บทที่ 4

ผลการทดลอง

หลังจากดำเนินการตามขั้นตอนในบทที่ 3 แล้วในบทนี้จะเป็นการบรรยายถึงผลการทดลองดังรายละเอียดต่อไปนี้

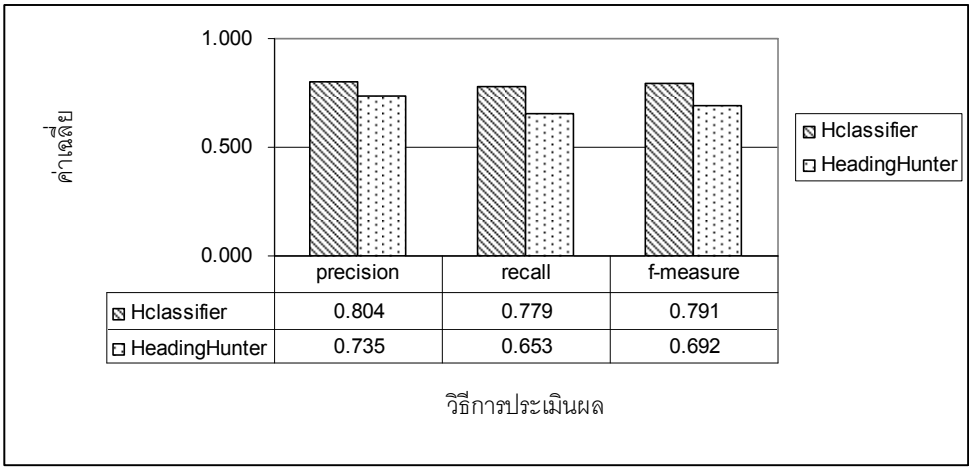
4.1 ผลการทดลอง

กำหนดให้ HClassifier แทนการจำแนกหัวเรื่องด้วยวิธีการที่นำเสนอในวิทยานิพนธ์นี้ และ HeadingHunter แทนวิธีการจำแนกหัวเรื่องของ (Brudvik, 2008) ซึ่งจะแบ่งการนำเสนอผลการทดลองออกเป็นหัวข้อดังต่อไปนี้ ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องจากลักษณะเด่น, ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร, ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องด้วยวิธีการแบ่งเว็บเพจเป็นส่วนย่อยและซัพพอร์ตเวกเตอร์แมชชีน, ผลการศึกษาลักษณะเด่นในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ซึ่งมีรายละเอียดดังต่อไปนี้

4.1.1 ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องจากลักษณะเด่น

ภาพที่ 4.1

แสดงการเปรียบเทียบผลการจำแนกหัวเรื่องจากลักษณะเด่นจากวิธีของ HClassifier กับลักษณะเด่นจากวิธีของ HeadingHunter โดยใช้ต้นไม้ตัดสินใจเป็นการเรียนรู้ด้วยเครื่องจักร



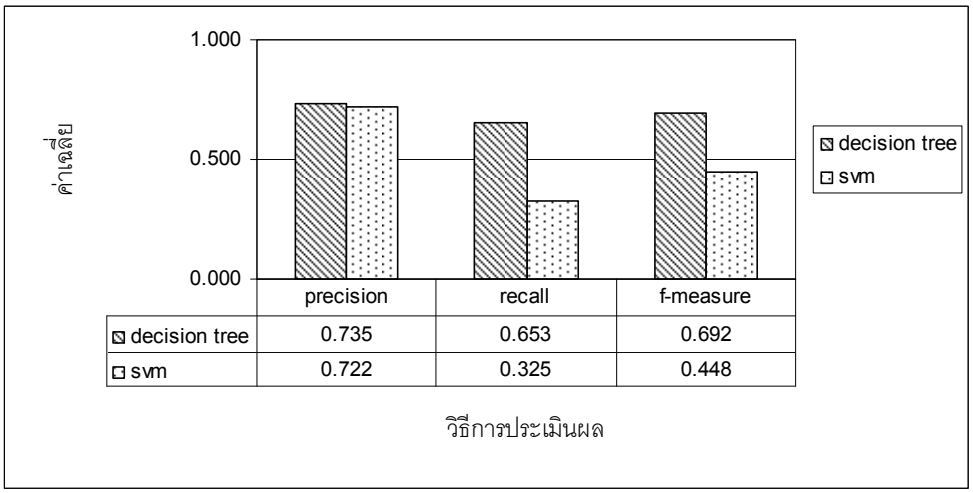
	precision	recall	f-measure
Hclassifier	0.804	0.779	0.791
HeadingHunter	0.735	0.653	0.692

จากภาพที่ 4.1 แสดงค่าเฉลี่ยรวมของวิธีการประเมินผลแบบต่างๆ ของการจำแนกหัวเรื่องของเว็บเพจโดยใช้ลักษณะเด่นจากวิธีของ HClassifier กับ ลักษณะเด่นจากวิธี HeadingHunter โดยใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร จะเห็นได้ว่าค่าความถูกต้องของ HClassifier มีค่า 0.804 และ HeadingHunter มีค่า 0.735 ส่วนค่าความระลึกของ HClassifier มีค่า 0.779 และ HeadingHunter มีค่า 0.653 ส่วนค่า f-measure ของ HClassifier มีค่า 0.791 และ HeadingHunter มีค่า 0.692 จากผลการทดลองสามารถสรุปได้ว่าลักษณะเด่นซึ่งได้จากวิธีของ HClassifier ทำให้ประสิทธิภาพของการจำแนกหัวเรื่องมีประสิทธิภาพสูงขึ้น ร้อยละ 9 ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 จากการวัดด้วยค่า f-measure ซึ่งเป็นค่าวัดประสิทธิภาพโดยรวมของระบบ

4.1.2 ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร

ภาพที่ 4.2

แสดงการเปรียบเทียบผลการจำแนกหัวเรื่องจากลักษณะเด่นของวิธี HeadingHunter โดยใช้ต้นไม้ตัดสินใจและซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้ด้วยเครื่องจักร

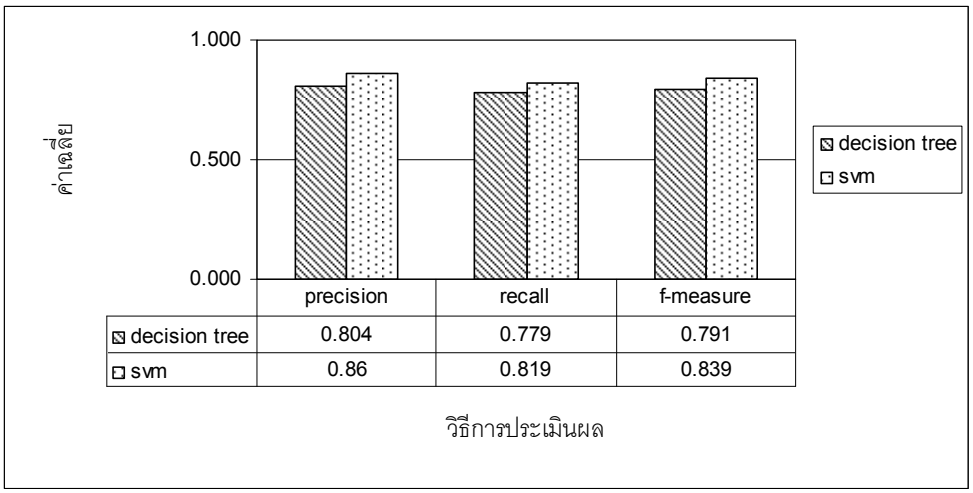


	precision	recall	f-measure
decision tree	0.735	0.653	0.692
svm	0.722	0.325	0.448

จากภาพที่ 4.2 แสดงค่าเฉลี่ยรวมของวิธีการประเมินผลแบบต่างๆ ของการจำแนกหัวเรื่องของเว็บเพจโดยใช้ลักษณะเด่นจากวิธีของ HeadingHunter โดยใช้ต้นไม้ตัดสินใจและ ซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้ด้วยเครื่องจักร จะเห็นได้ว่าค่าความถูกต้องของการใช้ต้นไม้ตัดสินใจ มีค่า 0.735 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.722 ส่วนค่าความระลึกของ ต้นไม้ตัดสินใจ มีค่า 0.653 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.325 ส่วนค่า f-measure ของ ต้นไม้ตัดสินใจ มีค่า 0.692 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.448 จากผลการทดลองสามารถสรุปได้ว่าเมื่อใช้ลักษณะเด่นจากวิธีของ HeadingHunter โดยใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักรจะมีประสิทธิภาพการจำแนกหัวเรื่องสูงกว่าการใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร ร้อยละ 24 ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 จากการวัดด้วยค่า f-measure ซึ่งเป็นค่าวัดประสิทธิภาพโดยรวมของระบบ

ภาพที่ 4.3

แสดงการเปรียบเทียบผลการจำแนกหัวเรื่องจากลักษณะเด่นของวิธี HClassifier โดยใช้ต้นไม้ตัดสินใจและซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้ด้วยเครื่องจักร



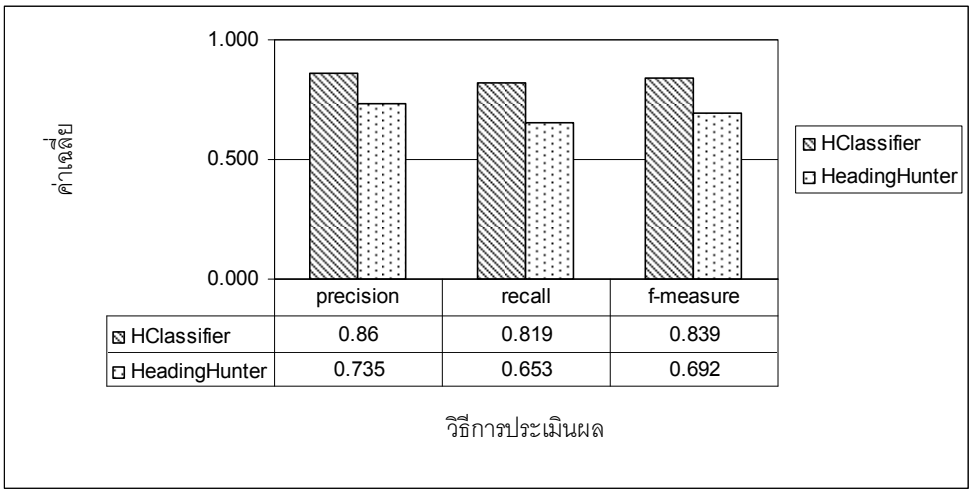
	precision	recall	f-measure
decision tree	0.804	0.779	0.791
svm	0.86	0.819	0.839

จากภาพที่ 4.3 แสดงค่าเฉลี่ยรวมของวิธีการประเมินผลแบบต่างๆ ของการจำแนกหัวเรื่องของเว็บเพจโดยใช้ลักษณะเด่นจากวิธีของ HClassifier โดยใช้ต้นไม้ตัดสินใจและซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้ด้วยเครื่องจักร จะเห็นได้ว่าค่าความถูกต้องของการใช้ต้นไม้ตัดสินใจ มีค่า 0.804 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.86 ส่วนค่าความระลึกของ ต้นไม้ตัดสินใจ มีค่า 0.779 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.819 ส่วนค่า f-measure ของ ต้นไม้ตัดสินใจ มีค่า 0.791 และซัพพอร์ตเวกเตอร์แมชชีน มีค่า 0.839 จากผลการทดลองสามารถสรุปได้ว่าเมื่อใช้ลักษณะเด่นจากวิธีของ HClassifier โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักรจะมีประสิทธิภาพการจำแนกหัวเรื่องสูงกว่าการใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร ร้อยละ 4 ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 จากการวัดด้วยค่า f-measure ซึ่งเป็นค่าวัดประสิทธิภาพโดยรวมของระบบ

4.1.3 ผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องด้วยวิธีการแบ่งเว็บเพจเป็นส่วนย่อยและซัพพอร์ตเวกเตอร์แมชชีน

ภาพที่ 4.4

แสดงการเปรียบเทียบผลการจำแนกหัวเรื่องระหว่างวิธี HClassifier กับ HeadingHunter

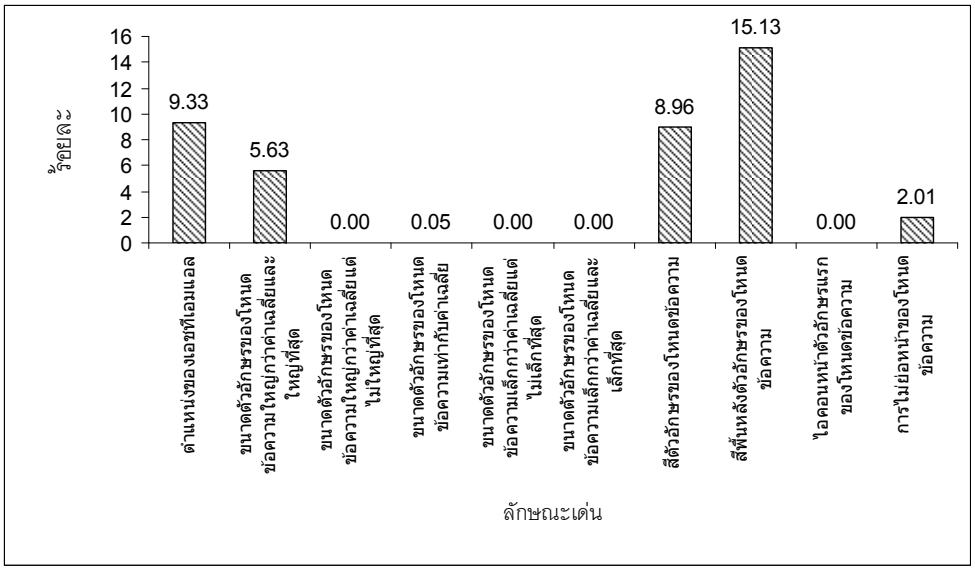


จากภาพที่ 4.4 แสดงค่าเฉลี่ยรวมของวิธีการประเมินผลแบบต่างๆ ของการจำแนกหัวเรื่องของเว็บเพจระหว่างวิธีของ HClassifier กับ วิธี HeadingHunter จะเห็นได้ว่าค่าความถูกต้องของ HClassifier มีค่า 0.86 และ HeadingHunter มีค่า 0.735 ส่วนค่าความระลึกของ HClassifier มีค่า 0.819 และ HeadingHunter มีค่า 0.653 ส่วนค่า f-measure ของ HClassifier มีค่า 0.839 และ HeadingHunter มีค่า 0.692 จากผลการทดลองสรุปได้ว่าการจำแนกหัวเรื่องโดยวิธี HClassifier สามารถปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องได้สูงขึ้นร้อยละ 14 เมื่อเปรียบเทียบกับวิธี HeadingHunter ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 จากการวัดด้วยค่า f-measure ซึ่งเป็นค่าวัดประสิทธิภาพโดยรวมของระบบ

4.1.4 ผลการทดลองการศึกษาลักษณะเด่นในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ

ภาพที่ 4.5

แสดงผลกระทบของแต่ละลักษณะเด่นของ HClassifier ในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ



จากภาพที่ 4.5 จากผลการทดลองผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจของแต่ละลักษณะเด่นเรียงตามความสำคัญในการจำแนกหัวเรื่อง พบว่าลักษณะเด่นสีพื้นหลังตัวอักษรของโหนดข้อความให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ร้อยละ 15.13, ลักษณะเด่นตำแหน่งของเอชทีเอมแอลให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ร้อยละ 9.33, ลักษณะเด่นของสีตัวอักษรของโหนดข้อความให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจร้อยละ 8.96, ลักษณะเด่นขนาดตัวอักษรของโหนดข้อความใหญ่กว่าค่าเฉลี่ยและใหญ่ที่สุดให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ร้อยละ 5.63, ลักษณะเด่นการไม่ย่อหน้าของโหนดข้อความให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ร้อยละ 2.01, ลักษณะเด่นขนาดตัวอักษรของโหนดข้อความเท่ากับค่าเฉลี่ยให้ผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ร้อยละ 0.05 และลักษณะเด่นซึ่งไม่ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ได้แก่ ลักษณะเด่นขนาดตัวอักษรของโหนดข้อความใหญ่กว่าค่าเฉลี่ยแต่ไม่ใหญ่ที่สุด, ลักษณะเด่นขนาดตัวอักษรของ

โหนดข้อความเล็กกว่าค่าเฉลี่ยแต่ไม่เล็กที่สุด, ลักษณะเด่นขนาดตัวอักษรของโหนดข้อความเล็กกว่าค่าเฉลี่ยและเล็กที่สุด และลักษณะเด่นไอคอนหน้าตัวอักษรตัวแรกของโหนดข้อความ

4.2 อภิปรายผลการทดลอง

ในบทนี้จะอภิปรายผลการทดลองจากผลการทดลองที่ได้นำเสนอไป ดังหัวข้อต่อไปนี้ ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องจากลักษณะเด่น, ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร, ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องด้วยวิธีการแบ่งเว็บเพจเป็นส่วนย่อยและซัพพอร์ตเวกเตอร์แมชชีน และ ผลการศึกษาลักษณะเด่นในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ ซึ่งมีรายละเอียดดังต่อไปนี้

4.2.1 อภิปรายผลการทดลองผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องจากลักษณะเด่น

จากผลการทดลองสามารถสรุปได้ว่าลักษณะเด่นซึ่งได้นำเสนอในวิทยานิพนธ์นี้ทำให้ประสิทธิภาพของการจำแนกหัวเรื่องมีประสิทธิภาพสูงขึ้น ร้อยละ 9 ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 จากการวัดด้วยค่า f-measure ซึ่งเป็นค่าวัดประสิทธิภาพโดยรวมของระบบ ซึ่งเป็นผลมาจากการนำลักษณะเด่นจากการที่มนุษย์ใช้ในการแยกแยะหัวเรื่องของเว็บเพจได้แก่ ขนาดตัวอักษร, สีตัวอักษร, สีพื้นหลังตัวอักษร, ไอคอนหน้าข้อความ, การไม่ย่อหน้าของข้อความ และ ลำดับของโหนดข้อความภายในส่วนย่อยมาใช้ในวิทยานิพนธ์นี้ทำให้จำแนกหัวเรื่องของเว็บเพจได้หลากหลายรูปแบบมากขึ้นและมีประสิทธิภาพสูงขึ้น เช่น หัวเรื่องซึ่งใช้สีพื้นหลังของข้อความในการนำเสนอ แต่วิธีการของ Brudvik ซึ่งใช้ลักษณะเด่น ได้แก่ ขนาดตัวอักษร, ความหนาของตัวอักษร, ความยาวของข้อความ, อัตราส่วนความกว้างเทียบกับความสูง จะไม่สามารถจำแนกหัวเรื่องประเภทนี้ได้

4.2.2 อภิปรายผลการทดลองผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร

จากผลการทดลองสามารถสรุปได้ว่าเมื่อใช้ลักษณะเด่นจากวิธีของ Brudvik โดยใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักรมีประสิทธิภาพการจำแนกหัวเรื่องสูงกว่าการใช้ซัพพอร์ต

เวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร และ เมื่อใช้ลักษณะเด่นจากวิธีของซึ่งได้นำเสนอในวิทยานิพนธ์นี้โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักรมีประสิทธิภาพการจำแนกหัวเรื่องสูงกว่าการใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร แสดงให้เห็นว่าจากลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้เหมาะสมกับการใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้ด้วยเครื่องจักรเนื่องจากซัพพอร์ตเวกเตอร์แมชชีนสามารถจำแนกประเภทข้อมูลแบบหลายมิติได้ดีกว่าต้นไม้ตัดสินใจ (multi-dimensions features) เช่น หัวเรื่องประกอบด้วยรูปแบบการแสดงที่ใช้จำแนกหัวเรื่องหลายรูปแบบ เช่น หัวเรื่องใช้ตัวอักษรขนาดใหญ่กว่าปกติและสีพื้นหลังตัวอักษรต่างจากส่วนอื่น เนื่องจากลักษณะเด่นจากวิธีของ Brudvik ประกอบด้วยลักษณะเด่นเพียง 4 ลักษณะเด่นเท่านั้นและใช้เพียงลักษณะเด่นที่เกี่ยวกับขนาดและจำนวนตัวอักษรเท่านั้นทำให้มิติของลักษณะเด่นน้อยกว่าจึงเหมาะสมกับการใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร แต่ลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้ประกอบด้วย 10 ลักษณะเด่นจากรูปแบบของการแสดงผลที่หลากหลายกว่าจึงเหมาะสมกับการใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร เช่น พิจารณาถึงสีตัวอักษรของข้อความที่นำมาพิจารณาว่าเป็นหัวเรื่องหรือไม่ ในขณะที่งานวิจัยของ Brudvik ไม่ได้พิจาณาในส่วนนี้

4.2.3 อภิปรายผลการทดลองการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร

จากผลการทดลองสามารถสรุปได้ว่าวิธีที่นำเสนอในวิทยานิพนธ์นี้สามารถปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องได้สูงขึ้นร้อยละ 14 เมื่อเปรียบเทียบกับวิธี Brudvik ซึ่งแตกต่างอย่างมีระดับนัยสำคัญที่ 0.01 ซึ่งเกิดจากการปรับปรุงลักษณะเด่นซึ่งสามารถเพิ่มประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจให้สูงขึ้นเมื่อเปรียบเทียบกับลักษณะเด่นจากวิธีของ Brudvik และการใช้ซัพพอร์ตซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักรซึ่งสามารถเพิ่มประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจให้สูงขึ้นเมื่อใช้กับประเภทของลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้

4.2.4 อภิปรายผลการทดลองผลการทดลองการศึกษาลักษณะเด่นในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ

จากผลการทดลองว่าสามารถสรุปได้ว่าลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้ ลักษณะเด่นซึ่งมาจากการที่ลักษณะเด่นแตกต่างจากส่วนอื่นของส่วนย่อยอย่างชัดเจน จะมีผลต่อประสิทธิภาพการจำแนกหัวเรื่องค่อนข้างสูง ได้แก่ ลักษณะเด่นของโหนดข้อความมีสีพื้นหลังตัวอักษรต่างจากสีพื้นหลังหลักของส่วนย่อยหรือไม่ (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 15.13), ลักษณะเด่นโหนดข้อความมีสีตัวอักษรต่างจากสีตัวอักษรหลักของส่วนย่อยหรือไม่ (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 8.98), ลักษณะเด่นของโหนดข้อความมีขนาดตัวอักษรใหญ่ที่สุดในส่วนย่อย (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 5.63) และลักษณะเด่นของโหนดข้อความมีการใช้รูปแบบการย่อหน้าหรือไม่ (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 2.01) ซึ่งเป็นผลมาจากผู้ออกแบบเว็บไซค์จะออกแบบส่วนที่เป็นหัวเรื่องให้เด่นชัดกว่าส่วนอื่นของส่วนย่อยนั้น ในขณะที่ลักษณะเด่นที่ไม่เด่นชัดกว่าส่วนอื่นในส่วนย่อย ได้แก่ ลักษณะเด่นของโหนดข้อความที่ขนาดตัวอักษรใหญ่กว่าค่าเฉลี่ยแต่ไม่ใช่ขนาดตัวอักษรที่มีขนาดใหญ่ที่สุดในส่วนย่อย, ลักษณะเด่นโหนดข้อความมีขนาดตัวอักษรเล็กกว่าค่าเฉลี่ยแต่ไม่ใช่ตัวอักษรที่มีขนาดเล็กที่สุดในส่วนย่อย, ลักษณะเด่นโหนดข้อความมีขนาดตัวอักษรเล็กที่สุดในส่วนย่อยจะไม่ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง และ ลักษณะเด่นของโหนดข้อความมีการใช้ไอคอนหน้าโหนดข้อความหรือไม่ ส่วนลักษณะเด่นของโหนดข้อความมีขนาดตัวอักษรเท่ากับค่าเฉลี่ยขนาดตัวอักษรของส่วนย่อยส่งผลกระทบเล็กน้อยต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 0.05) เนื่องจากผู้ออกแบบเว็บไซค์อาจจะออกแบบส่วนที่เป็นหัวเรื่องของเว็บเพจให้มีขนาดตัวอักษรเท่ากันกับข้อความอื่นๆ และใช้สีพื้นหลังตัวอักษรที่แตกต่างจากส่วนอื่นในการนำเสนอส่วนที่เป็นหัวเรื่อง และจากการที่หัวเรื่องจะปรากฏที่เริ่มต้นของแต่ละส่วนย่อยทำให้ลักษณะเด่นโหนดข้อความมีลำดับภายในส่วนย่อยมีค่าน้อยกว่า 10 หรือไม่ จะมีผลต่อประสิทธิภาพการจำแนกหัวเรื่องค่อนข้างสูงเช่นเดียวกัน (ส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่อง ร้อยละ 9.33)