลไปใช้ในจำแนกบทวิจารณ์ด้วย KNN

$$\sqrt{(\boldsymbol{p_1} - \boldsymbol{q_1})^2 + (\boldsymbol{p_2} - \boldsymbol{q_2})^2 + \cdots + (\boldsymbol{p_n} - \boldsymbol{q_n})^2} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{p_i} - \boldsymbol{q_i})^2} \quad (23)$$

จากสมการที่ 20 เป็นการหาค่าระยะทางระหว่างเอกสารที่เข้ามาใหม่ กับทุกเอกสารที่อยู่ในโมเดลว่าเอกสารใดมีความใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุด โดยค่าระยะทางยิ่งน้อยแสดงว่าเอกสารที่เข้ามาใหม่ใกล้เคียงกับเอกสารนั้นๆ มาก ซึ่งในโครงงานปริญญานิพนธ์นี้จะพิจารณาเอกสารที่ใกล้เคียงมากที่สุด 3 และ 5 เอกสาร โดยเรียงจากน้อยไปมาก

การใช้โมเดลการจำแนกบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย KNN โดยใช้การให้น้ำหนักคำแบบ *tf-idf* อ้างอิงค่าที่ใช้พิจารณาจากตารางที่ 3.2

ตัวอย่างเอกสารที่เข้ามาใหม่

D_{new} : impressive and good device.

ตารางที่ 3.18 คำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ TF-IDF

Word	impressive	good	device
D_{New}	1	1	1

ให้น้ำหนักคำในเอกสารตัวอย่างด้วย *tf-idf*

$$W_{good} = 1 * 0.477 = 0.477$$

$$W_{device} = 1 * 0.301 = 0.301$$

พิจารณาใน *Class* = “*Negative*”

$$D_1 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.477)^2 + (0 - 0.301)^2}$$
$$= 0.31813$$

$$D_2 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.477)^2 + (0.301 - 0.301)^2}$$
$$= 0.227529$$

$$D_3 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.477)^2 + (0 - 0.301)^2}$$
$$= 0.31813$$

พิจารณาใน *Class* = “*Positive*”

$$D_4 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0.477 - 0.477)^2 + (0 - 0.301)^2}$$
$$= 0.090601$$

$$D_5 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0.477 - 0.477)^2 + (0.301 - 0.301)^2}$$
$$= 0.0$$

$$D_6 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.477)^2 + (0.301 - 0.301)^2}$$
$$= 0.227529$$

พิจารณาโดยใช้ $K = 3$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4 และ D_2 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 2 เอกสาร และอยู่ในกลุ่ม *Negative* จำนวน 1 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

พิจารณาโดยใช้ $K = 5$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4, D_2, D_1 และ D_1 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 3 เอกสาร อยู่ในกลุ่ม *Negative* จำนวน 2 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive* การใช้โมเดลการจำแนกระดับคะแนนบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย KNN โดยใช้การให้น้ำหนักคำแบบ *Delta TF-IDF* อ้างอิงค่าที่ใช้พิจารณาจากตารางที่ 3.3

ตัวอย่างเอกสารที่เข้ามาใหม่

D_{new} : impressive and good device.

ตารางที่ 3.19 คำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ Delta TF-IDF

Word	impressive	good	device
D_{New}	1	1	1

ให้น้ำหนักคำในเอกสารตัวอย่างด้วย *Delta TF-IDF*

$$W_{good} = 1 * log_2\left(\frac{1 * 0 + 0.5}{1 * 0 + 0.5}\right) = 0$$
$$W_{device} = 1 * log_2\left(\frac{1 * 0 + 0.5}{1 * 0 + 0.5}\right) = 0$$

พิจารณาใน *Class* = “*Negative*”

$$D_1 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0-0)^2 + (0-0)^2}$$
$$= 0.0$$

$$D_2 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0-0)^2 + (0.893-0)^2}$$
$$= 0.893$$

$$D_3 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0-0)^2 + (0-0)^2}$$
$$= 0.0$$

พิจารณาใน *Class* = “*Positive*”

$$D_4 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(-3.700-0)^2 + (0-0)^2}$$
$$= 3.7$$

$$D_5 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(-3.700-0)^2 + (-0.893-0)^2}$$
$$= 4.593$$

$$D_6 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0-0)^2 + (-0.893-0)^2}$$
$$= 0.893$$

พิจารณาโดยใช้ K = 3 จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_1, D_2 และ D_6 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Negative* จำนวน 2 เอกสาร และอยู่ในกลุ่ม *Positive* จำนวน 1 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Negative*

พิจารณาโดยใช้ K = 5 จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_1, D_2, D_6, D_3 และ D_4 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Negative*

จำนวน 3 เอกสาร อยู่ในกลุ่ม *Positive* จำนวน 2 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Negative*

การใช้โมเดลการจำแนกระดับคะแนนบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย KNN โดยใช้การให้น้ำหนักคำแบบ *TF-IDF-ICF* อ้างอิงค่าที่ใช้พิจารณาจากตารางที่ 3.4

ตัวอย่างเอกสารที่เข้ามาใหม่

D_{new} : impressive and good device.

ตารางที่ 3.20 คำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ TF-IDF-ICF

Word	impressive	good	device
D_{New}	1	1	1

ให้น้ำหนักคำในเอกสารตัวอย่างด้วย *TF-IDF-ICF*

$$W_{good} = 1 * 0.477 * 1.301 = 0.620$$

$$W_{device} = 1 * 0.301 * 1 = 0.301$$

พิจารณาใน *Class = "Negative"*

$$D_1 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.620)^2 + (0 - 0.301)^2}$$
$$= 0.921$$

$$D_2 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.620)^2 + (0.301 - 0.301)^2}$$
$$= 0.620$$

$$D_3 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.620)^2 + (0 - 0.301)^2}$$
$$= 0.921$$

พิจารณาใน *Class = "Positive"*

$$D_4 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0.620 - 0.620)^2 + (0 - 0.301)^2}$$
$$= 0.301$$

$$D_5 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0.620 - 0.620)^2 + (0.301 - 0.301)^2}$$
$$= 0$$
$$D_6 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 0.620)^2 + (0.301 - 0.301)^2}$$
$$= 0.620$$

พิจารณาโดยใช้ $K = 3$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4 และ D_2 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 2 เอกสาร และอยู่ในกลุ่ม *Negative* จำนวน 1 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

พิจารณาโดยใช้ $K = 5$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4, D_2, D_6 และ D_1 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 3 เอกสาร อยู่ในกลุ่ม *Negative* จำนวน 2 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

การใช้โมเดลการจำแนกระดับคะแนนบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย KNN โดยใช้การให้น้ำหนักคำแบบ *TF-RF* อ้างอิงค่าที่ใช้พิจารณาจากตารางที่ 3.5

ตัวอย่างเอกสารที่เข้ามาใหม่

D_{new} : impressive and good device.

ตารางที่ 3.21 คำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ TF-RF

Word	impressive	good	device
D_{New}	1	1	1

ให้น้ำหนักคำในเอกสารตัวอย่างด้วย *TF-RF*

$$W_{good} = 1 * log_2\left(2 + \frac{1}{\max(1,0)}\right) = 1.584$$
$$W_{device} = 1 * log_2\left(2 + \frac{1}{\max(1,0)}\right) = 1.584$$

พิจารณาใน *Class = "Negative"*

$$D_1 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$

$= \sqrt{(0-1.584)^2 + (0-1.584)^2}$

$= 2.2401$

$D_2 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$

$= \sqrt{(0-1.584)^2 + (1.584-1.584)^2}$

$= 1.584$

D_3

$= \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$

$= \sqrt{(0-1.584)^2 + (0-1.584)^2}$

$= 2.2401$

พิจารณาใน *Class = "Positive"*

$D_4 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$

$= \sqrt{(0-1.584)^2 + (2.000-1.584)^2}$

$= 1.6377$

$D_5 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$

$= \sqrt{(2.000-1.584)^2 + (2.000-1.584)^2}$

$= 0$

$D_6 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$

$= \sqrt{(2.000-1.584)^2 + (0.0-1.584)^2}$

$= 1.6377$

พิจารณาโดยใช้ $K = 3$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4 และ D_2 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 2 เอกสาร และอยู่ในกลุ่ม *Negative* จำนวน 1 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

พิจารณาโดยใช้ $K = 5$ จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_2, D_4, D_6 และ D_1 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 3 เอกสาร อยู่ในกลุ่ม *Negative* จำนวน 2 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

การใช้โมเดลการจำแนกระดับคะแนนบทวิจารณ์สินค้าอิเล็กทรอนิกส์ด้วย KNN โดยใช้การให้น้ำหนักคำแบบ *TF-IGM* อ้างอิงค่าที่ใช้พิจารณาจากตารางที่ 3.6

ตัวอย่างเอกสารที่เข้ามาใหม่

D_{new} : impressive and good device.

ตารางที่ 3.22 คำสำคัญที่ได้หลังจากผ่านกระบวนการ pre-processing ในการทดสอบ TF-IGM

Word	impressive	good	device
D_{New}	1	1	1

ให้น้ำหนักคำในเอกสารตัวอย่างด้วย *TF-IGM*

$$W_{good} = 1 * (1 + 7.0 * 1) = 7.0$$

$$W_{device} = 1 * (1 + 7.0 * 0.5) = 3.5$$

พิจารณาใน *Class = "Negative"*

$$D_1 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 7.0)^2 + (0 - 3.5)^2}$$
$$= 7.82623792125$$

$$D_2 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 7.0)^2 + (3.5 - 3.5)^2}$$
$$= 7.0$$

$$D_3 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(0 - 7.0)^2 + (0 - 3.5)^2}$$
$$= 7.82623792125$$

พิจารณาใน *Class = "Positive"*

$$D_4 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(7.0 - 7.0)^2 + (0 - 3.5)^2}$$
$$= 3.5$$

$$D_5 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$
$$= \sqrt{(7.0 - 7.0)^2 + (3.5 - 3.5)^2}$$

$$= 0$$

$$D_6 = \sqrt{(Old_{good} - New_{good})^2 + (Old_{device} - New_{device})^2}$$

$$= \sqrt{(0 - 7.0)^2 + (3.5 - 3.5)^2}$$

$$= 7.0$$

พิจารณาโดยใช้ K = 3 จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4 และ D_2 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 2 เอกสาร และอยู่ในกลุ่ม *Negative* จำนวน 1 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

พิจารณาโดยใช้ K = 5 จะเห็นว่า เอกสารที่มีความใกล้เคียงกับ D_{NEW} มากที่สุด คือ D_5, D_4, D_2, D_6 และ D_1 ตามลำดับ ซึ่งเอกสารที่ใกล้เคียงกับเอกสารที่เข้ามาใหม่มากที่สุดอยู่ในกลุ่ม *Positive* จำนวน 3 เอกสาร อยู่ในกลุ่ม *Negative* จำนวน 2 เอกสาร ดังนั้นจึงสามารถสรุปได้ว่า D_{NEW} จัดอยู่ในกลุ่ม *Positive*

3.4.2 การวัดประสิทธิภาพของตัวจัดกลุ่มเอกสาร (Evaluation)

การวัดประสิทธิภาพของตัวจัดกลุ่มเอกสารเป็นขั้นตอนการประเมินโมเดลเพื่อใช้ในการจัดกลุ่มเอกสารก่อนการนำไปใช้งานจริงที่โดยทั่วไป จะใช้เทคนิคมาตรฐานที่นิยมใช้กันอย่างแพร่หลาย ที่เรียกว่า การวัดค่าความระลึก (Recall) การวัดค่าความแม่นยำ (Precision) และการวัดค่า F-Measure ตัวอย่าง

ตารางที่ 3.23 ตัวอย่าง Confusion Matrix

N=50		Prediction	
		Class 1	Class 2
Actual	Class 1	42	5
	Class 2	8	45

1. การวัดค่าความระลึก (Recall)

สมมุติให้ แต่ละ class มีเอกสารจำนวน 50 เอกสาร ซึ่งรวมทั้งสิ้น 100 เอกสาร และในการจำแนกเอกสารอัตโนมัติทำนายได้ถูกต้องตามความจริง (TP) และทำนายผิด (FN) จะได้ค่าความระลึกดังต่อไปนี้

$$R(class\ 1) = 42/(42+8) = 0.84$$

R(class 2) = 45/(45+5) = 0.90

ดังนั้น Average Recall = (0.84+0.90)/2 = 0.87

2. การวัดค่าความแม่นยำ (Precision)

สมมุติให้ แต่ละ class มีเอกสารจำนวน 50 เอกสาร ซึ่งรวมทั้งสิ้น 100 เอกสาร และในการจำแนกเอกสารอัตโนมัติทำนายได้ถูกต้องตามความจริง (TP) และทำนายไม่ถูกต้องตามความจริง (FP) จะได้ค่าความแม่นยำดังต่อไปนี้

P(class 1) = 42/(42+5) = 0.8936

P(class 2) = 45/(45+8) = 0.8490

ดังนั้น Average Precision = (0.8936+0.8490)/2 = 0.8713

3. การวัดค่า F-Measure

คือผลเฉลี่ยระหว่างค่าความแม่นย่าและค่าความระลึกสามารถแสดงตัวอย่างค่านวณได้ดังต่อไปนี้

F-measure = 2 * (0.87*0.8713)/(0.87+0.8713)

= 0.8706

3.5 การปรับปรุงประสิทธิภาพโมเดลเพื่อการจำแนก

3.5.1 ปัญหาจากการทำ Lemmatization

เนื่องจากการทำ Lemma เป็นการเปลี่ยนคำให้อยู่ในรูปแบบดั้งเดิม ตัวอย่างเช่น คำว่า This’s จะถูกเปลี่ยนเป็น this, be และด้วยเหตุนี้เอง ทำให้คำบางคำที่มีผลต่อการแสดงความรู้สึกอาจถูกเปลี่ยนแปลงไป เช่นคำว่า don’t เมื่อผ่านกระบวนการเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิมแล้วจะได้คำว่า do และคำว่า not ซึ่งจะเห็นว่า หากสองคำนี้ถูกแยกออกจากกันทำให้ความหมาย หรือค่าน้ำหนักของคำเปลี่ยนแปลงไป เช่น

I don’t like this device. จะได้คำว่า I / do / ’/ not / like / this / device / .

และจากตัวอย่างข้างต้นจะเห็นได้ว่า หลังจากผ่านการทำ Lemma จะได้อักขระพิเศษเข้ามาในการประมวลผลด้วยดังตัวอย่าง ทำให้มีคำมากยิ่งขึ้นซึ่งคำที่ได้ไม่ได้มีผลกับการแสดงความรู้สึกแต่ถูกนำมาคำนวณ ทำให้ใช้ระยะเวลาในการประมวลผลมากขึ้น

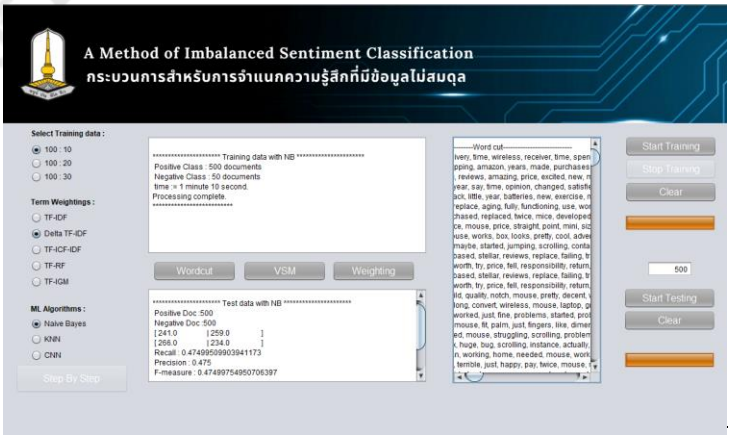
3.5.2 ปัญหาด้านการใช้ภาษา

เนื่องจากข้อมูลที่ใช้ในการสร้างโมเดลเป็นเอกสารข้อความแสดงความคิดเห็นเกี่ยวสินค้าอิเล็กทรอนิกส์ ที่เปิดให้ทุกคนสามารถเข้ามาเขียนแสดงความคิดเห็นและให้คะแนนสินค้านั้นๆ ได้ ทำให้เกิดปัญหาด้านการใช้ภาษา คือการใช้คำที่ไม่มีความหมาย หรือไม่มีในพจนานุกรม (Unknown word) ดังนั้น จึงได้มีการนำพจนานุกรมมาใช้เพื่อคัดกรองคำเหล่านั้นออกไป เพราะคำเหล่านั้นไม่ได้มีความหมาย หรือส่งผลต่อการจัดกลุ่มเอกสาร

ภาพประกอบที่ 3.6 ตัวอย่าง Unknown word

3.6 ตัวอย่างหน้าจอโปรแกรม

ตัวอย่างหน้าจอการทำงานของโปรแกรมที่เราจะนำเสนอระบบการควบคุมข้อมูลไม่สมดุลในการจำแนกความรู้สึก



ภาพประกอบที่ 3.7 ตัวอย่างหน้าจอโปรแกรม