โปสเตอร์โครงงาน

Computer Science Department
Faculty of Informatics, Mahasarakham University

A Method of Imbalanced Sentiment Classification

กระบวนการสำหรับการจำแนกความรู้สึกที่มีข้อมูลไม่สมดุล

ผู้พัฒนา : พีระวัฒน์ บุญบ้านงิ้ว (Pheerawat Bunbanngio)
อาจารย์ที่ปรึกษา : จันทิมา พลพินิจ (Jantima Polpinij)
Intellect Laboratory สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
knarf.pheerawat@gmail.com, jantima.p@msu.ac.th

ที่มาและความสำคัญ

การจำแนกความรู้สึก (Sentiment Classification) คือการจำแนกเอกสารตามขั้นความรู้สึกซึ่งโดยทั่วไปอาจจะจำแนกเป็นความรู้สึกที่เป็นบวก (Positive) ความรู้สึกที่เป็นลบ (Negative) และความรู้สึกที่เป็นกลาง (Neutral) โดยการจำแนกความรู้สึกนั้น ได้รับการศึกษามาอย่างต่อเนื่อง เพราะการประยุกต์ใช้ในหลายลักษณะ แต่โดยทั่วไปมักจะนิยมใช้ในการจำแนกความรู้สึกที่มีการแสดงไว้ในรูปแบบข้อความ (Text) เช่น ประยุกต์ใช้ในการจัดอันดับความรู้สึกจากข้อความแสดงความคิดเห็นของผู้คนที่มีต่อสินค้าและบริการ การประยุกต์ใช้เพื่อวิเคราะห์ความรู้สึกของผู้เรียน การประยุกต์ใช้เพื่อวิเคราะห์ความรู้สึกของผู้คนในเรื่องการเมือง เป็นต้น ซึ่งปัญหาความไม่สมดุลของข้อมูลในคลาสนั้น เกิดจากกลุ่มตัวอย่างที่ใช้ในการเรียนรู้มีข้อมูลไม่สมดุลกัน โดยกลุ่มที่มีข้อมูลมากกว่าจะเรียกว่า "ข้อมูลกลุ่มหลัก (Majority Class)" ขณะที่กลุ่มตัวอย่างที่มีข้อมูลจำนวนน้อยกว่าจะเรียกว่า "ข้อมูลกลุ่มรอง (Minority Class)" เมื่อนำเอาชุดข้อมูลในลักษณะนี้ไปเรียนรู้เพื่อสร้างตัวจำแนกความรู้สึก (Sentiment Classifier) ข้อมูลใหม่ๆ ที่อ่านเข้ามาเพื่อวิเคราะห์เพื่อจำแนกกลุ่มด้วยตัวจำแนกความรู้สึกดังกล่าว ก็มีแนวโน้มที่จะทำนายกลุ่มของข้อมูลนั้นไปยังทิศทางของข้อมูลกลุ่มหลักที่ใช้ในการเรียนรู้ตัวจำแนกความรู้สึก ดังนั้น ในโครงงานปริญญานิพนธ์ฉบับนี้ จึงได้นำเสนอการศึกษาการแก้ปัญหาความไม่สมดุลของข้อมูลในการจำแนกความรู้สึกด้วยเทคนิคการให้น้ำหนักคำ 5 เทคนิค คือ TF-IDF, Delta TF-IDF, TF-IDF-ICF, TF-RF และ TF-IGM ร่วมกับแมชชีนเลิร์นนิง 3 ตัว คือ Naive Bayes, K-Nearest Neighbor และสุดท้าย Convolution Neural Network

คำสำคัญ: การจำแนกเอกสาร, การให้น้ำหนักคำ, ข้อมูลไม่สมดุล, เซ็พพอร์ตเวคเตอร์แมชชีน

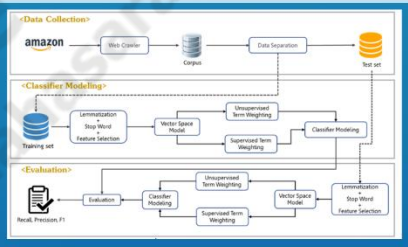
วัตถุประสงค์

นำเสนอกระบวนการสำหรับการจำแนกความรู้สึกที่มีข้อมูลไม่สมดุลโดยมีเครื่องมือหลักคือเทคนิคการให้น้ำหนักคำแบบมีผู้สอน

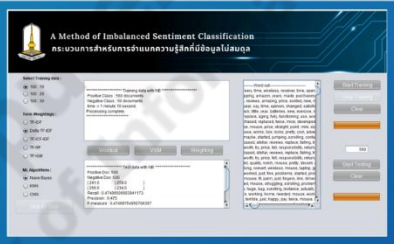
ประโยชน์ที่คาดว่าจะได้รับ

ได้กระบวนการในการจำแนกข้อความแสดงความรู้สึกที่มีข้อมูลแบบไม่สมดุล

กรอบการดำเนินงาน



ตัวอย่างหน้าจอการทำงาน



สรุป

เนื่องจากบ่อยครั้งที่ การจำแนกเอกสารที่ไม่สมดุลกันนั้นมีการเอนเอียงการให้คำแนนไปฝั่งที่มีข้อมูลมากกว่าเนื่องจากมีข้อมูลที่ครอบคลุมการทำนายที่ดีกว่า

ดังนั้นงานวิจัยฉบับนี้จึงได้นำเสนอวิธีการการจำแนกข้อมูลที่ไม่สมดุลด้วยการให้น้ำหนักคำเปรียบเทียบ 2 รูปแบบหลักคือ UTW และ STW โดย UTW ใช้รูปแบบการให้น้ำหนักคำที่ได้รับความนิยมมากที่สุดคือ TF-IDF และ STW ใช้ทั้งหมด 4 รูปแบบคือ Delta TF-IDF, TF-ICF-IDF, TF-RF และ TF-IGM โดยผลที่ได้คือการให้น้ำหนักคำแบบ STW มีประสิทธิภาพในการจำแนกข้อมูลที่ไม่สมดุลมากกว่ารูปแบบการให้น้ำหนักคำแบบ UTW ซึ่งได้แก่การให้น้ำหนักคำแบบ TF-IGM โดยใช้อัลกอริทึม CNN ในการสร้างโมเดล มีค่าเฉลี่ย F-measure สูงที่สุดอยู่ที่ 74.41%

Computer Science Department
Faculty of Informatics, Mahasarakham University