บทที่ 5

สรุปและอภิปรายผลการทดลอง

ในบทนี้จะเป็นการสรุปภาพรวมของการสร้างโมเดลการจำแนกเอกสารข้อความที่ไม่สมดุลจากข้อความแสดงความคิดเห็นของลูกค้าที่ซื้อสินค้าอิเล็กทรอนิกส์ต่างๆ ที่ได้ทำการรวบรวมไว้ดังนี้

5.1 สรุปผลและอภิปรายผล

โครงงานฉบับนี้ เป็นงานวิจัยทางด้านการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุล โดยใช้ชุดข้อมูลที่เป็นบทวิจารณ์สินค้าอิเล็กกทรอนิกส์ ซึ่งเป็นการสร้างโมเดลที่ไม่มีความสมดุลของข้อมูล ที่มีสัดส่วน 100 : 10, 100 : 20 และ 100 : 30 เพื่อคัดแยกกลุ่มข้อความ โดยใช้อัลกอริทึมอีฟเบย์ (*Naïve Bayes*) อัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด (*KNN*) และอัลกอริทึมโครงข่ายประสาทแบบคอนโวลูชัน *(Convolutional Neural Network: CNN)* ส่วนการให้น้ำหนักคำจะมีอยู่ 5 รูปแบบหลักคือ *TF-IDF, Delta TF-IDF, TF-ICF-IDF, TF-RF* และ *TF-IGM*

ขั้นตอนในการสร้างโมเดลการจำแนกข้อมูลที่ไม่สมดุลนั้น ในขั้นตอนแรกจะเป็นการรวบรวมข้อมูลที่เป็นบทวิจารณ์สินค้าอิเล็กทรอนิกส์มาจากเว็บไซต์ Amazon จากนั้นคัดแยกออกเป็นสองกลุ่ม โดยจะมีการแบ่งเอกสารออกเป็น 2 ชุด คือ ชุดข้อมูลสอน (Training) และ ชุดข้อมูลทดสอบ (Test set) โดยชุดข้อมูลสอนจะแบ่งเป็นชุดย่อยสามชุดที่มีขนาดข้อมูลที่ไม่สมดุลกันโดยสัดส่วนข้อมูลชุดหลักที่เป็น Positive class มากกว่าข้อมูลที่เป็นชุดรอง Negative class คือ 100 : 10, 100 : 20 และ 100 : 30 คัดแยกข้อมูลเสร็จเรียบร้อยแล้ว ก็จะนำข้อมูลเข้าสู่ขั้นตอนก่อนการประมวลผลต่อไป

ขั้นตอนก่อนการประมวลผล (Text pre-processing) เป็นการนำเอาเอกสารที่ได้จากขั้นตอนก่อนหน้านี้มาทำการตัดคำ การตัดคำหยุด การคัดเลือกคำด้วยพจนานุกรม และการเลือกคุณลักษณะเพื่อกรองคำที่มีความเกี่ยวข้องกับการจัดกลุ่มเอกสารน้อยที่สุดออกด้วย *IG* และเพื่อหาจำนวนคำทั้งหมดในเอกสาร โดยเอกสารที่ผ่านกระบวนการนี้จะอยู่ในรูปแบบ *Vector Space Model* เพื่อแสดงให้เห็นถึงความสัมพันธ์ระหว่างเอกสารและคำที่ปรากฏในเอกสาร พร้อมทั้งการให้น้ำหนักของคำเพื่อแสดงว่าคำๆ นั้นมีความสำคัญกับเอกสารมากน้อยเพียงใด โดยการให้น้ำหนักคำจะมีอยู่ 5 รูปแบบคือ *TF-IDF, Delta TF-IDF, TF-ICF-IDF, TF-RF* และ *TF-IGM* ซึ่งการให้น้ำหนักคำในเอกสารจะให้น้ำหนักแยกตาม class สำหรับการสร้างโมเดลจะมีอัลกอริทึมที่ใช้ในการสร้างโมเดล 3 อัลกอริทึม คืออัลกอริทึมอีฟเบย์ (*Naïve Bayes*) อัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด (*KNN*) และอัลกอริทึมโครงข่ายประสาทแบบคอนโวลู

ชัน *(Convolutional Neural Network: CNN)* ต่อไปจะเข้าสู่ขั้นตอนการจัดเก็บโมเดลเพื่อใช้ในการประมวลผลถัดไป

ขั้นตอนการทดสอบโมเดลเมื่อได้โมเดลการจำแนกข้อมูลที่ไม่สมดุลเกี่ยวกับสินค้าอิเล็กทรอนิกส์เป็นที่เรียบร้อยแล้วจากขั้นตอนข้างต้น สามารถนำมาใช้จัดระดับคะแนนข้อความบทวิจารณ์สินค้าอิเล็กทรอนิกส์ของผู้ซื้อสินค้าอื่นๆ เพื่อให้ทราบว่าบทวิจารณ์นั้นๆ จัดควรอยู่ในกลุ่ม Positive class หรือ Negative class

สำหรับขั้นตอนการวัดประสิทธิภาพในโครงงานนี้จะใช้ การวัดค่าความระลึก ค่าความแม่นยำ และการวัดค่า *F-measure* โดยค่าความระลึกจะเป็นอัตราส่วนของเอกสารที่จัดกลุ่มได้จากเอกสารทั้งหมดที่มีอยู่ ส่วนค่าความแม่นยำเป็นอัตราส่วนของเอกสารที่จัดกลุ่มได้ถูกต้อง จากจำนวนของเอกสารทั้งหมดที่จัดกลุ่มได้ ค่า *F-measure* เป็นการพิจารณาค่าความสัมพันธ์ระหว่างค่าความระลึกและค่าความแม่นยำ

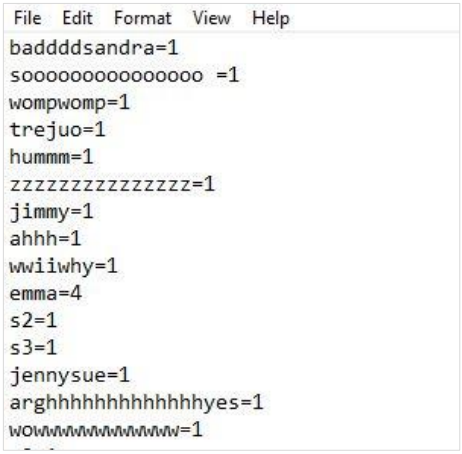
5.2 ปัญหาและอุปสรรคในการดำเนินงาน

5.2.1 ปัญหาเกี่ยวกับอัลกอริทึมในการสร้างโมเดล

เนื่องจากอัลกอริทึม *CNN* และ *KNN* นั้นเหมาะกับการทดลองกับชุดข้อมูลชุดสอนที่มีขนาดใหญ่ แต่ในโครงงานปริญญานิพนธ์นี้ได้ ทำการทดลองกับชุดข้อมูลชุดสอนที่มีขนาดเล็ก จึงทำให้ไม่สามารถดึงประสิทธิภาพสูงสุดของอัลกอริทึม *CNN* และ *KNN* ออกมาได้ ทั้งนี้เวลาในการสร้างโมเดลนั้นค่อนข้างนาน ดังนั้นควรจะทำการ save model ไว้หากได้ค่าเฉลี่ยที่พึ่งพอใจแล้ว

5.2.2 ปัญหาเกี่ยวกับชุดข้อมูลที่ใช้ในการสร้างโมเดล

เนื่องจากเอกสารข้อความแสดงความคิดเห็นเกี่ยวกับสินค้าอิเล็กทรอนิกส์ที่รวบรวมมานั้น เป็นข้อความที่ทุกคนที่ซื้อสินค้า สามารถเข้ามาเขียนแสดงความรู้สึกต่อสินค้านั้นๆ ได้ ทำให้เกิดการใช้คำที่ไม่มีความหมาย (Unknown word) และไม่พบในพจนานุกรม ทำให้การสร้างและการทดสอบโมเดลมีความไม่เสถียร ถึงแม้ในโครงงานนี้จะใช้พจนานุกรมในการคัดกรองคำเหล่านั้นแล้วก็ตาม แต่ในอนาคตก็อาจจะมีคำเหล่านี้หลุดเข้าในขั้นตอนการสร้างโมเดลได้



ภาพประกอบที่ 5.1 ตัวอย่างคำที่ไม่มีความหมาย (Unknown word)

5.3 ข้อเสนอแนะ

1. การให้น้ำหนักคำแต่ละรูปแบบ STW ควรมีชุดข้อมูล 2 กลุ่มเป็นต้นไปและมีขนาดข้อมูลที่มีขนาดใหญ่
2. ประสิทธิภาพของโมเดลจะขึ้นอยู่กับจำนวนเอกสารที่ใช้ในการสร้างโมเดล และความถูกต้องของเอกสารที่ใช้สร้างโมเดลด้วย

ดังนั้น การสร้างโมเดลหรือตัวจัดกลุ่มเอกสาร ควรมีจำนวนคำศัพท์ที่จำเป็นสำหรับการจัดกลุ่มปริมาณไม่น้อยจนเกินไป และถ้าหากคำศัพท์ที่รวบรวมมาตรงกับเอกสารข้อความที่ต้องการนำมาวิเคราะห์เพื่อจัดกลุ่ม จะทำให้ประสิทธิภาพในการวิเคราะห์ของโปรแกรมมีมากยิ่งขึ้น รวมไปถึงจำนวนเอกสารที่ใช้ในขั้นตอนการสร้างโมเดล เพราะโครงงานฉบับนี้นำเสนออัลกอริทึมการเรียนรู้แบบมีผู้สอน