การพัฒนาโมเดลตรวจจับคำหยาบภาษาไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์นิง

DEVELOPMENT OF THAI LANGUAGE PROFANITY INVESTIGATION MODEL FOR ONLINE MEDIA USING DATA MINING TECHNIQUE

ณัฐาศิริ เชาว์ประสิทธิ์[1] สมชาย เล็กเจริญ[2]

[1] สาขาวิชาเทคโนโลยีสารสนเทศ (วิทยาลัยเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยรังสิต) nachow@rpu.ac.th

[2] สาขาวิชาเทคโนโลยีสารสนเทศ (วิทยาลัยเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยรังสิต) somchai.l@rsu.ac.th

---

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบกระบวนการวิเคราะห์คำหยาบภาษาไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์นิง โดยใช้โมเดลในการตรวจจับคำหยาบภาษาไทยด้วยพจนานุกรมคำหยาบที่ผ่านการปรับปรุงใช้เทคนิค TFICF (Term Frequency - Inverse Class Frequency) จากการวิจัยครั้งนี้พบว่า การศึกษาเปรียบเทียบกระบวนการวิเคราะห์คำหยาบด้วยเทคนิคดาต้าไมน์นิง ได้แก่ เทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.96 และค่าความคลาดเคลื่อนเฉลี่ย กำลังสอง (RMSE) เท่ากับ 0.19 รองลงมา เทคนิคนาอีฟเบย์ มีค่าความถูกต้อง เท่ากับ 0.96 และค่าความคลาดเคลื่อนเฉลี่ย กำลังสอง (RMSE) เท่ากับ 0.21 และ เทคนิคเคเนียร์เรสเนเบอร์ให้ค่าความถูกต้องที่น้อยกว่า โดยมีค่าความถูกต้อง (Accuracy) เท่ากับ 0.95 และค่าความคลาดเคลื่อนเฉลี่ย กำลังสอง (RMSE) เท่ากับ 0.22 แม้ว่าเทคนิคต้นไม้ตัดสินใจและเทคนิคนาอีฟเบย์จะให้ค่าความถูกต้องที่เท่ากัน แต่พบว่าการวิเคราะห์คำหยาบด้วยเทคนิคต้นไม้ตัดสินใจมีค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) น้อยที่สุด และมีรูปแบบการวิเคราะห์ข้อความที่ง่ายต่อความเข้าใจมากกว่าเทคนิคอื่นๆ

คำสำคัญ: คำหยาบ, เว็บบอร์ด, พจนานุกรม, ต้นไม้ตัดสินใจ, เคเนียร์เรสเนเบอร์, นาอีฟเบย์

ABSTRACT

This research aims to comparatively study the process of Thai language profanity investigation for online media with data mining techniques. These models were used to investigate Thai language profanity by using a profanity dictionary that improved by Term Frequency-Inverse Class Frequency technique (or TFICF). According to this research, the comparatively study process of Thai profanity investigation with data mining techniques such as decision tree technique which gave the accuracy at 0.96 and root mean square error (RMSE) equal to 0.19, followed by Naive Bayes technique which gave the accuracy at 0.96 and RMSE equal to 0.21, and K-Nearest Neighbor technique which gave the lowest accuracy at 0.95 and RMSE equal to 0.22. Although the decision tree and Naive Bayes techniques gave the similar accuracy, profanity investigation using decision tree technique had the lowest RMSE and easy analysis pattern to more understand when compared with the other techniques.

Keywords: Profanity, web board, dictionary, Decision tree, K-Nearest Neighbors (K-NN), Naïve Bayes

1. บทนำ

การใช้ข้อความภาษาไทยบนสื่อออนไลน์ที่ผู้เขียนมีจุดประสงค์ให้ผู้พบเห็นรู้สึกเจ็บปวด เสื่อมเสียชื่อเสียง คำด่าหยาบคาย การสร้างความแตกแยก จนถึงการกระทบถึงสถาบันอันเป็นที่เคารพในสังคมออนไลน์ทั่วโลก ยังคงเป็นปัญหาที่สำคัญของสังคมเป็นอย่างมากโดยเฉพาะการเรียนรู้จำของเยาวชนไทย ในปัจจุบันยังสร้างความหนักใจสำหรับการตรวจสอบของผู้ดูแลเว็บไซต์ เนื่องจากภาษาที่ใช้มีลักษณะเป็นทั้งภาษาพูด และภาษาเขียนผสมกัน อีกทั้งการดัดแปลงการเขียนโดยการตั้งใจสะกดผิด การใช้ตัวย่อ สัญลักษณ์ คำสแลง (Sood, Sara Owsley, Judd Antin, and Elizabeth Churchill,2012) เพื่อหลีกเหลี่ยงโปรแกรมคัดกรอง หรือตรวจจับคำหยาบ มีนักวิจัย ได้พยายามศึกษาพฤติกรรมการใช้ภาษาในเนื้อหาที่ไม่เหมาะสมของผู้ใช้บริการสังคมออนไลน์เพื่อหาวิธีการตรวจจับคำหยาบให้แม่นยำมากยิ่งขึ้น (Xiang, G., Hong, J., and Rose, C. P.,2012) ศูนย์ข้อมูลกฎหมายและคดีเสรีภาพโดย ไอลอว์ เมื่อวันที่ 2 กุมภาพันธ์ พ.ศ. 2559 สถิติการระงับการเผยแพร่เนื้อหาและการสั่งปิดเว็บไซต์ ได้รวบรวมข้อมูลผ่านฐานข้อมูลคดีออนไลน์ของศาลอาญา ปรากฏว่า ในช่วงเดือนมกราคม 2556 จนถึงเดือนธันวาคม 2557 ศาลอาญามีคำสั่งระงับการเผยแพร่เนื้อหาทั้งสิ้น 123 ฉบับ รวมจำนวน 9,328 URL พบว่าเนื้อหาที่ถูกปิดกั้นการเข้าถึงเป็นอันดับหนึ่งคือ เนื้อหาและภาพซึ่งมีลักษณะดูหมิ่น หมิ่นประมาท พระมหากษัตริย์ ราชินี และรัชทายาท ร้อยละ 82.83 อันดับที่สองคือ เนื้อหาและภาพซึ่งมีลักษณะลามกอนาจาร ร้อยละ 16.58 และอับดับที่สามคือ เนื้อหาและภาพที่มีลักษณะขัดต่อความสงบเรียบร้อยหรือศีลธรรมอันดีของประชาชน คิดเป็นร้อยละ 0.32 ส่วนเนื้อหาที่เป็นการหมิ่นประมาทบุคคลธรรมดามีการขอคำสั่งศาลให้ปิดกั้นเป็นจำนวนน้อยมาก เมื่อเทียบกับทั้งหมด ปัจจุบันปัญหาการใช้คำพูดที่รุนแรงหรือ ด่า หมายถึงใช้ถ้อยคำว่าคนอื่นด้วยถ้อยคำที่หยาบช้าเลวทราม (Tanasanti Jirapon, Phokharatkul Pisit, Buntilov Vladimir, and Kanoksilpatham Budsaba.,2012) จนถึง ส่อเสียด หมิ่นประมาทผู้อื่น หรือหมิ่นกระทบถึงสถาบันเบื้องสูงมีมากมายในสื่อออนไลน์ ซึ่งยากต่อการตรวจจับคำหยาบจนบางเว็บไซต์ไม่สามารถปิดกั้นได้ ด้วยปัญหาการใช้ภาษาไทยมีคำศัพท์ใหม่ๆเกิดขึ้นอยู่เสมอด้วยจากการที่มีการปรับเปลี่ยนไปตามแต่ละยุคสมัยอย่างรวดเร็ว และการใช้ภาษาไทยมักจะใช้ภาษาพูดในการเขียน ใช้คำสแลง การลากเสียง การสะกดคำผิด หรือบางคำที่พิมพ์ด้วยความรีบเร่งและเป็นคำที่ใกล้เคียงกับกลุ่มคำพ้องเสียงแต่การกดแป้น Shift ทำให้เสียเวลาจึงไม่กด และเปลี่ยนคำที่ต้องการเป็นอีกคำที่ออกเสียงคล้ายกันแทน เช่น กูเกลียดมึง เป็น กุเกียดมึงงงง ไอ้สัตว์นรก เป็น ไอสัสนรก คำเหล่านี้ผู้เขียนต้องการหลีกเหลี่ยงระบบตรวจจับคำไม่เหมาะสม หรือคำหยาบ สาเหตุดังกล่าวผู้วิจัยจึงนำเสนอแนวทางการตรวจจับคำหยาบ โดยใช้โมเดลการตรวจจับคำหยาบด้วยพจนานุกรมคำหยาบที่ผ่านการปรับปรุงปรับเพิ่มลดคำศัพท์ (adaptive) โดยนำข้อมูลที่มนุษย์ และพจนานุกรมเห็นตรงกันนำมาวิเคราะห์ข้อความคำหยาบด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียร์เรสเนเบอร์ และนาอีฟเบย์ เพื่อนำมาเปรียบเทียบหาโมเดลที่ดีสุด มีรูปแบบเข้าใจง่ายและแม่นยำในการตรวจจับคำหยาบภาษาไทย

2. วัตถุประสงค์การวิจัย

เพื่อพัฒนาโมเดลที่ให้ค่าความแม่นยำ และถูกต้องที่สุดในการตรวจจับคำหยาบภาษาไทยบนสื่อออนไลน์ด้วยการใช้พจนานุกรมที่ผ่านการปรับปรุงโดยมีผลการวิเคราะห์จากผู้เชี่ยวชาญเป็นข้อมูลป้อนกลับ และนำมาวิเคราะห์ข้อความคำหยาบด้วยเทคนิคการจำแนกประเภทข้อมูล 3 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียร์เรสเนเบอร์ และเทคนิคนาอีฟเบย์

3. การดำเนินการวิจัย

คลังข้อความ

การสร้างคลังข้อความสำหรับการศึกษาครั้งนี้ได้ใช้ข้อมูลประเภทตัวอักษรบนกระดานสนทนา หรือข้อความแสดงความคิดเห็นตามสื่อบันเทิงออนไลน์ต่างๆ จากเว็บไซต์ ข่าว บันเทิง เกมออนไลน์ และประมูลขายสินค้า เว็บไซต์ดังกล่าวมีการเปิดเสรี รวมถึงการไม่ปิดกั้นคำที่ไม่เหมาะสม หรือคำหยาบ ในการแสดงความคิดของผู้ใช้บริการ ผู้วิจัยได้เก็บตัวอย่างที่ใช้ทดลองได้จำนวนข้อความ 1,214 โพสต์ (1 โพสต์ความยาวไม่เกิน 300 ตัวอักษร) อย่างไรก็ตามข้อมูลที่นำมาใช้ในการทดสอบนี้เป็นการสุ่มตัวอย่างเพื่อนำมาวิเคราะห์ และมีขนาดไม่ใหญ่นัก ผู้วิจัยก็พบว่าข้อมูลเพียงเท่านี้ก็เพียงพอในการทดสอบเพื่อศึกษาเปรียบเทียบการวิเคราะห์คำหยาบสำหรับนำไปพัฒนาโมเดลตรวจจับคำหยาบภาษาไทยในลักษณะต่างๆ ได้ดีพอสมควร

การเตรียมข้อมูล

ข้อมูลที่ใช้สำหรับในการวิจัยครั้งนี้เป็นข้อมูลที่มาจากเว็บไซต์ 4 กลุ่ม ได้แก่ ข่าว บันเทิง เกมออนไลน์ และประมูลขายสินค้า ซึ่งข้อมูลดังกล่าวประกอบด้วยคำสั่ง HTML และภาษาสคริปต์ ตามหลักโครงสร้างเว็บเพจ ผู้วิจัยจึงต้องนำแท็ก HTML ออกให้เหลือเฉพาะข้อความเท่านั้น เมื่อได้ข้อมูลที่เป็นเฉพาะข้อความ ขั้นตอนสำคัญ คือการประมวลผลทางภาษาด้วยการตัดคำภาษาไทย (Word Segmentation) แบบอิงพจนานุกรมผ่านการใช้โปรแกรมเล็กซ์โต (Lexto) งานวิจัยนี้ได้เลือกวิธีการตัดคำภาษาไทยที่เหมาะสมกับข้อมูลที่ใช้ในชีวิตประจำวัน ด้วยการตัดคำแบบยาวที่สุด (Longest Matching) (ยืน ภู่วรวรรณ,2529) ซึ่งมีความแม่นยำสำหรับการตัดคำในเอกสารได้สูงถึง 99% และนำข้อความที่ผ่านการตัดคำภาษาไทยมาดำเนินการกำจัดคำที่ไม่มีนัยสำคัญออก หรือ คำหยุด คือคำที่ไม่มีความสำคัญ เช่น กับ ของ หรือ ที่ เป็นต้น เพื่อประหยัดพื้นที่ และเวลาในการประมวลผล (C. Haruechaiyasak, S. Kongyoung and M. Dailey,2008)

พจนานุกรมคำหยาบ

พจนานุกรมคำหยาบภาษาไทยสำหรับใช้ในการศึกษาในงานวิจัยครั้งนี้ ได้ข้อมูลคำศัพท์จากทีมปฏิบัติการวิจัยวิทยาการมนุษยภาษา (HLT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์ และคอมพิวเตอร์แห่งชาติ (เนคเทค) เก็บรวบรวมคำศัพท์จากเว็บไซต์ agoda, pantip, facebook, twitter และ youtube ตั้งแต่ปี พ.ศ. 2552-2555 มีจำนวนทั้งสิ้น 663 คำ ถือเป็นพจนานุกรมต้นฉบับที่ใช้สำหรับการศึกษาในงานวิจัยในครั้งนี้ ซึ่งงานวิจัยก่อนหน้าผู้วิจัยได้นำพจนานุกรมคำหยาบมาทำการทดลองปรับปรุงด้วยเทคนิค TFICF (Term Frequency - Inverse Class Frequency) เพื่อปรับเพิ่มลดคำศัพท์ในพจนานุกรม ซึ่งพจนานุกรมคำหยาบที่ปรับปรุงด้วยการลดคำศัพท์มีความความแม่นยำมากที่สุด มีค่าความถูกต้อง (Accuracy) 0.87 ผู้วิจัยนำใช้พจนานุกรมลดคำศัพท์เนื่องจากมีค่าความแม่นยำในการตรวจจับคำหยาบสูงที่สุด และสามารถนำมาศึกษาเปรียบเทียบถึงการวิเคราะห์คำหยาบด้วยเทคนิคการจำแนกประเภทได้เป็นอย่างดี ซึ่งจำนวนคำศัพท์คำหยาบที่ใช้ในการศึกษาครั้งนี้มีจำนวนทั้งหมด 593 คำ (ณัฐาศิริ เชาว์ประสิทธิ์ กฤษฎา ศรีแผ้ว, 2557)

โมเดลตรวจจับคำหยาบด้วยพจนานุกรมลดคำหยาบ

โมเดลการตรวจจับคำหยาบภาษาไทยในการศึกษาเปรียบเทียบการวิเคราะห์คำหยาบด้วยเทคนิคการจำแนกประเภท 3 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียร์เรสเนเบอร์ และเทคนิคนาอีฟเบย์ ผู้วิจัยได้ใช้พจนานุกรมลดคำหยาบที่ผ่านการปรับปรุงจากพจนานุกรมต้นฉบับด้วยเทคนิค TFICF (Term Frequency - Inverse Class Frequency)ในการสกัดคำสำคัญภายในคลาส หรือต่างคลาสกัน ซึ่งโมเดลตรวจจับคำหยาบภาษาไทยด้วย

พจนานุกรมลดคำหยาบ ให้ความสำคัญกับความถี่ (Term Frequency) และความถี่ผกผัน (Inverse Frequency) ของคำที่ปรากฏอยู่ในข้อความเพราะมีผลต่อการสกัดคำสำคัญภายในคลาส (intra-class) และต่างคลาส (inter-class) ความสำคัญของวิธีการนี้คือการหาค่าน้ำหนักของคำในแต่ละคลาส ดังนิยามในสมการ (1) – (2) กำหนดให้มี N ข้อความ M คำที่แตกต่างกัน และแต่ละข้อความอยู่ในคลาสใดคลาสหนึ่งตั้งแต่ 1 ถึง 4

$$ICF_i^k = log\frac{N^k}{N_i^k} \quad (1)$$

$$TFICF_i^k = TF_{ij} \times ICF_i^k \quad (2)$$

โดย i คือ คำ = {1, 2, ...,M}
j คือ ข้อความ = {1, 2, ...,N}
k คือ คลาส = {1, 2, 3, 4}
$i \subseteq j \in k$ คือ คำอยู่ในข้อความ โดยแต่ละข้อความเป็น สมาชิกของคลาสใดคลาสหนึ่ง
N^k คือ จำนวนข้อความที่อยู่คลาส k
N_i^k คือ จำนวนข้อความที่มีคำ i ปรากฏอยู่ในคลาส k
ICF_i^k คือ ความถี่คลาสผกผันของคำ i ที่ปรากฏในคลาส k
TF_{ij} คือ ค่าความถี่ของคำ i ที่ปรากฏในข้อความ j

พจนานุกรมลดคำหยาบได้มีการประยุกต์แนวทางการสร้างพจนานุกรมด้วยการป้อนกลับความรู้ที่ได้จากการให้มนุษย์วิเคราะห์แล้วนำมาพิจารณาในลักษณะของ Confusion Matrix ดังรูปที่ 1

	กำกับด้วยพจนานุกรม (Predicted Class)	
กำกับด้วยผู้เชี่ยวชาญ (Actual class)	**หยาบ-หยาบ** (True Positive) $k = 1$	**หยาบ-ไม่หยาบ** (False Negative) $k = 2$
	ไม่หยาบ-หยาบ (False Positive) $k = 3$	**ไม่หยาบ-ไม่หยาบ** (True Positive) $k = 4$

รูปที่ 1 Confusion Matrix ระหว่างการกำกับของผู้เชี่ยวชาญและการกำกับโดยพจนานุกรมคำหยาบ

พจนานุกรมลดคำหยาบที่ผ่านการปรับปรุง เริ่มต้นด้วยการแบ่งกลุ่มข้อมูลโดยสามารถแบ่งกลุ่มได้จากการนำคำตอบของผู้เชี่ยวชาญที่ได้จากการกำกับข้อความ และผลลัพธ์ที่ได้จากพจนานุกรมคำหยาบต้นฉบับนำมาทำการแบ่งกลุ่ม โดยแต่ละกลุ่มสามารถแทนด้วย “A-B” โดยที่ A คือ การกำกับของผู้เชี่ยวชาญ และ B คือ การกำกับโดยพจนานุกรมคำหยาบต้นฉบับ ซึ่งสามารถแบ่งออกได้ 4 กลุ่ม คือ k = 1 คือ “หยาบ-หยาบ” k = 2 คือ “หยาบ-ไม่หยาบ” k = 3 คือ “ไม่หยาบ-หยาบ และ k = 4 คือ “ไม่หยาบ-ไม่หยาบ” ส่วนวิธีการปรับปรุงพจนานุกรมโดยการลดคำหยาบ

ผู้วิจัยได้ดำเนินการสกัดคำที่ไม่น่าจะเป็นคำหยาบที่ปรากฏอยู่ในพจนานุกรม โดยคำนั้นจะได้จากกลุ่มที่คนกำกับว่า “ไม่หยาบ” แต่พจนานุกรมกำกับว่า “หยาบ (k = 3) มาทำการวัดน้ำหนักของคำภายในคลาสด้วยสมการที่ (3)

$$V_i = \frac{1}{e^{ICF_i^3}} = \frac{N_i^3}{N_i} \qquad (3)$$

โดย V_i คือ น้ำหนักของคำ i ที่ปรากฏภายในคลาส k=3, V_i มีค่าตั้งแต่ 0 ถึง 1

เงื่อนไขในการลบคำออกจากพจนานุกรมคำหยาบ เป็นดังสมการที่ (4) นั่นคือ

$$V_i \geq \sigma \qquad (4)$$

โดย σ เป็นพารามิเตอร์ที่จะทำการศึกษาถึงค่าที่เหมาะสม ซึ่งค่าที่เหมาะสมสำหรับพจนานุกรมลดคำหยาบ ที่ค่า σ = 0.3 ให้ค่าความถูกต้องและค่าอัตราการรู้จำสูงที่สุดสามารถตรวจจับคำได้แม่นยำมีค่าเท่ากับ 0.87 จึงสามารถนำพจนานุกรมลดคำหยาบมาศึกษาการเปรียบเทียบวิเคราะห์ข้อความคำหยาบได้อย่างมีประสิทธิภาพ

การทดลอง และการศึกษาเปรียบเทียบเทคนิคจำแนกประเภทข้อมูล

ชุดข้อมูล

ในการศึกษาการเปรียบเทียบการวิเคราะห์ข้อความคำหยาบในครั้งนี้ผู้วิจัยได้ใช้ข้อความแสดงความคิดเห็นทั้งหมด 515 โพสต์ จากข้อมูลทั้งหมด 1,214 โพสต์ โดยเลือกกลุ่มข้อมูลกลุ่มที่ 1 (k=1) คน และพจนานุกรมกำกับเป็น “หยาบ-หยาบ” จำนวน 199 โพสต์ และ กลุ่มที่ 4 คน และพจนานุกรมกำกับเป็น “ไม่หยาบ-ไม่หยาบ” จำนวน 316 โพสต์

การเตรียมข้อมูล

ข้อมูลที่ใช้สำหรับในการศึกษาครั้งนี้ได้ผ่านกระบวนการประมวลผลด้วยภาษาธรรมชาติ และการเทียบคู่คำ (Matching) ระหว่างพจนานุกรมลดคำหยาบ และคำในข้อความ รวมถึงการใช้เทคนิค TFIC ในการสกัดคำภายในคลาสหรือต่างต่างคลาสกัน เครื่องมือที่ใช้ได้พัฒนาด้วยภาษาจาวา และการวิเคราะห์ข้อความคำหยาบ ผู้วิจัยได้ใช้โปรแกรม RapidMiner Studio เวอร์ชั่น 7.5 ในการสร้างโมเดลเพื่อศึกษาการเปรียบเทียบด้วยเทคนิคการจำแนกประเภท 3 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียร์เรสเนเบอร์ และเทคนิคนาอีฟเบย์

การทดลอง

นำข้อมูลที่ใช้สำหรับการทดลองในครั้งนี้มีจำนวนข้อความ 515 โพสต์ และ คำศัพท์คำหยาบจากพจนานุกรมลดคำหยาบจำนวน 593 คำ แล้วทำการวิเคราะห์ข้อความคำหยาบด้วยการนำข้อมูลไปเรียนรู้ในโปรแกรม RapidMiner Studio โดยแบ่งข้อมูลออกเป็น 80% สำหรับการเรียนรู้ และ 20% เป็นการทดสอบ และใช้เทคนิคการจำแนกข้อมูลในการวิเคราะห์คำหยาบ 3 เทคนิค

1) ต้นไม้ตัดสินใจ (Decision Tree) การสร้างโมเดลในการวิเคราะห์ข้อความคำหยาบโดยการจำแนกกระบวนการทำงานที่มีเงื่อนไขการตัดสินใจอยู่ในรูปของโหนด (Nodes) โหนดจะมีคุณลักษณะ (Attribute) กิ่งของต้นไม้ (Branch) แสดงค่าที่เป็นไปได้ของคุณลักษณะ และใบ (Leaf) ที่อยู่ด้านล่างสุดของต้นไม้ตัดสินใจ จะแสดงกลุ่ม หรือ คลาส (Class) การวิเคราะห์ด้วยเทคนิคดังกล่าว กำหนดคลาสเป็น 2 กลุ่ม คือ กลุ่ม Negative (หยาบคาย) และกลุ่ม Positive (ไม่หยาบ)

2) เทคนิคเคเนียร์เรสเนเบอร์(K-Nearest Neighbors หรือ K-NN) เป็นการวิเคราะห์ข้อความคำหยาบที่มีการจำแนกประเภทข้อมูล โดยการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายข้อมูลที่อยู่ใกล้เคียงเป็น

จำนวน K ตัว ซึ่งคำตอบที่ทำนายได้ คือ คลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้านทั้ง K ตัว ในเทคนิคนี้ใช้วิธีการวัดระยะห่างแบบ Euclidean ซึ่งเกิดจากรากที่สองของผลต่างระหว่างแอตทริบิวต์ต่างๆ ยกกำลัง

3) เทคนิคนาอีฟเบย์ (Naive Bayes) เป็นอีกวิธีที่ใช้ในการจำแนกประเภทข้อมูล ซึ่งใช้ทฤษฎีความน่าจะเป็น (Probability) เป็นหลักในการวิเคราะห์คำหยาบ โดยเทคนิคดังกล่าวจะทำนายผลลัพธ์ โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์

การวัดประสิทธิภาพ

การทดสอบเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกประเภททั้ง 3 เทคนิค ในการศึกษางานวิจัยนี้ ใช้วิธีการตามหลักแนวคิดเดียวกับการจำแนกประเภท โดยใช้ค่าความถูกต้อง (Accuracy) และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง ดังสมการ (5)

ค่าความถูกต้อง $$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \quad (5)$$

โดย TP คือ ข้อความหยาบคาย และผลการทำนายบอกว่าหยาบคาย

TN คือ ข้อความไม่หยาบคาย และผลการทำนายบอกว่าไม่หยาบคาย

FP คือ ข้อความหยาบคาย แต่ผลการทำนายบอกว่าไม่หยาบคาย

FN คือ ข้อความไม่หยาบคาย แต่ผลการทำนายบอกว่าหยาบคาย

ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) เป็นวิธีการวัดความคลาดเคลื่อนจากโมเดลที่ใช้สำหรับศึกษาเปรียบเทียบในงานวิจัยนี้ โดยเป็นค่าที่ทำนายจากโมเดลกับค่าที่เกิดขึ้นจริง หากค่า RMSE มีค่าน้อยถือได้ว่าโมเดลนั้นสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง หากแต่ว่าถ้าค่านี้มีค่าเท่ากับศูนย์แล้วนั้นจะหมายความว่า ไม่เกิดความคลาดเคลื่อนในโมเดลเลย ดังสมการ (6) (ณัฐภัทร แก้วรัตนภัทร์ และคณะ,2555)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(Y_i - \hat{Y}_i)^2}{N^k}} \quad (6)$$

กำหนดให้ ตัวแปร $\hat{Y}_i$ คือ ค่าข้อมูลที่แท้จริงที่ได้จากการคำนวณ

ตัวแปร Y_i คือ ค่าผลลัพธ์ที่ได้จากการทำนาย

ตัวแปร N^k คือ จำนวนข้อมูลตัวอย่างที่อยู่ภายในคลาสสำหรับประมาณการโมเดล

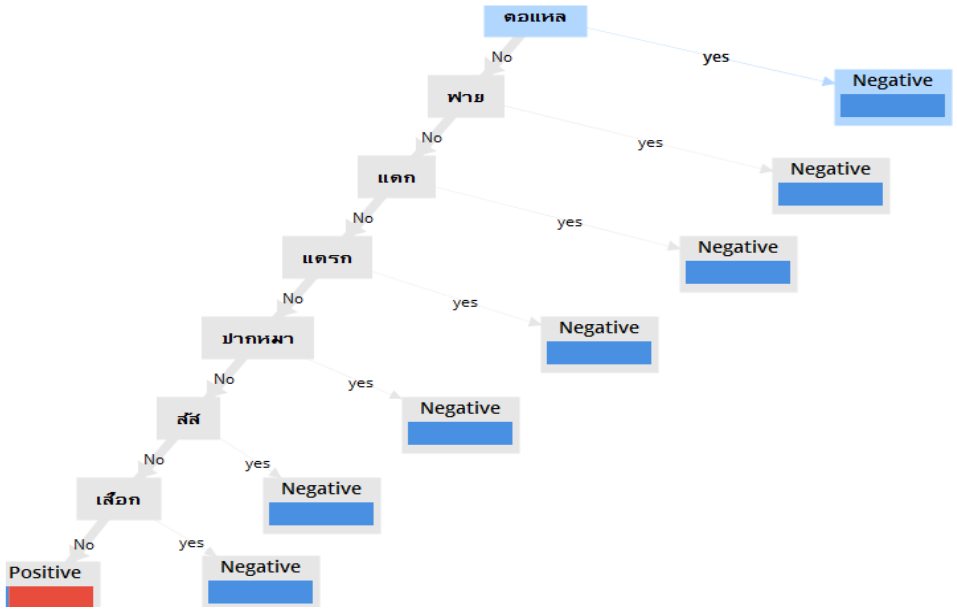
ตารางที่ 1 ผลการทดลองโมเดลเปรียบเทียบประสิทธิภาพการวิเคราะห์ข้อความคำหยาบ

Model	Accuracy	RMSE
Decision Tree	0.96	0.19
Naive Bayes	0.96	0.20
K-Nearest Neighbors	0.95	0.22

4. ผลการวิจัย

การพัฒนาโมเดลตรวจจับคำหยาบภาษาไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์นิง ได้ใช้ค่าความถูกต้อง (Accuracy) และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) ในการแสดงผลเปรียบเทียบความแม่นยำเพื่อหาโมเดลการวิเคราะห์คำหยาบที่มีประสิทธิภาพมากที่สุดสำหรับการตรวจจับคำ

ในตารางที่ 1 แสดงผลการเปรียบเทียบด้วยเทคนิคดาต้าไมน์นิง 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ เทคนิค เคเนียร์เรสเนเบอร์ และเทคนิคนาอีฟเบย์ ผลการทดลองพบว่า เทคนิคต้นไม้ตัดสินใจ สามารถวิเคราะห์คำหยาบให้ค่าความถูกต้องเท่ากับ 0.96 และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง เท่ากับ 0.19 รองลงมาเป็นเทคนิคนาอีฟเบย์ ให้ค่าความถูกต้องเท่ากับ 0.96 และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง เท่ากับ 0.20 เทคนิคสุดท้ายเทคนิคเคเนียร์เรสเนเบอร์ให้ค่าความถูกต้องเท่ากับ 0.95 และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง เท่ากับ 0.22 ซึ่งโมเดลที่ดีที่สุดสำหรับการพัฒนาโมเดลในงานวิจัยนี้ คือ เทคนิคต้นไม้ตัดสินใจ ดังรูปที่ (2)



รูปที่ 2 แสดงผลลัพธ์การวิเคราะห์คำหยาบภาษาไทยด้วยเทคนิคต้นไม้ตัดสินใจ โดยใช้โปรแกรม Rapidminer

จากรูปที่ 2 การสร้างโมเดลเทคนิคต้นไม้ตัดสินใจ สามารถยกตัวอย่างได้ตามรูปแบบดังนี้

IF ควย = yes Type = Negative Then

Else IF แตด = yes Type = Negative Then

Else IF แม่ง = yes Type = Negative Then

Else IF แมร่ง = yes Type = Negative Then

Else IF กวนตีน = yes Type = Negative Then

Else IF ตอแหล = yes Type = Negative Then

Else IF ฟาย = yes Type = Negative Then

Else IF แดก = yes Type = Negative Then

Else IF แดรก = yes Type = Negative Then

Else IF ปากหมา = yes Type = Negative Then

Else IF สัส = yes Type = Negative Then

Else IF เสือก = yes Type = Negative Then

Else Type = Positive

ดังตัวอย่างวิธีการระบุมาตราจากกิ่งในต้นไม้ตัดสินใจด้วยกฎ IF Then ดังนี้

กฎข้อที่ 1 ถ้ามีคำว่า “ควย” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 2 ถ้ามีคำว่า “แตด” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 3 ถ้ามีคำว่า “แม่ง” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 4 ถ้ามีคำว่า “แมร่ง” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 5 ถ้ามีคำว่า “กวนตีน” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 6 ถ้ามีคำว่า “ตอแหล” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 7 ถ้ามีคำว่า “ฟาย” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 8 ถ้ามีคำว่า “แดก” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 9 ถ้ามีคำว่า “แดรก” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 10 ถ้ามีคำว่า “ปากหมา” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 11 ถ้ามีคำว่า “สัส” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ

กฎข้อที่ 12 ถ้ามีคำว่า “เสือก” จะเป็นผลทางด้านลบ หรือเป็นคำหยาบ ถ้าไม่มีคำเหล่านี้จะมีผลเป็นด้านบวก หรือไม่เป็นคำหยาบ

ผลการเปรียบเทียบประสิทธิภาพของโมเดลทั้ง 3 เทคนิค คือ รูปแบบโมเดลต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.96 ค่าคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) 0.19 ซึ่งมีประสิทธิภาพและรูปแบบที่ง่ายต่อความเข้าใจมากกว่า เทคนิค นาอีฟเบย์ และ เคเนียร์เรสเนเบอร์

5. บทสรุปและข้อเสนอแนะ

การศึกษาเปรียบเทียบกระบวนการวิเคราะห์คำหยาบภาษาไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์นิง โดยใช้โมเดลในการตรวจจับคำหยาบภาษาไทยด้วยพจนานุกรมคำหยาบที่ผ่านการปรับปรุงโดยใช้เทคนิค TFICF (Term Frequency - Inverse Class Frequency) งานวิจัยนี้ได้นำเสนอเทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียร์เรสเนเบอร์และเทคนิคนาอีฟเบย์ เพื่อนำมาสร้างโมเดลการวิเคราะห์คำหยาบเพื่อหาโมเดลที่เหมาะสมที่สุดไปเป็นแนวทางการตรวจจับคำหยาบให้แม่นยำขึ้น จากการทดลองพบว่า โมเดลที่พัฒนาด้วยเทคนิคต้นไม้ตัดสินใจนั้นให้ประสิทธิภาพในการวิเคราะห์คำหยาบได้ค่าความถูกต้อง 0.96 และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง 0.19 ซึ่งถือว่าเป็นโมเดลที่ดีที่สุดเมื่อเปรียบเทียบกับเทคนิค นาอีฟเบย์ และเคเนียร์เรสเนเบอร์

งานวิจัยที่เกี่ยวกับการตรวจจับคำหยาบภาษาไทย ในอนาคตอาจเป็นแนวทางที่ดีในการพัฒนาโมเดลวิเคราะห์คำหยาบเชิงลึก หรือคำที่ยากต่อการตรวจจับ เพื่อสามารถนำไปพัฒนาโปรแกรมตรวจจับคำหยาบที่แม่นยำต่อไป

6. กิตติกรรมประกาศ

ขอขอบคุณทีมปฏิบัติการวิจัยวิทยาการมนุษยภาษา (HLT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์ และคอมพิวเตอร์แห่งชาติ (เนคเทค) สำหรับข้อมูลพจนานุกรมคำหยาบภาษาไทย และขอขอบคุณ ผศ.ดร.สมชาย เล็กเจริญ สำหรับคำแนะนำ และคำปรึกษา ตลอดจนวิธีการในดำเนินงานศึกษาค้นคว้างานวิจัยในครั้งนี้จนสำเร็จลุล่วงไปได้ด้วยดี

เอกสารอ้างอิง

ณัฐาศิริ เชาว์ประสิทธิ์ กฤษฎา ศรีแผ้ว. 2557. โมเดลตรวจจับคำหยาบภาษาไทยด้วยการปรับปรุงพจนานุกรม. เอกสารการประชุมวิชาการระดับชาติด้านเทคโนโลยีสารสนเทศครั้งที่ 6. 359-363.

ณัฐภัทร แก้วรัตนภัทร์ ปรีดาวรรณ เกษเมธีการุณและชนินทร์มโนชญากร. 2555. การเปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลเพื่อแทนค่าสูญหาย. เอกสารการประชุมทางวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่ 8. บทความวิจัยสืบค้นจาก http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20142810144316.pdf

ศูนย์ข้อมูลกฎหมายและคดีเสรีภาพโดยไอลอว์. “สถิติการปิดกั้นเว็บไซต์ในประเทศไทยตั้งแต่ปี 2556-2557 สืบค้นจาก [เว็บบล็อก] สืบค้นจาก https://freedom.ilaw.or.th/blog/webblockstat20132014

Thai Lexitron (Lexto), [ข้อมูลอิเล็กทรอนิกส์] สืบค้นจาก : http://lexitron.nectec.or.th

Sood, Sara Owsley, Judd Antin, and Elizabeth Churchill. (2012). Profanity Use in Online Communities. In proceedings of ACM SIGCHI.

Xiang, G., Hong, J., and Rose, C. P. (2012). Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. Proceedings of The 21st ACM Conference on Information and Knowledge Management, Sheraton, Maui Hawaii.

Tanasanti Jirapon, Phokharatkul Pisit, Buntilov Vladimir, and Kanoksilpatham Budsaba. (2012). Thai insult Detection system based on linguistic features analysis. Proceedings: the 6th Symposium on Advances in Science and Technology, Kuala Lumpur, Malaysia.

C. Haruechaiyasak, S. Kongyoung and M. Dailey. (2008). A comparative study on Thai word segmentation approaches. Processing of the ECTI-CON.

Y. Poovarawan.(2529) "Thai Syllable Separator by Dictionary". Electrical Engineering Conference.