การจำแนกข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์

นายรัชกฤต อารีราษฎร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

DEFAMATORY TEXT CLASSIFICATION ON ONLINE SOCIAL MEDIA

Mr. Ratchakrit Arreerard

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

หัวข้อวิทยานิพนธ์	การจำแนกข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์
โดย	นายรัชกฤต อารีราษฎร์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.วีระ บุญจริง)

รัชกฤต อารีราษฎร์ : การจำแนกข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์. (DEFAMATORY TEXT CLASSIFICATION ON ONLINE SOCIAL MEDIA) อ.ที่ปรึกษาหลัก : รศ. ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา

การสื่อสารผ่านสื่อสังคมออนไลน์ในปัจจุบันเป็นที่นิยมกันอย่างแพร่หลาย การแสดงความคิดเห็นหรือแบ่งปันข้อมูลที่มีเนื้อหาก้าวร้าว โจมตี หรือดูหมิ่นผู้ใช้งานคนอื่นบนสื่อสังคมออนไลน์ อาจส่งผลกระทบต่อสังคมในด้านลบ โดยเนื้อหาดังกล่าวอาจผิดกฎหมายอาญาหมวด 3 ความผิดฐานหมิ่นประมาท มาตรา 326 วิทยานิพนธ์นี้เสนอคุณลักษณะเพื่อใช้ในการจำแนกข้อมูลเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์ด้วยขั้นตอนวิธีเพอเซ็ปตรอนหลายชั้น ซับพอร์ทเวคเตอร์แมชชีน และการถดถอยโลจิสติกส์ โดยเปรียบเทียบประสิทธิภาพแต่ละขั้นตอนวิธี ซึ่งการทดลองพบว่าเอ็น-แกรม คลังคำศัพท์จากศาลฎีกา และโครงสร้างไวยากรณ์แบบขึ้นต่อกันเป็นคุณลักษณะที่สามารถใช้ในการจำแนกข้อความหมิ่นประมาทได้โดยได้ค่าความเที่ยงสูง แต่ค่าเรียกคืนต่ำ แต่เมื่อมีการจัดการข้อมูลที่ไม่สมดุลด้วยแล้ว จะพบว่าการจำแนกมีประสิทธิภาพดีขึ้นโดยที่ขั้นตอนวิธีเพอเซ็ปตรอนหลายชั้นมีความสามารถในการจำแนกได้ดีที่สุดโดยมีค่าความเที่ยงเป็น 0.93 ค่าเรียกคืนเป็น 0.98 และค่าเอฟวันเป็น 0.95 นอกจากนี้จำนวนมิติของเอ็น-แกรมมีผลต่อประสิทธิภาพของการจำแนกข้อความ โดยจำนวนมิติที่เหมาะสมของเอ็น-แกรมแต่ละชนิดขึ้นอยู่กับขั้นตอนวิธีที่ใช้

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....
ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

5970292621 : MAJOR COMPUTER SCIENCE

KEYWORD: machine learning, support vector machine, logistic regression, text classification, social media, defamatory text, Multi Layer Perceptron

Ratchakrit Arreerard : DEFAMATORY TEXT CLASSIFICATION ON ONLINE SOCIAL MEDIA. Advisor: Assoc. Prof. Twittie Senivongse, Ph.D.

Communication on online social media is popular nowadays. Expressing opinions and sharing information with offensive or defamatory contents that target other social media users may have negative societal impact. The contents may violate the criminal code, Chapter 3 Offence of Defamation, Section 326. In this thesis, features are proposed to classify defamatory text on online social media with machine learning algorithms, i.e. multi-layer perceptron, support vector machine, and logistic regression. The performance of these algorithms are compared. The experiment reveals that n-grams, dictionary of judgment terms, and dependency structure of sentence are features that can be used to classify defamatory text, yielding high precision but low recall. After the imbalanced data problem is handled, performance of the classifiers improves substantially. In particular, multi-layer perceptron has the best performance with precision of 0.93, recall of 0.98, and F1 of 0.95. Moreover, the number of n-grams dimension affects performance of classification. The best number of dimension for each type of n-grams dimension varies by the algorithms used.

Field of Study: Computer Science Student's Signature .....

Academic Year: 2018 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์นี้จะสำเร็จลุล่วงไม่ได้หากปราศจากความกรุณาจาก รองศาสตราจารย์ ดร. ทวิตีย์ เสนีวงศ์ ณ อยุธยาอาจารย์ที่ปรึกษา ผู้ให้คำแนะนำสำหรับแนวทางการทำงานวิจัย และเป็นผู้ให้คำแนะนำในด้านวิชาการและอื่นๆ รวมถึงเป็นผู้ตรวจทานแก้ไขวิทยานิพนธ์ฉบับนี้ให้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ และ รองศาสตราจารย์ ดร. วีระ บุญจริง ผู้ให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ และชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้ดียิ่งขึ้น

ขอขอบพระคุณ ดร.สุขแสง คูกนก ดร.สวียา สุรมณี และ อ.พิทักษ์ ธรรมะ ที่ให้ความกรุณาตรวจสอบข้อความเข้าข่ายหมิ่นประมาท และ ขอขอบพระคุณ ผศ.ดร.ชมพูนุช เมฆเมืองทอง ผศ.ดร.ธรงวิทย์ ทองเสี่ยน และ อ.วินัย แสงกล้า ที่ให้ความกรุณาตรวจสอบชนิดของข้อความ

สุดท้ายนี้ขอขอบคุณสมาชิกครอบครัวทุกคน ที่คอยเป็นกำลังใจและสนับสนุนในทุกด้านจนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

รัชกฤต อารีราษฎร์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การสื่อสารผ่านสื่อสังคมออนไลน์ (Online social media) บนอินเทอร์เน็ตในปัจจุบัน เป็นที่นิยมกันอย่างแพร่หลาย เนื่องจากสื่อสังคมออนไลน์ช่วยให้การติดต่อสื่อสาร หรือการรับส่งข้อมูลในระยะทางไกลมีความสะดวกและรวดเร็วมากยิ่งขึ้น ด้วยรูปแบบที่ง่ายต่อการใช้งานผ่านอุปกรณ์อิเล็กทรอนิกส์ จึงส่งผลให้อัตราการใช้งานเพิ่มสูงขึ้นอย่างรวดเร็ว โดยผู้ใช้งานสามารถพูดคุย แบ่งปัน แลกเปลี่ยนความคิดเห็น ความรู้ หรือสิ่งที่ตัวเองสนใจได้อย่างไม่มีขีดจำกัด

ในทางกลับกันการเติบโตและการขยายตัวของสื่อสังคมออนไลน์ ส่งผลกระทบต่อสังคมในเชิงลบได้เช่นกัน เช่น ด้านธุรกิจ การศึกษา หรือต่อสังคม เป็นต้น เนื่องจากการใช้งานสื่อสังคมออนไลน์ผู้ใช้งานสามารถเลือกที่จะเปิดเผยชื่อข้อมูลจริงของผู้ใช้งาน หรือปกปิดชื่อและข้อมูลจริงโดยใช้ข้อมูลปลอมก็ได้ ทำให้ผู้ใช้งานขาดความระมัดระวังในการสื่อสาร การแสดงความคิดเห็น การแบ่งปันข้อมูล ซึ่งเนื้อหาของข้อมูลอาจมีเนื้อหาที่ก้าวร้าว โจมตี หรือดูหมิ่นผู้ใช้งานคนอื่นบนสื่อสังคมออนไลน์

จากประมวลกฎหมายอาญาหมวด 3 ความผิดฐานหมิ่นประมาท มาตรา 326 “ผู้ใดใส่ความผู้อื่นต่อบุคคลที่สาม โดยประการที่น่าจะทำให้ผู้อื่นเสื่อมเสียชื่อเสียง ถูกดูหมิ่น หรือถูกเกลียดชัง ผู้นั้นกระทำความผิดฐานหมิ่นประมาท” และมาตรา 328 “ถ้าความผิดฐานหมิ่นประมาทได้กระทำโดยการโฆษณา ด้วยเอกสาร ภาพวาด ภาพระบายสี ภาพยนตร์ ภาพหรือตัวอักษรที่ ทำให้ปรากฏด้วยวิธีใด ๆ แผ่นเสียง หรือสิ่งบันทึกเสียง บันทึกภาพ หรือบันทึกอักษร กระทำโดยการกระจายเสียง หรือการกระจายภาพ หรือโดยกระทำการป่าวประกาศด้วยวิธีอื่น” ผู้ที่ถูกดูหมิ่นสามารถฟ้องเพื่อเอาผิดผู้ที่เผยแพร่ข้อมูลดังกล่าวได้ แต่ผู้ใช้งานสื่อสังคมออนไลน์กลับไม่มีความระมัดระวัง และอาจไม่ได้ตระหนักถึงความผิดตามประมวลกฎหมายอาญา ทำให้ข้อความในลักษณะดังกล่าวยังสามารถพบเห็นได้ทั่วไป บนสื่อสังคมออนไลน์ต่างๆ ไม่ว่าจะเป็น เฟซบุ๊ก (Facebook) หรือ ทวิตเตอร์ (Twitter) เป็นต้น ในทางกลับกันการดำเนินคดีทางกฎหมายเป็นกระบวนการที่ใช้เวลา และทรัพยากรเป็นอย่างมาก ไม่ว่าจะเป็นการรวบรวมหลักฐาน การจ้างทนาย การขึ้นศาล จากการเก็บข้อมูลคำพิพากษาศาลฎีกาพบว่าบางคดีถูกยกฟ้องจากความเข้าใจผิดเนื่องจากความกำกวมของคำ หรือการแปลความหมายของประมวลกฎหมายฐานหมิ่นประมาท

จากปัญหาดังกล่าว อาจทำให้สื่อสังคมออนไลน์ถูกใช้เป็นเครื่องมือกระทำความผิดฐานหมิ่นประมาทผู้อื่น ผู้วิจัยเห็นความสำคัญในการตรวจสอบข้อความที่ผู้ใช้แสดงความคิดเห็นบนสื่อสังคมออนไลน์ โดยใช้หลักการการประมวลผลภาษาธรรมชาติ (Natural language processing - NLP)

และการเรียนรู้ของเครื่อง (Machine learning) ซึ่งเป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial intelligence) เพื่อสร้างขั้นตอนวิธีเพื่อจำแนกประเภทข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์ ออกจากข้อความทั่วไป ข้อความที่เข้าข่ายหมิ่นประมาทหมายถึง ข้อความที่มีองค์ประกอบครบตามความผิดฐานหมิ่นประมาทและถูกส่งต่อเข้าสู่กระบวนการยุติธรรมเพื่อการตัดสินความผิดต่อไป ซึ่งการจำแนกข้อความช่วยให้ผู้ใช้งานสื่อสังคมออนไลน์ตระหนักถึงผลกระทบที่อาจเกิดขึ้นต่อตนเองและผู้อื่นจากการแสดงความคิดเห็นผ่านการใช้งานสื่อสังคมออนไลน์ Ikonomakis et al. (2005) [1] ศึกษาการใช้ขั้นตอนวิธีการเรียนรู้ของเครื่องในการจำแนกประเภทข้อความ ด้วยขั้นตอนวิธีต่างๆ การเลือกคุณลักษณะ (Feature) ของข้อความเพื่อใช้ในการจำแนกข้อความ ส่งผลต่อประสิทธิภาพของการจำแนกประเภทอย่างมาก

แบบสำรวจการจำแนกข้อความ Aggarwal, C. C., & Zhai, C. (2012) [2] พบว่าขั้นตอนวิธีที่ใช้ในการจำแนกข้อความที่เป็นที่นิยมได้แก่ ต้นไม้การตัดสินใจ (Decision tree) การจำแนกข้อมูลด้วยกฎ (Rule-based classifiers) ซับพอร์ทเวคเตอร์แมชชีน (Support Vector Machine - SVM) โครงข่ายประสาทเทียม (Neural network) การเรียนรู้เบย์อย่างง่าย (Naïve Bayes classification) และการจำแนกด้วยการถดถอย (Regression-based classifiers) ซึ่งต้นไม้การตัดสินใจและการจำแนกข้อมูลด้วยกฎมีความคล้ายคลึงกัน ขั้นตอนวิธีจะสร้างกฎหรือเงื่อนไขจากข้อความเพื่อใช้ในการจำแนกข้อความ ในขณะที่ ซับพอร์ทเวคเตอร์แมชชีน โครงข่ายประสาทเทียม และการจำแนกด้วยการถดถอยถูกจัดอยู่ในหมวดหมู่การจำแนกเชิงเส้น (Linear classifiers) ส่วนการเรียนรู้เบย์อย่างง่ายนั้นจำแนกข้อความด้วยความน่าจะเป็น

การเรียนรู้เบย์อย่างง่ายไม่เหมาะสมกับการทดลองเนื่องจากคุณลักษณะที่ใช้ในการทดลองมีค่าเป็นจำนวนจริงซึ่งไม่เหมาะสำหรับการเรียนรู้เบย์อย่างง่าย ในขณะที่จำนวนข้อความที่ใช้ในการทดลองมีปัญหาความไม่สมดุล คือ จำนวนข้อความเข้าข่ายหมิ่นประมาทมีจำนวนน้อยกว่าข้อความไม่เข้าข่ายหมิ่นประมาท ทำให้ต้นไม้การตัดสินใจ และการจำแนกข้อมูลด้วยกฎ ไม่สามารถสร้างกฎ เพื่อใช้ในการจำแนกข้อความได้ ซึ่งคุณลักษณะและจำนวนข้อมูลที่ใช้ในการทดลองจะถูกอธิบายในบทที่ 4

จากขั้นตอนวิธีที่กล่าวมาข้างต้น ผู้วิจัยเลือกใช้ขั้นตอนวิธีซับพอร์ทเวคเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น (Multi-layer perceptron) และการถดถอยโลจิสติกส์ (Logistic regression) โดยที่เพอร์เซ็ปตรอนหลายชั้นและการถดถอยโลจิสติกส์คือขั้นตอนวิธีจำแนกข้อความซึ่งมีพื้นฐานมาจากโครงข่ายประสาทเทียมและการจำแนกด้วยการถดถอยตามลำดับ

1.2 วัตถุประสงค์

1.2.1 เพื่อพัฒนาวิธีการสำหรับการจำแนกประเภทข้อความที่เข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์

1.2.2 เพื่อทดสอบประสิทธิภาพและเปรียบเทียบวิธีการสำหรับการจำแนกประเภทข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์

1.3 ขอบเขตการดำเนินงาน

1.3.1 ข้อมูลจากเฟซบุ๊ก เป็นความคิดเห็นเกี่ยวข้องกับบุคคลทางการเมือง บุคคลในหน่วยงานราชการ ตัวอย่างเช่น นายกรัฐมนตรี พรรคการเมือง ตำรวจ ทหาร เป็นต้น

1.3.2 ข้อมูลคำพิพากษาศาลฎีกาที่ใช้ในการสร้างคลังคำศัพท์เป็นข้อความซึ่งถูกฟ้องในมาตราที่ 326 ตั้งแต่ ปี พ.ศ. 2503 ถึง พ.ศ.2557

1.3.3 วิธีที่ใช้สร้างแบบจำลองเพื่อจำแนกข้อความเข้าข่ายหมิ่นประมาท ได้แก่ ขั้นตอนวิธีเพอร์เซ็ปตรอนหลายชั้น ซับพอร์ทเวคเตอร์แมชชีน และการถดถอยโลจิสติกส์

1.3.4 การสร้างต้นไม้ทางไวยากรณ์แบบขึ้นต่อกัน ประยุกต์ใช้ตามงานวิจัย [3] ร่วมกับข้อมูล Universal dependency data [4] สำหรับภาษาไทย

1.3.5 การตรวจหาชื่อเฉพาะใช้ Polyglot [5] ซึ่งเป็นโปรแกรมสำเร็จที่ใช้สำหรับตรวจหาชื่อเฉพาะในภาษา Python

1.3.6 ขั้นตอนวิธีการเรียนรู้ของเครื่อง ใช้โปรแกรมสำเร็จ Scikit-learn [6] ซึ่งใช้งานด้วยภาษา Python

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ได้วิธีการสำหรับการจำแนกประเภทข้อความที่เข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์ที่มีประสิทธิภาพ

1.5 ขั้นตอนการดำเนินงาน

1.5.1 ศึกษาองค์ความรู้และค้นคว้างานวิจัยต่าง ๆ ที่เกี่ยวข้อง

1.5.2 ออกแบบวิธีการสร้างคุณลักษณะข้อความ

1.5.3 จัดเตรียมข้อมูลการวิจัย

1.5.4 ทำการทดลอง และวัดประสิทธิภาพ

1.5.5 สรุปผลการทดลอง

1.5.6 จัดทำรูปเล่มวิทยานิพนธ์

1.6 ผลงานวิจัยที่ได้รับการตีพิมพ์

R. Arreerard and T. Senivongse, “Thai Defamatory Text Classification on Social Media,” in Third IEEE/ACIS International Conference on Big Data Cloud Computing and Data Science Engineering, BCD 2018, Japan, from 10 – 18 July 2018.

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 หลักประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาท

สหรัฐ กิติ ศุภการ (2017) [7] อธิบายว่า มาตรา 326 “ผู้ใดใส่ความผู้อื่นต่อบุคคลที่สาม โดยประการที่น่าจะทำให้ผู้อื่นเสื่อมเสียชื่อเสียง ถูกดูหมิ่น หรือถูกเกลียดชัง ผู้นั้นกระทำความผิดฐานหมิ่นประมาท” มีองค์ประกอบทั้งหมด 5 องค์ประกอบคือ

2.1.1 ผู้กระทำ “ผู้ใด” อาจเป็นนิติบุคคล เช่น นาย ก. กรรมการบริษัท ข. แถลงข่าวในนามบริษัท ข. หมิ่นประมาทนาย ค. ถือว่า บริษัท ข. หมิ่นประมาทนาย ค. นอกจากนี้การแจ้งความเท็จยังสามารถเป็นความผิดฐานหมิ่นประมาทได้เช่นกัน ถ้าหากความเท็จนั้นเป็นการใส่ความผู้อื่นผู้ถูกใส่ความ

2.1.2 “ผู้อื่น” คือบุคคลธรรมดาหรือนิติบุคคล เช่น นาย ก. หมิ่นประมาทบริษัทที่มีฐานะเป็นนิติบุคคล เช่นหมิ่นประมาทกองทัพบก หมิ่นประมาทสำนักงานตำรวจแห่งชาติ ในกรณีหมิ่นประมาทบุคคลหลายบุคคล ไม่สามารถเจาะจง หรือเข้าใจได้ว่า บุคคลดังกล่าวเป็นใคร ก็ไม่อาจมีผู้ถูกใส่ความได้ เช่น ฎีกาที่ 295/2505 “ทนายความเมืองร้อยเอ็ดคบไม่ได้ เป็นนกสองหัว เหยียบเรือสองแคม เป็นมวยล้ม ว่าความทีแรกดีได้รับเงินแล้วก็ว่าเป็นอย่างอื่น” เป็นการกล่าวถึงทนายความเมืองร้อยเอ็ด ซึ่งทนายความมี 10 คน การไม่เจาะจงบุคคล ทำให้ไม่สามารถหาผู้ถูกใส่ความได้ จึงไม่มีความผิดฐานหมิ่นประมาท

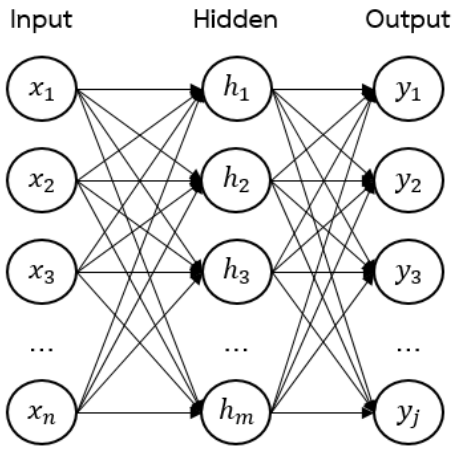
2.1.3 บุคคลที่สาม เป็นใครก็ได้แต่ต้องไม่ใช่ตัวการร่วมกระทำความผิด นอกจากนี้บุคคลที่สามอาจจะอยู่คนละแห่งกับผู้ถูกใส่ความและบุคคลที่สามอาจเป็นคนๆ เดียวก็ได้ไม่จำเป็นต้องมีหลายคน

2.1.4 การใส่ความ การพูดเหตุร้ายหรือกล่าวหาเรื่องร้ายให้ผู้อื่นได้รับความเสียหาย อาจกระทำด้วยวาจา ลายลักษณ์อักษร กิริยาท่าทาง หรือพฤติการณ์อย่างอื่น แต่จะต้องมีลักษณะยืนยันข้อเท็จจริงให้คนอ่าน คนฟัง คนเห็นเชื่อ แล้วเกิดความรู้สึกดูหมิ่น เกลียดชัง

2.1.5 การยืนยันข้อเท็จจริง การกล่าวข้อเท็จจริง แล้วยืนยันข้อเท็จจริงนั้น เช่น ฎีกาที่ 1034/2533 กล่าวข้อเท็จจริงว่า “โจทก์หากินกับตำรวจในการเรียกรับสินบน” ในขณะที่การกล่าวข้อเท็จจริงแล้วไม่ยืนยันข้อเท็จจริง เช่น ฎีกาที่ 2180,2155/2531 “การที่จำเลยถามว่า ป. มีความสัมพันธ์ชู้สาวกับโจทย์หรือไม่” ไม่เป็นการยืนยันข้อเท็จจริง เป็นเพียงการคาดคะเนของจำเลย จึงไม่เป็นความผิดฐานหมิ่นประมาท

2.2 เพอร์เซ็ปตรอนหลายชั้น (Multi-layer perceptron)

เพอร์เซ็ปตรอนหลายชั้น [8] คือ รูปแบบหนึ่งของโครงข่ายประสาทเทียม ที่เลียนแบบวิธีการทำงานของสมองมนุษย์เพื่อใช้ในการจำแนกข้อมูล เพอร์เซ็ปตรอนหลายชั้นดังรูปที่ 2.1



รูปที่ 2.1 เพอร์เซ็ปตรอนหลายชั้น

จากรูปที่ 2.1 แสดงถึงรูปแบบการทำงานของเพอร์เซ็ปตรอนหลายชั้น เริ่มจากแบบจำลองได้รับข้อมูลนำเข้า (Input) x_i จากนั้นแบบจำลองทำการคำนวณข้อมูลและเก็บไว้ที่ชั้นซ่อน (Hidden layer) h_i ชั้นซ่อนของเพอร์เซ็ปตรอนหลายชั้นมีได้มากกว่า 1 ชั้น หากชั้นซ่อนมีมากกว่า 1 ชั้นการคำนวณข้อมูลจะใช้ข้อมูลของชั้นซ่อนก่อนหน้าและส่งต่อไปเรื่อยๆจนได้แบบจำลองข้อมูลส่งออก (Output) y_i ซึ่งการคำนวณนี้สามารถทำได้ตามสมการดังต่อไปนี้

$$z = (\sum_i w_i x_i) \quad (1)$$

เมื่อ w_i คือ ค่าน้ำหนัก (Weight)

x_i คือ ข้อมูลจากโหนดชั้นก่อนหน้า

$$g(z) = \frac{1}{1+e^{-z}} \quad (2)$$

จากรูปที่ 2.1 ลูกศรแต่ละเส้น ในแต่ละชั้นมีค่าน้ำหนักเป็นของตัวเองซึ่งแทนด้วย w ซึ่งค่าน้ำหนักถูกคำนวณร่วมกับข้อมูลในชั้นก่อนหน้า และนำผลลัพธ์มาแปลงให้อยู่ในช่วง 0 ถึง 1 ด้วยสมการที่ 2

จากสมการที่ 1 และ 2 แสดงถึงการทำงานของเพอร์เซ็ปตรอนหลายชั้น ซึ่งการเรียนรู้ของเพอร์เซ็ปตรอนหลายชั้น คือ การใช้ข้อมูลนำเข้าเพื่อเรียนรู้ และปรับค่าน้ำหนักของแบบจำลองเพื่อให้แบบจำลองสามารถจำแนกข้อมูลได้อย่างถูกต้อง ซึ่งการเรียนรู้ของแบบจำลองใช้วิธีการหาค่าความผิดพลาด (E) เพื่อใช้ในการปรับค่าน้ำหนักให้แบบจำลองตามสมการต่อไปนี้

$$E = \frac{1}{2}(t-y)^2 \tag{3}$$

เมื่อ t คือ ข้อมูลเป้าหมาย (Target output)

y คือ ข้อมูลส่งออกที่แบบจำลองคำนวณได้

จากสมการที่ 3 เมื่อคำนวณค่าความผิดพลาดแล้ว แบบจำลองจึงทำการปรับค่าน้ำหนักตามสมการต่อไปนี้

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} = \delta_j y_i \tag{4}$$

เมื่อ w_{ij} คือ ค่าน้ำหนักจากโหนด i ไปโหนด j

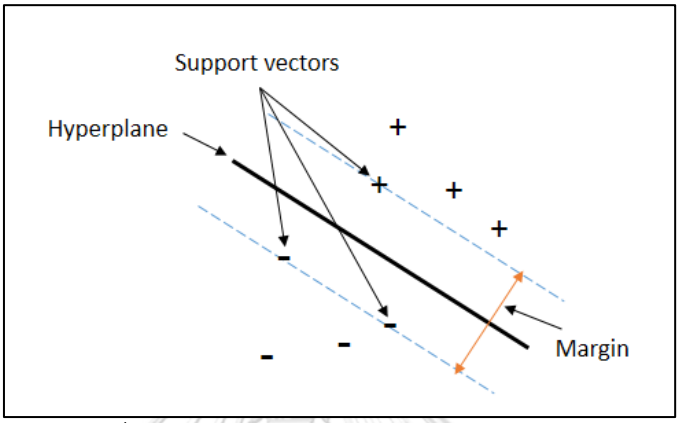
$$\delta_j = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j} = \begin{cases} (o_j - t_j)o_j(1-o_j) & \text{เมื่อ } j \text{ คือข้อมูลส่งออกที่แบบจำลองคำนวณ} \\ \left(\sum_{l\epsilon L} w_{jl}\delta_l\right)o_j(1-o_j) & \text{เมื่อ } j \text{ คือโหนดภายในชั้นของเพอร์เซ็ปตรอนหลายชั้น} \end{cases} \tag{5}$$

$$\Delta w_{ij} = -\eta\frac{\partial E}{\partial w_{ij}} = -\eta\delta_j y_i \tag{6}$$

เมื่อ η คือ ค่าเรียนรู้ (Learning rate)

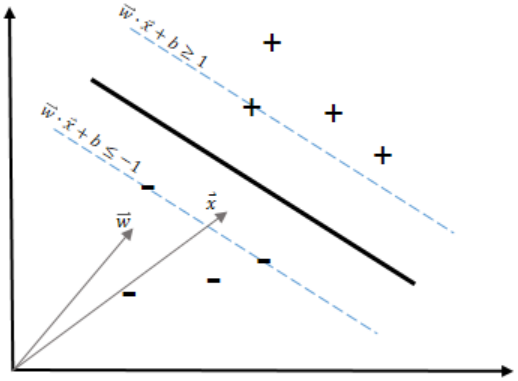
2.3 ซับพอร์ทเวคเตอร์แมชชีน (Support Vector Machine - SVM)

SVM [9] คือ ขั้นตอนวิธีที่ใช้จำแนกประเภทของข้อมูล โดยแทนค่าข้อมูลด้วยเวคเตอร์ขนาด n มิติ เมื่อ n คือจำนวนคุณลักษณะของข้อมูล SVM ทำการสร้างเส้นแบ่ง (Hyperplane) ระหว่างข้อมูลทั้ง 2 ประเภทเมื่อพิจารณาการจำแนกประเภทแบบทวิภาค (Binary) ดังรูปที่ 2.2



รูปที่ 2.2 ตัวอย่างการจำแนกประเภทของ SVM

เนื่องจากการจำแนกด้วยขั้นตอนวิธี SVM สามารถสร้างเส้นแบ่งระหว่างประเภทของข้อมูลได้หลายวิธี SVM จะเลือกเส้นแบ่งที่มีความกว้างระหว่างประเภทของข้อมูลมากที่สุด หรือเรียกว่า มาร์จิน (Margin) โดยข้อมูลที่อยู่บนขอบของมาร์จิน เรียกว่า ซับพอร์ทเวคเตอร์ (Support vectors) เส้นแบ่งของ SVM นิยามด้วยค่าคงที่ b และ เวคเตอร์ $\vec{w}$ เมื่อ $\vec{w}$ คือเวคเตอร์ที่ตั้งฉากกับเส้นแบ่ง ดังรูปที่ 2.3



รูปที่ 2.3 มาร์จินของ SVM

จากรูปที่ 2.3 เมื่อ $\vec{x}$ คือเวคเตอร์ของข้อมูลใดๆที่ต้องการจำแนก การจำแนกด้วย SVM ใช้การหาผลคูณของเวคเตอร์ $\vec{x}$ และ $\vec{w}$ หากผลลัพธ์มากกว่าหรือเท่ากับ 1 ข้อมูลจะถูกจำแนกเป็นประเภทบวก ในขณะที่หากผลลัพธ์น้อยกว่าหรือเท่ากับ -1 ข้อมูลจะถูกจำแนกเป็นตัวอย่างประเภทลบดังสมการต่อไปนี้

$$y = \begin{cases} +1 \ for + Samples \\ -1 \ for - Samples \end{cases} \tag{7}$$

$$y_i(\overrightarrow{x_i} \cdot \vec{w} + b) - 1 \geq 0 \tag{8}$$

เมื่อ y คือ ตัวแปรที่ใช้กำหนดสัญลักษณ์ของประเภทข้อมูล

การหาเส้นแบ่งที่มีความกว้างมากที่สุด สามารถทำได้โดยการหาผลต่างของเวคเตอร์ข้อมูลที่อยู่บนขอบของมาร์จิน ซึ่งจากสมการที่ 8 เวคเตอร์ที่อยู่บนขอบของมาร์จินจะมีค่าเท่ากับ 0 แสดงได้ดังสมการต่อไปนี้

$$width = (x_+ - x_-) \cdot \frac{\vec{w}}{||\vec{w}||} = \frac{2}{||\vec{w}||} \tag{9}$$

เมื่อ $\vec{w}$ คือเวคเตอร์ที่ตั้งฉากกับเส้นแบ่ง ผลต่างของเวคเตอร์ที่อยู่บนขอบของมาร์จินจึงถูกคูณด้วยเวคเตอร์ $\frac{\vec{w}}{||\vec{w}||}$ เพื่อให้ทิศทางของผลต่างมีทิศทางเดียวกันกับเส้นแบ่ง จากสมการที่ 8 แทนค่าในสมการที่ 9 จึงได้ว่า ความกว้างของมาร์จิน คือ $\frac{2}{||\vec{w}||}$ เพราะฉะนั้นการหาความกว้างของมาร์จินที่มากที่สุดคือการหาค่าที่น้อยที่สุดของ $\left|\left|\vec{w}\right|\right|$ ซึ่งการหาค่าความกว้างที่มากที่สุดของมาร์จิน ใช้ตัวคูณลากรองจ์ (Lagrange multiplier) ในการแก้ปัญหา โดยหาค่า α_i ใดๆที่เกี่ยวข้องกับข้อมูล $y_i(\overrightarrow{x_i} \cdot \vec{w} + b) - 1 \geq 0$ ดังสมการต่อไปนี้

$$L = \sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \overrightarrow{x_i} \overrightarrow{x_j} \tag{10}$$

เมื่อ $\sum_i \alpha_i y_i = 0$ และ $\alpha_i \geq 0$ สำหรับ $1 \leq i \leq n$

จากการใช้ตัวคูณลากรองจ์ ทำให้ได้ค่าของ $\overrightarrow{w}$ ซึ่งเท่ากับ $\overrightarrow{w} = \sum_i \alpha_i y_i \overrightarrow{x_i}$ เมื่อแทนค่า $\overrightarrow{w}$ กลับเข้าไปในสมการที่ 8 ทำให้ได้ฟังก์ชันในการจำแนกประเภทของข้อมูลโดยใช้ SVM ดังต่อไปนี้

$$f(\vec{x}) = sign(\sum_i \alpha_i y_i \overrightarrow{x_i} \vec{x} + b) \quad (11)$$

เมื่อ $sign$ คือ ฟังก์ชันเพื่อหาค่าเครื่องหมายของจำนวน

2.4 การถดถอยโลจิสติกส์ (Logistic regression)

การถดถอยโลจิสติกส์ [10] คือ ขั้นตอนวิธีที่จำแนกประเภทข้อมูลด้วยการประมาณค่าความน่าจะเป็นของข้อมูลว่าเป็นประเภท 0 หรือ 1 เมื่อข้อมูลสามารถจำแนกได้เพียงแค่สองประเภทคือ 0 และ 1 จากการประมาณค่าความน่าจะเป็น ประมาณได้จากค่าอัตราส่วนความน่าจะเป็น (or) ตามสมการดังต่อไปนี้

$$or = \frac{p}{(1-p)} \quad (12)$$

เมื่อ p คือ ความน่าจะเป็นของข้อมูลประเภท 1

$(1-p)$ คือ ความน่าจะเป็นของข้อมูลประเภท 0

เนื่องจากความน่าจะเป็นมีค่าระหว่าง 0 และ 1 การประมาณค่าความน่าจะเป็นจึงจำเป็นต้องปรับค่าของสมการที่ 12 ให้มีค่าระหว่าง 0 ถึง 1 และใช้รูปแบบของการถดถอยโลจิสติกส์ตามสมการต่อไปนี้

$$\ln(or) = \ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (13)$$

เมื่อ β คือ สัมประสิทธิ์ของ x

x คือ ตัวแปรที่ใช้ในการประมาณค่าความน่าจะเป็นในการจำแนกข้อมูล หรือ คุณลักษณะของข้อมูลซึ่งมีทั้งหมด n ตัว

การประมาณค่าความน่าจะเป็น (p) สามารถประมาณได้ตามสมการที่ 14 ซึ่งมาจากการแก้สมการที่ 13 ดังนี้

$$p = \frac{1}{1+e^{\beta x}} \tag{14}$$

เมื่อ $\beta x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

ซึ่งความน่าจะเป็นที่ประมาณออกมาจากสมการที่ 14 จะถูกใช้ในการจำแนกประเภทของข้อมูลว่าเป็นประเภท 1 หากความน่าจะเป็นมีค่ามากกว่าหรือเท่ากับ 0.5 และเป็นประเภท 0 หากค่าความน่าจะเป็นมีค่าน้อยกว่า 0.5

2.5 Synthetic Minority Over-sampling Technique (SMOT)

SMOT [11] คือ กระบวนการการจัดการกับประเภทข้อมูลที่มีจำนวนน้อย (Minority class) เมื่อจำนวนของประเภทข้อมูลมีความไม่สมดุลกัน (Imbalance data) โดย SMOT ใช้วิธีการสังเคราะห์ข้อมูลโดยการสร้างข้อมูลขึ้นมาใหม่ (Oversampling) จากคุณลักษณะของข้อมูล การสังเคราะห์ข้อมูลตามขั้นตอนวิธี SMOT จะใช้การหาเพื่อนบ้านที่ใกล้ที่สุดของข้อมูลที่มีจำนวนน้อย (K nearest neighbor) เพื่อทำการสร้างข้อมูลขึ้นมาใหม่ระหว่างข้อมูลสองจุด ตามสมการดังต่อไปนี้

$$dif = (x_2, y_2) - (x_1, y_1) \tag{15}$$

เมื่อ (x_2, y_2) คือ ข้อมูลเพื่อนบ้านที่ใกล้ที่สุดของข้อมูล (x_1, y_1)

$$gap = random(0,1) \tag{16}$$

เมื่อ $random(0,1)$ คือ ฟังก์ชันการสุ่มเลขจำนวนจริงระหว่าง 0 ถึง 1

$$(x, y) = (x_1, y_1) + gap * dif \tag{17}$$

เมื่อ (x, y) คือ ข้อมูลที่ถูกสร้างขึ้นมาใหม่โดย SMOT

จากสมการข้างต้น สามารถนำไปใช้ในการสร้างข้อมูลใหม่ได้ดังตารางที่ 2.1

ตารางที่ 2.1 ตัวอย่างของการสร้างข้อมูลใหม่

1	พิจารณาตัวอย่างข้อมูล (6,4) ซึ่งมีเพื่อนบ้านที่ใกล้ที่สุด คือ (4,3)
2	เพราะฉะนั้น
3	ค่า dif จะมีค่าเท่ากับ (4,3) - (6,4) = (-2,-1)
4	กำหนดให้ $random(0,1)$มีค่าเท่ากับ 0.5 จึงได้ว่า
5	ค่า gap มีค่าเท่ากับ 0.5
6	จากค่าของ gap และ dif จึงได้ข้อมูลใหม่
7	ข้อมูลใหม่ (x, y) จะถูกคำนวณดังนี้ (6,4) + 0.5*(-2,-1)
8	จึงได้ว่า ข้อมูลใหม่ (x, y) มีค่าเท่ากับ (5.0,3.5)

2.6 วิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition - SVD)

การแยกค่าแบบเดี่ยว หรือ SVD [12] สามารถใช้ในการลดมิติ (Dimension) ของแมทริกซ์ กล่าวคือ SVD สามารถลดจำนวนคุณลักษณะของข้อมูลที่ใช้ในการจำแนกประเภท เมื่อข้อมูลอยู่ในรูปของแมทริกซ์ A_{n*m} เมื่อ n คือจำนวนของข้อมูล และ m คือจำนวนของคุณลักษณะ ซึ่ง SVD ของ A_{n*m} สามารถหาได้ตามสมการดังต่อไปนี้

$$A_{n*m} = U_{n*n} S_{n*m} V^T_{m*m} \quad (18)$$

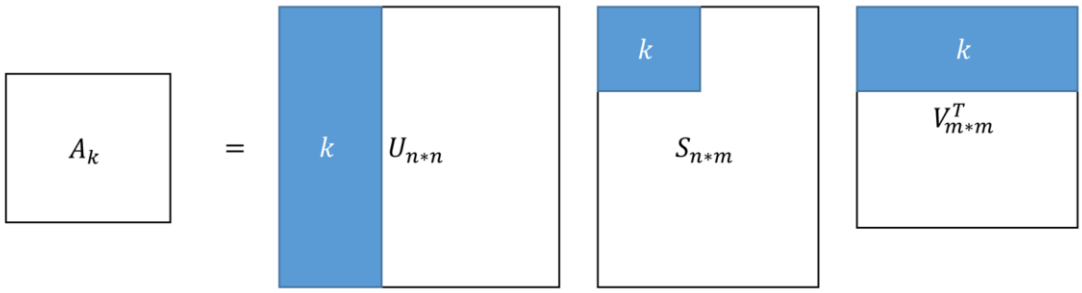
เมื่อ U คือ แมทริกซ์ตั้งฉากขนาด $n * n$

V^T คือ แมทริกซ์ตั้งฉากขนาด $m * m$

S คือ แมทริกซ์ทแยงมุมขนาด $n * m$ ซึ่งประกอบด้วยจำนวนจริงที่ไม่ใช่จำนวนลบ เรียกว่า ค่าแบบเดี่ยว (Singular value) และเรียงลำดับจากมากไปน้อย

จากสมการที่ 18 ค่าแบบเดี่ยวที่เรียงลำดับจากมากไปน้อย แสดงถึงค่าความแปรปรวน (Variance) ของการขึ้นต่อกันของแต่ละคุณลักษณะ กล่าวคือ หากค่าแบบเดี่ยวมีค่ามาก คุณลักษณะนั้นจะสามารถใช้ในการจำแนกข้อมูลได้ดี เพราะข้อมูลมีการกระจายตัว ไม่กระจุกอยู่ที่ใดที่หนึ่ง

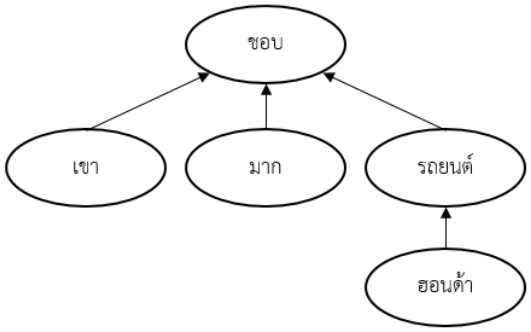
การลดจำนวนคุณลักษณะของข้อมูลสามารถทำได้โดยการเลือกจำนวนคุณลักษณะที่ต้องการทั้งหมด k ชนิด จากคุณลักษณะทั้งหมด m ชนิด ดังรูปที่ 2.4



รูปที่ 2.4 การเลือก k คุณลักษณะของแมทริกซ์ A

2.7 ต้นไม้ทางไวยากรณ์แบบขึ้นต่อกัน (Dependency tree)

ต้นไม้ทางไวยากรณ์แบบขึ้นต่อกัน [13] คือ ต้นไม้ที่แสดงความสัมพันธ์ของคำแต่ละคำในประโยค ว่ามีความสัมพันธ์แบบขึ้นต่อกันกับคำใดในประโยคดังรูปที่ 2.5



รูปที่ 2.5 ต้นไม้ทางไวยากรณ์แบบขึ้นต่อกันของประโยค “เขาชอบรถยนต์ฮอนด้ามาก”

จากรูปที่ 2.5 คือตัวอย่างของต้นไม้ทางไวยากรณ์แบบขึ้นต่อกันของประโยค คำว่า “ชอบ” หรือปม (Node) ที่อยู่บนสุดเรียกว่าปมราก (Root node) และลูกศรของต้นไม้แสดงถึงความสัมพันธ์แบบขึ้นต่อกันของคำ เช่น คำว่า “เขา” ขึ้นกับ คำว่า “ชอบ” เป็นต้น ซึ่งคำๆหนึ่งในประโยคสามารถขึ้นได้กับเพียงคำเดียวเท่านั้น กล่าวคือแต่ละคำจะมีเพียงหนึ่งลูกศรชี้ออก ยกเว้นเพียงปมรากซึ่งเป็นปมหลักของประโยค ความสัมพันธ์แบบขึ้นต่อกันของคำมีหลายชนิด จากรูปที่ 2.5 “เขา” และ “ชอบ” คือความสัมพันธ์ระหว่างประธานกับกริยา “รถยนต์” และ “ชอบ” คือความสัมพันธ์ระหว่างกริยากับกรรม ในขณะที่ “ฮอนด้า” และ “มาก” มีความสัมพันธ์เป็นส่วนขยายของคำ “รถยนต์” และ “ชอบ” ตามลำดับ

การสร้างต้นไม้ทางไวยากรณ์แบบขึ้นต่อกันในงานวิจัยนี้ได้ประยุกต์ใช้วิธีการสร้างต้นไม้ของ Tongchim และ คณะ [3] ซึ่งแบ่งออกเป็น 3 ส่วน คือ การหาปมราก การหาความสัมพันธ์ และการสร้างต้นไม้ทางไวยากรณ์แบบขึ้นต่อกัน

2.7.1 การหาปมรากของประโยคใดๆ สามารถหาได้โดยการใช้การเรียนรู้ของเครื่องด้วยวิธีการ SVM เพื่อหาความน่าจะเป็นของคำว่าคำใดมีความน่าจะเป็นปมรากมากที่สุด จากนั้นเลือกคำดังกล่าวเป็นปมราก ซึ่งแต่ละคำในประโยคจะถูกเก็บข้อมูลคุณลักษณะดังต่อไปนี้

1) ประเภทของคำ (Part of speech)

2) ตำแหน่งของคำในประโยค

3) จำนวนของคำกริยาทั้งหมด

4) จำนวนของประเภทของคำที่เหมือนกันข้างหน้าคำที่พิจารณา

5) จำนวนของประเภทของคำที่เหมือนกันข้างหลังคำที่พิจารณา

6) จำนวนของประเภทของคำที่อยู่หมวดเดียวกันข้างหน้าคำที่พิจารณา

7) จำนวนของประเภทของคำที่อยู่หมวดเดียวกันข้างหลังคำที่พิจารณา

2.7.2 การหาความสัมพันธ์ของคำ คือ การหาความน่าจะเป็นของคำ ว่าคำสองคำใดๆในประโยคมีความน่าจะเป็นที่มีความสัมพันธ์กันเท่าใด โดยใช้การเรียนรู้ของเครื่องด้วยวิธีการ SVM โดยทุกๆคำในประโยคจะถูกจับคู่เพื่อหาความน่าจะเป็น และสร้างแมทริกซ์ของความสัมพันธ์ ด้วยคุณลักษณะดังต่อไปนี้

1) ประเภทของคำที่หนึ่ง

2) ประเภทของคำที่สอง

3) ทิศทางของความสัมพันธ์ หากคำที่หนึ่งอยู่ทางซ้ายของคำที่สอง (0) หรือคำที่หนึ่งอยู่ทางขวาของคำที่สอง (1)

4) ระยะห่างระหว่างทั้งสองคำ

5) หมวดหมู่ของคำที่หนึ่ง คือเป็นคำไวยากรณ์ (Function words เช่น คำสรรพนาม, คำเชื่อม, คำบุพบท) หรือคำสำคัญ (Content words เช่น คำนาม, คำกริยา)

6) หมวดหมู่ของคำที่สอง คือเป็นคำไวยากรณ์ (Function words เช่น คำสรรพนาม, คำเชื่อม, คำบุพบท) หรือคำสำคัญ (Content words เช่น คำนาม, คำกริยา)

7) หมวดหมู่ประเภทของคำที่หนึ่ง

8) หมวดหมู่ประเภทของคำที่สอง

9) ตำแหน่งของคำที่หนึ่งในประโยค

10) ตำแหน่งของคำที่สองในประโยค

เนื่องจากการจำแนกประเภทของคำ (Part of speech tagging) ในภาษาไทย นอกจากจำแนกคำประเภทต่างๆแล้ว ประเภทของคำยังถูกจำแนกอย่างละเอียด เช่น คำนาม สามารถจำแนก

ประเภทของคำออกได้เป็น NPRP, NCNM, NONM, NLBL, NCMN, NTTL ซึ่งความหมายของแต่ละประเภทของคำนามแสดงดังตารางที่ 2.2

ตารางที่ 2.2 ประเภทของคำนาม

ประเภทของคำ	คำอธิบาย	ตัวอย่าง
NPRP	คำเฉพาะ	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
NCNM	ตัวเลข	หนึ่ง, สอง, สาม, 1, 2, 3
NONM	ลำดับเลข	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
NLBL	หัวข้อ	1, 2, 3, 4, ก, ข, a, b
NCMN	คำนามทั่วไป	หนังสือ, อาหาร, อาคาร, คน
NTTL	ตำแหน่ง	ดร., พลเอก

จากประเภทของคำนามแต่ละแบบในตารางที่ 2.2 หากคำใดๆมีประเภทของคำตามตารางที่ 1 จะถูกจัดให้เป็นคำที่อยู่ในหมวดหมู่เดียวกัน ในขณะที่ คำสำคัญ คือ คำที่ถูกจำแนกประเภทเป็น คำนาม คำกริยา และคำคุณศัพท์ นอกเหนือจากนั้นจะถูกจัดให้เป็นคำไวยากรณ์ จากประโยคในรูปที่ 2.5 สามารถจำแนกประเภทของคำออกมาได้เป็น (เขา, PPRS), (ชอบ, VSTA), (รถยนต์, NCMN), (ฮอนด้า, NPRP), (มาก, ADVN) โดยที่ PPRS, VSTA และ ADVN คือ คำสรรพนาม คำกริยา และคำคุณศัพท์ ซึ่งคุณลักษณะของคำว่า “รถยนต์” ที่ใช้ในการหาปมรากของประโยค และความสัมพันธ์ของคำว่า “ชอบ” และ “รถยนต์”แสดงดังตารางที่ 2.3 และ 2.4

ตารางที่ 2.3 คุณลักษณะของคำว่า “รถยนต์” สำหรับการหาปมราก

คุณลักษณะ	1	2	3	4	5	6	7
ผลลัพธ์	NCMN	3	1	0	0	0	1

ตารางที่ 2.4 คุณลักษณะของความสัมพันธ์ของคำว่า “ชอบ” และ “รถยนต์”

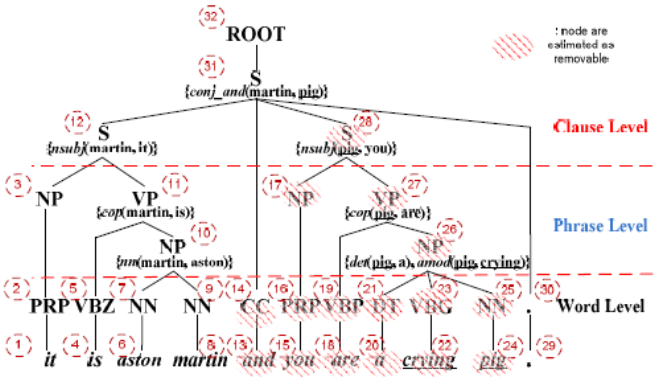
คุณลักษณะ	1	2	3	4	5	6	7	8	9	10
ผลลัพธ์	VSTA	NCMN	0	1	Content	Content	Verb	Noun	2	3

2.7.3 การสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกัน สามารถทำได้โดยการหาค่าความน่าจะเป็นที่มากที่สุดที่เป็นไปได้ของต้นไม้ ด้วยข้อมูลความน่าจะเป็นจากแมทริกซ์ที่ได้ก่อนหน้านี้ ซึ่งการสร้างเริ่มจากปมรากที่หาได้จากขั้นตอนแรก

2.8 งานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยเกี่ยวกับการจำแนกข้อความ จะเน้นไปที่เรื่องคุณลักษณะของข้อมูลที่ใช้ในการจำแนกข้อความ การประมวลผลข้อความ และประสิทธิภาพของการจำแนกข้อความ ซึ่งมีงานวิจัยดังต่อไปนี้

Xu และ Zhu (2010) [14] นำเสนอการลบคำรุนแรงออกจากข้อความเมื่อตรวจพบ โดยมีจุดประสงค์เพื่อลบคำรุนแรงออกจากข้อความ โดยที่ข้อความยังคงมีใจความสำคัญอยู่ ซึ่งการลบคำรุนแรงออกจากข้อความ หลังจากตรวจพบ ขั้นตอนวิธีจะวิเคราะห์คำในข้อความเกี่ยวกับความสัมพันธ์ทางไวยากรณ์ (Grammatical relation) ของแต่ละคำในข้อความ และทำการลบโครงสร้างทางไวยากรณ์ที่เกี่ยวข้องกับคำรุนแรงออกไป เพื่อให้ประโยคที่เหลืออยู่ยังอ่านได้ใจความอยู่ดังรูปที่ 2.6 ผู้วิจัยมีแนวคิดในการนำหลักการวิเคราะห์ความสัมพันธ์ทางไวยากรณ์ระหว่างคำในข้อความดังเช่นในงานวิจัยดังกล่าวมาใช้ในการวิเคราะห์ข้อความที่อาจเข้าข่ายหมิ่นประมาท



รูปที่ 2.6 โครงสร้างของประโยคในการลบคำรุนแรง [14]

Chen และ คณะ (2012) [15] นำเสนอวิธีการคัดแยกข้อความที่มีเนื้อหาก้าวร้าว จากข้อความปกติ ซึ่งมีชื่อว่า LSF โดยสร้างคลังของคำที่มีความหมายก้าวร้าว ซึ่งแต่ละคำในคลังคำศัพท์จะถูกให้คะแนนความรุนแรงว่ามาก หรือน้อย และงานวิจัยมีแนวคิดว่าคำรุนแรงที่มีความสัมพันธ์กับคำระบุตัวตนของผู้ใช้งาน (User identifiers) หรือชื่อคน จะมีความรุนแรงมากยิ่งขึ้น จึงหาความสัมพันธ์ระหว่างคำกับคำระบุตัวตน (Grammatical dependencies) เพื่อหาคะแนนความก้าวร้าวของประโยคในการจำแนกประเภทข้อความรุนแรงออกจากข้อความทั่วไป นอกจากนี้งานวิจัยยังเปรียบเทียบกับวิธีต่างๆได้แก่ Bag of words (BoW), 2-grams, 3-grams, 5-grams และ Appraisal approach ซึ่ง Appraisal approach คือคลังของคำศัพท์ที่เก็บข้อมูลรายละเอียดของแต่ละคำได้แก่ Attitude, Orientation, Polarity, และ Graduation เพื่อใช้ในการหาคุณลักษณะของ

ประโยคและจำแนกประเภทของประโยคต่อไป ผลปรากฏว่า ทุกวิธีที่เปรียบเทียบทั้งหมดมีความแม่นยำมากกว่า 80 เปอร์เซ็นต์ แต่ทุกวิธีนอกจาก LSF มี recall ต่ำกว่า 70 เปอร์เซ็นต์เนื่องจากข้อความรุนแรงมีกฎเกณฑ์ ที่เกี่ยวข้องกับคำระบุตัวตน จากวิธีการทั้งหมด BoW ใช้คำระบุตัวตนกับคำหยาบคายที่ไม่เกี่ยวข้องกันมาเป็นคุณลักษณะในการจำแนก ทำให้มีค่า False positive สูงที่สุด ในขณะที่ n-grams มี recall ที่ต่ำเมื่อค่า n น้อย อย่างไรก็ตามเมื่อค่า n มากขึ้นส่งผลให้ค่า false positive เพิ่มขึ้นตามไปด้วย

Van Hee และ คณะ (2015) [16] เสนอวิธีการจำแนกประเภทของการข่มเหงรังแก (Cyberbullying) บนสื่อสังคมออนไลน์ผ่านข้อความ ซึ่งประเภทของการข่มแหงรังแกจำแนกออกเป็น 8 ประเภทได้แก่ การข่มขู่ (Threat/Blackmail) ข้อความดูถูก (Insult) ข้อความสาปแช่ง (Curse/Exclusion) ข้อความที่ทำให้เสียชื่อเสียง (Defamation) ข้อความเรื่องเพศ (Sexual Talk) ข้อความปกป้องผู้ถูกข่มเหงรังแก (Defense) ข้อความสนับสนุนการคุกคาม (Encouragement to Harasser) และอื่นๆ (Other) ซึ่งวิธีที่ใช้ในการสร้างคุณลักษณะในการจำแนกข้อความได้แก่ n-grams ระดับคำ, n-grams ระดับตัวอักษร และการวิเคราะห์ความรู้สึก ซึ่งจำแนกโดยใช้ SVM แต่ประสิทธิภาพในการจำแนกข้อความที่ทำให้เสื่อมเสียชื่อเสียงค่อนข้างต่ำคือ F-score เท่ากับ 7.41% ต่อมา Van Hee และ คณะ (2018) [17]ได้เพิ่มวิธีในการสร้างคุณลักษณะในการจำแนกข้อความคือ การวิเคราะห์หัวข้อของข้อความ (Topic Model) และการหาเงื่อนไขของประโยค ประกอบด้วย ชื่อเฉพาะ (Proper Name) คำที่ใช้แสดงความเหมารวม (‘allness’ Indicators) คำที่แสดงความลดลง (Diminishers) คำที่ใช้เพิ่มความหนักแน่น (Intensifiers) คำที่มีความหมายเชิงลบ (Negation Words) หรือคำที่รุนแรง (Aggressive Language) หรือคำต้องห้าม (Profanity Words) และ การเปลี่ยนแปลงตามบุคคล (Person Alternation) ซึ่งหมายถึงคุณลักษณะของการปรากฏของคำสรรพนามบุคคลที่หนึ่งและบุคคลที่สอง จากการเพิ่มวิธีการสร้างคุณลักษณะของข้อความทำให้ประสิทธิภาพในการจำแนกประเภทของข้อความทั้ง 8 ประเภทโดยรวมเฉลี่ยแล้วดีขึ้น คือ F-score เท่ากับ 64.32% แต่ผลการทดลองไม่ได้ระบุเฉพาะเจาะจงลงไปว่าการจำแนกข้อความที่ทำให้เสื่อมเสียชื่อเสียงมีประสิทธิภาพดีขึ้นหรือไม่อย่างไร

บทที่ 3
การจำแนกข้อความเข้าข่ายหมิ่นประมาท

3.1 แนวคิดในการจำแนกข้อความเข้าข่ายหมิ่นประมาท

การจำแนกข้อความส่วนใหญ่มักจะใช้การเรียนรู้ของเครื่องในการจำแนกข้อความ ซึ่งการหาคุณลักษณะที่สามารถแสดงถึงเอกลักษณ์ของข้อความเข้าข่ายหมิ่นประมาทออกมาได้นั้น ผู้วิจัยจึงมีแนวคิดที่จะใช้หลักประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาท ซึ่งประกอบไปด้วยองค์ประกอบที่สำคัญ คือ ผู้กระทำ ผู้ถูกใส่ความ บุคคลที่สาม การใส่ความ และการยืนยันข้อเท็จจริง ร่วมกับคุณลักษณะของข้อความที่มีการใช้งานกันอย่างแพร่หลาย คือ n-grams ซึ่งประกอบไปด้วย n-grams ระดับคำ และ n-grams ระดับตัวอักษร

ในงานวิจัยนี้ผู้วิจัยทำการเก็บข้อมูลจากเฟซบุ๊กเพจและกลุ่มในเฟซบุ๊กเกี่ยวกับข่าวสารในการเมืองปัจจุบัน จำนวนทั้งหมด 18 เพจ 2 กลุ่ม ซึ่งประกอบด้วยโพสต์ทั้งหมด 95 โพสต์ โดยข้อมูลที่เก็บจะเป็นความคิดเห็นของผู้ใช้งานเฟซบุ๊กที่มีต่อโพสต์เหล่านั้น

จากการเก็บข้อมูลในรอบที่หนึ่งภายในเดือนตุลาคม 2560 นั้น ผู้วิจัยสามารถเก็บข้อมูลได้ทั้งหมด 1035 ข้อความ ประกอบด้วยข้อความเข้าข่ายหมิ่นประมาทจำนวน 189 ข้อความ และข้อความไม่เข้าข่ายหมิ่นประมาทจำนวน 846 ข้อความ แต่เนื่องจากข้อมูลไม่สมดุลจึงทำการเก็บข้อมูลรอบที่สองในเดือนตุลาคม 2561 ได้ข้อมูลมาทั้งหมดจำนวน 305 ข้อความ ประกอบด้วยข้อความเข้าข่ายหมิ่นประมาทจำนวน 104 ข้อความ และ ข้อความไม่เข้าข่ายหมิ่นประมาทจำนวน 201 ข้อความ รวมจำนวนข้อความทั้งหมด 1,340 ข้อความ ประกอบด้วยข้อความเข้าข่ายหมิ่นประมาท 293 ข้อความ และข้อความไม่เข้าข่ายหมิ่นประมาท 1,047 ข้อความ จากข้อความที่เก็บมาทั้งหมดนั้น จะเห็นว่าข้อความเข้าข่ายหมิ่นประมาทมีจำนวนน้อยมากเมื่อเทียบกับข้อความไม่เข้าข่ายหมิ่นประมาท ซึ่งอาจจะส่งผลกระทบถึงการจำแนกข้อความ

จากสมมติฐานดังกล่าวจึงทำให้ผู้วิจัยออกแบบการทดลองเป็นสองขั้นตอน ซึ่งแต่ละขั้นตอนใช้คุณลักษณะของข้อความแบบเดียวกัน โดยการทดลองที่หนึ่ง ทดลองจำแนกข้อความกับชุดข้อความที่ไม่มีการจัดการกับปัญหาความไม่สมดุล และการทดลองที่สอง ทดลองกับชุดข้อความที่มีการจัดการกับปัญหาความไม่สมดุลแล้ว ผู้วิจัยเลือกใช้ขั้นตอนวิธี SMOT เพื่อจัดการกับข้อมูลที่มีความไม่สมดุลกัน เนื่องจาก SMOT เป็นขั้นตอนวิธีที่ใช้วิธีการสร้างข้อมูลที่มีน้อยขึ้นมาใหม่ (Oversampling) ไม่ใช่การลดจำนวนข้อมูลที่มีจำนวนมากลง (Undersampling)

3.2 คุณลักษณะที่ใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาท

จากแนวคิดในการจำแนกข้อความเข้าข่ายหมิ่นประมาทที่ได้กล่าวในหัวข้อที่ 3.1 ผู้วิจัยได้นำแนวคิดมาออกแบบคุณลักษณะเพื่อใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทดังนี้

3.2.1 n-grams

ข้อความประกอบไปด้วยคำเรียงต่อกัน ซึ่งข้อความที่เข้าข่ายหมิ่นประมาทมักมีคำเฉพาะที่ใช้ในการหมิ่นประมาท จึงทำให้ผู้วิจัยเลือกใช้ n-grams ในการตัดคำเพื่อนำมาเป็นคุณลักษณะในการจำแนกข้อความ โดยที่ Grams คือ หน่วยที่ใช้ในการแบ่งคำของข้อความ เพราะฉะนั้น n-grams ระดับคำ คือ การแบ่งคำในข้อความออกเป็นทั้งหมด n คำ ในขณะที่ n-grams ระดับตัวอักษร คือ การแบ่งตัวอักษรออกเป็น n ตัวโดยไม่เกินขอบเขตของคำ ซึ่งการเพิ่มคุณลักษณะ n-grams ระดับตัวอักษรอาจจะช่วยเพิ่มประสิทธิภาพของการจำแนกประเภทข้อความในกรณีที่เกิดความผิดพลาดในขั้นตอนการตัดคำภาษาไทย เช่น หากการตัดคำของประโยค “มันโกงเข้ามาเป็นผบ.” เกิดความผิดพลาดในการตัดคำเป็น “มันโกง|เข้า|มา|เป็น|ผบ.” n-grams ระดับตัวอักษรยังสามารถได้คุณลักษณะเป็นคำว่า “โกง” จากคำว่า “มันโกง” ในขณะที่ n-grams ระดับคำไม่ได้คุณลักษณะเป็นคำว่า “โกง” ตัวอย่างของ n-grams ระดับตัวอักษร และ n-grams ระดับคำแสดงดังตารางที่ 3.1

ตารางที่ 3.1 แสดงตัวอย่างของ n-grams ระดับคำศัพท์ และ ตัวอักษร

ประโยค	มัน โกง เข้า มา เป็น ผบ.
คำศัพท์ 2-grams	“มัน โกง” “โกง เข้า” “เข้า มา” “มา เป็น” “เป็น ผบ.”
ตัวอักษร 3-grams	“มัน” “โกง” “เข้” “ข้า” “เป็” “ป็น” “ผบ.”

n-grams จะสร้างคลังของคำศัพท์ที่แบ่งคำตามจำนวน n สำหรับแต่ละข้อความเมื่อตัดแบ่งออกมาด้วย n-grams แล้ว จะเทียบคำหรือตัวอักษรที่ตัดออกมาว่าปรากฏในคลังของ n-grams หรือไม่ หากปรากฏให้แทนด้วย 1 หากไม่ปรากฏ ให้แทนด้วย 0 เพราะฉะนั้น คุณลักษณะของประโยคที่ใช้ n-grams จะมีความยาวทั้งหมดเท่ากับ N เมื่อ N คือจำนวนคำหรือตัวอักษรที่ตัดแบ่งออกมา ในงานวิจัยนี้ใช้ 1-grams 2-grams และ 3-grams สำหรับคำศัพท์ และ 2-grams 3-grams และ 4-grams สำหรับตัวอักษร ร่วมกันในการสร้างแบบจำลองในการจำแนกข้อความเข้าข่ายหมิ่นประมาท โดยมีจำนวนคุณลักษณะ n-grams ทั้งหมด 47,517 คุณลักษณะ แบ่งเป็น 35,015

คุณลักษณะสำหรับ n-grams ระดับคำ และ 12,502 คุณลักษณะสำหรับ n-grams ระดับตัวอักษร ตัวอย่างของคุณลักษณะ n-grams ของประโยค “มันโกงเข้ามาเป็นผบ.” ที่ใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทแสดงดังตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างของคุณลักษณะ n-grams

n-grams ระดับคำ	1-grams	มัน	โกง	เข้า	มา	เป็น	ผบ.	...							
		1	1	1	1	1	1	0							
	2-grams	มัน โกง	โกง เข้า	เข้า มา	มา เป็น	เป็น ผบ.	...								
		1	1	1	1	1	0								
	3-grams	มัน โกง เข้า	โกง เข้า มา	เข้า มา เป็น	มา เป็น ผบ.	...									
		1	1	1	1	0									
n-grams ระดับตัวอักษร	2-grams	ม	ัน	โก	กง	เข	ข้	้า	มา	เป	ป็	น	ผบ	บ.	...
		1	1	1	1	1	1	1	1	1	1	1	1	1	0
	3-grams	มัน	โกง	เข้	ข้า	เป็	ป็น		ผบ.	...					
		1	1	1	1	1	1		1	0					
	4-grams	เข้า	เป็น	...											
		1	1	0											

3.2.2 คลังคำศัพท์จากคำพิพากษาศาลฎีกา (Dictionary of Judgment Terms)

คลังคำศัพท์จากคำพิพากษาศาลฎีกา คือ คลังของคำศัพท์ที่ผู้วิจัยเก็บรวบรวมจากคำพิพากษาศาลฎีกา เพื่อนำมาสร้างคุณลักษณะเพื่อใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาท โดยมีองค์ประกอบทั้งหมด 5 อย่าง คือ ผู้กระทำ ผู้ถูกใส่ความ บุคคลที่สาม การใส่ความ และการยืนยันข้อเท็จจริง ตามที่ประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาทได้กล่าวไว้ ซึ่งผู้วิจัยได้ออกแบบคุณลักษณะที่ใช้ในการจำแนกข้อความตามองค์ประกอบดังนี้

1) ผู้กระทำ คือ ผู้ใส่ความ หมายถึงผู้ที่แสดงความคิดเห็นบนสื่อสังคมออนไลน์ และสามารถไปปรากฏอยู่ในข้อความได้เช่นเดียวกันดังตัวอย่างต่อไปนี้ เช่น “**ผม**คิดว่าถ้าใครอยากศึกษาเรื่องความบ้าอำนาจ ความเกลียดชังต่อคนเห็นต่าง จนนำไปสู่การหลอกตัวเอง หลงตัวเองในระดับจิตใต้สำนึก สุเทพเนี่ยเหมาะสำหรับเป็นกรณีศึกษามากที่สุด” จากข้อความข้างต้นจะเห็นว่าผู้กระทำสามารถใช้สรรพนามบุคคลที่ 1 คำว่า “ผม” ในการแสดงความคิดเห็นแทนตัวเองเพื่อหมิ่นประมาทผู้อื่นได้เช่นกัน

2) ผู้ถูกใส่ความ คือ เหยื่อที่ถูกผู้กระทำหมิ่นประมาทผ่านการแสดงความคิดเห็นบนสื่อสังคมออนไลน์ ซึ่งผู้กระทำสามารถระบุถึงผู้ถูกใส่ความได้ด้วยชื่อ หรือ การใช้สรรพนามบุคคลที่ 2 และ 3 เช่นตัวอย่างดังต่อไปนี้ “ผมคิดว่าถ้าใครอยากศึกษาเรื่องความบ้าอำนาจ ความเกลียดชังต่อคนเห็นต่าง จนนำไปสู่การหลอกตัวเอง หลงตัวเองในระดับจิตใต้สำนึก **สุเทพ**เนี่ยเหมาะสำหรับเป็น

กรณีศึกษามากที่สุด” และ “คนอย่าง**มึง**นี่ไม่น่าเกิดมาเลยจริงๆ ไอ้สวะ อัปปรีย์ เสนียดจัญไรบ้านเมือง” จากตัวอย่างจะเห็นได้ว่าผู้ใส่ความอ้างอิงถึงผู้ถูกใส่ความโดยใช้ชื่อ “สุเทพ” และ สรรพนามบุคคลที่ 2 “มึง” ในการกล่าวถึงผู้ถูกใส่ความ

3) บุคคลที่สาม คือ ผู้ที่อยู่ในเหตุการณ์ในการแสดงความคิดเห็นเพื่อหมิ่นประมาทบนสื่อสังคมออนไลน์ หรือก็คือ ผู้ใช้งานสื่อสังคมออนไลน์ เพราะฉะนั้น บุคคลที่สามผู้วิจัยจึงไม่ได้นำมาเป็นคุณลักษณะ

4) การใส่ความ คือ การพูดหาเหตุร้ายทำให้ผู้ถูกใส่ความเสียหาย อาจกระทำด้วยวาจา ลายลักษณ์อักษร กริยาท่าทาง หรือพฤติกรรมอื่นๆ แต่การแสดงความคิดเห็นนั้นผู้วิจัยสนใจเฉพาะข้อความ จึงทำให้ผู้วิจัย ใช้คำศัพท์ หรือ วลี ที่ทำให้ผู้ถูกใส่ความเสียหาย โดยผู้วิจัยได้เก็บข้อมูลของคำศัพท์จากคำพิพากษาศาลฎีกา ซึ่งแบ่งคำออกเป็น 3 ประเภท คือ คำกริยา คำนาม และคำหยาบ

4.1) คำกริยา คือ การกระทำที่ผู้ใส่ความใส่ความผู้เสียหาย เช่น ฎีกาที่ 1006/2542 “นางประทิน กับพวก **สมคบ**กัน**ขู่**ให้พยานให้การ**ปรักปรำ**นายดิเรก โดยโจทก์ได้**เรียกร้อง *เงิน*** จำนวน 30000 บาท จากจำเลยที่ 1 เพื่อเป็นการตอบแทนในการปั้นพยานอันเป็นการสร้างพยานหลักฐานที่ไม่เป็นความจริงให้แก่จำเลยที่ 1” จากเหตุการณ์ข้างต้น ผู้วิจัยได้เก็บคำกริยาได้แก่ สมคบ ข่มขู่ ปรักปรำ และ เรียกร้อง เพื่อเก็บเป็นคำที่สามารถใส่ความให้ผู้อื่นเสียหายได้ นอกจากนี้ผู้วิจัยได้เก็บคำนามซึ่งเป็นคำที่เกี่ยวข้องกับคำกริยา จากตัวอย่างข้างต้น ผู้วิจัยเก็บคำนามคำว่าเงิน เนื่องจากเป็นคำที่เกี่ยวข้องกับคำว่าเรียกร้อง ซึ่งคำว่าเรียกร้องคำเดียวไม่สามารถส่อไปถึงการใส่ความได้ แต่หากรวมกับคำว่าเงินแล้วอาจทำให้หมายถึงการใส่ความผู้อื่นให้เสียหายได้

นอกจากนี้จากการเก็บข้อมูลคำศัพท์ ผู้วิจัยแบ่งคำกริยาออกเป็น 2 ประเภท คือ คำกริยาที่กรรมไม่มีผล และคำกริยาที่กรรมมีผล คำกริยาที่กรรมไม่มีผล คือ คำกริยาที่มีความหมายรุนแรง หากกล่าวออกมา ถือว่าผู้กล่าวหมิ่นประมาทผู้อื่นแน่นอน เช่น จากฎีกาที่ 5797/2545 “นางเสาวลักษณ์ได้**ขโมย**เศษทองแดงสายไฟฟ้าชำรุดของห้างหุ้นส่วนจำกัดสายไฟไทยอุตสาหกรรมไปขายให้แก่นายประสาน” คำว่าขโมยจากตัวอย่างข้างต้น หากกล่าวออกมาว่าขโมย ไม่ว่าจะขโมยอะไรก็ตาม ก็ถือเป็นการหมิ่นประมาท ในขณะที่บางข้อความมีเนื้อหาหมิ่นประมาทผู้อื่นแต่คำกริยานั้นต้องการกรรมจึงจะมีความหมายหมิ่นประมาท เรียกว่า คำกริยาที่กรรมมีผล ดังเช่น ฎีกาที่ 97/2543 “มึง**เป็น**เมียน้อยสารวัตร ส. อย่ามาทำใหญ่ให้กูเห็นนะ” คำว่าเป็น หากใช้คำว่าเป็นเดี่ยวๆ นั้นไม่สามารถระบุได้ว่าเป็นการหมิ่นประมาทได้ เช่น คำว่าเป็นในประโยค “ฉัน**เป็น**คนดี” แต่หากกล่าวหาว่า “เป็น**เมียน้อย**” คำว่าเมียน้อยนี้ทำให้คำข้อความดังกล่าวถือเป็นการหมิ่นประมาท

4.2) คำหยาบ คือ คำที่ทำให้ผู้ที่ถูกว่ารู้สึกอับอายแต่ไม่เสื่อมเสียชื่อเสียง ไม่เป็นการหมิ่นประมาท ซึ่งในคำพิพากษาศาลฎีกาสามารถพบเห็นได้บ่อยครั้ง เช่น ฎีกาที่ 2324/2518 “ไอ้

ทนาย**กระจอก** ไอ้ทนาย**เฮงซวย**” ศาลได้ตัดสินว่าการใช้ถ้อยคำดังกล่าวเป็นการพูดดูหมิ่นเหยียดหยามโจทก์ให้ได้รับความอับอายและเจ็บใจเท่านั้น หาใช่เป็นการใส่ความโจทก์โดยประการที่น่าจะทำให้โจทก์เสียชื่อเสียง ถูกดูหมิ่น หรือถูกเกลียดชังไม่ ไม่เป็นความผิดฐานหมิ่นประมาทตามประมวลกฎหมายอาญา มาตรา 326 เช่นเดียวกันกับคำด่า เช่น “ไอ้เหี้ย” “ไอ้สัตว์” เป็นต้น

4.3) วลี คือ กลุ่มคำ หรือข้อความที่ปรากฏอยู่ในคำพิพากษาศาลฎีกา ที่ทำให้ผู้ถูกกล่าวหาเสื่อมเสียชื่อเสียง เช่น ฎีกาที่ 302/2507 “ทนายความเมืองร้อยเอ็ดคบไม่ได้ เป็น**นกสองหัว เหยียบเรือสองแคม เป็นมวยล้ม** ว่าความทีแรกดีได้รับเงินแล้วก็ว่าเป็นอย่างอื่น” วลี นกสองหัว เหยียบเรือสองแคม เป็นมวยล้ม นั้นทำให้ผู้ถูกกล่าวหาเสื่อมเสียชื่อเสียง แต่จากฎีกาดังกล่าวผู้กล่าวหาไม่ได้ระบุว่าเป็นทนายร้อยเอ็ดคนใดจึงทำให้ไม่มีความผิดฐานหมิ่นประมาท

5) การยืนยันข้อเท็จจริง การยืนยันข้อเท็จจริง คือ การใส่ความผู้อื่นเช่น ฎีกาที่ 97/2543 “มึง**เป็น**เมียน้อยสารวัตร ส. อย่ามาทำใหญ่ให้กูเห็นนะ” ได้กล่าวว่าผู้เสียหายนั้นเป็นเมียน้อยของสารวัตร ในขณะที่ ฎีกาที่ 2155/2531 “น้อย (นายประกอบ) มีอะไรกับติ๋ม (โจทก์) จริงหรือเปล่า” ศาลฎีกาวินิจฉัยว่าการกระทำของจำเลยตามที่ได้บรรยายมาในฟ้องเป็นเรื่องที่จำเลยถามนายประกอบว่ามีความสัมพันธ์ทางชู้สาวกับโจทก์จริงหรือไม่ถ้าจริงก็ให้เลิกเสียเท่านั้นไม่ได้ยืนยันถึงว่านายประกอบมีความสัมพันธ์ทางชู้สาวกับโจทก์ยังไม่เข้าลักษณะเป็นการใส่ความอันจะเป็นหมิ่นประมาทโจทก์ตามประมวลกฎหมายอาญามาตรา 326 จากตัวอย่างดังกล่าวทำให้ผู้วิจัยมีแนวคิดที่ว่าการยืนยันข้อเท็จจริงมีความเกี่ยวข้องกับชนิดของประโยคภาษาไทย

จากองค์ประกอบตามประมวลกฎหมายอาญาความผิดฐานหมิ่นประมาทที่ได้กล่าวมาข้างต้น ผู้วิจัยได้รวบรวมคำศัพท์จากคำพิพากษาศาลฎีกาตามองค์ประกอบข้างต้น และเพิ่มจำนวนคำศัพท์โดยการเพิ่มคำเหมือน (Synonyms) เข้าไปเพื่อให้คำศัพท์มีหลากหลายขึ้น ผู้วิจัยได้สรุปและออกแบบเป็นคุณลักษณะเพื่อใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาท 8 ชนิดดังนี้

1) คำสรรพนามบุรุษที่หนึ่ง

2) คำสรรพนามบุรุษที่สอง และคำสรรพนามบุรุษที่สาม

3) ชื่อเฉพาะ คือ ชื่อของบุคคล หน่วยงาน หรือสถานที่

4) คำกริยา คือ คำกริยาที่เก็บรวบรวมจากคำพิพากษาศาลฎีกา เช่น โกง ขโมย เป็น

5) คำนาม คือ คำนามที่เก็บรวบรวมจากคำพิพากษาศาลฎีกา เช่น เงิน ชู้ เมียน้อย

6) วลี คือ กลุ่มคำที่เก็บรวบรวมจากคำพิพากษาศาลฎีกา เช่น นกสองหัว

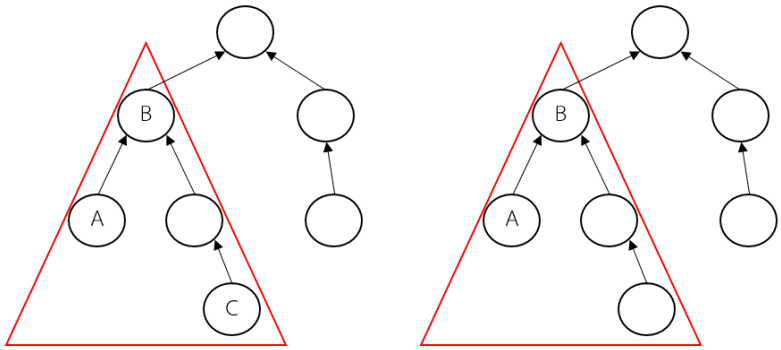
7) ชนิดของประโยคภาษาไทย ได้แก่ ประโยคบอกเล่า ประโยคคำถาม ประโยคปฏิเสธ ประโยคขอร้องหรือชักชวน ประโยคคำสั่ง และประโยคอุทาน

8) คำหยาบ คือ คำหยาบที่เก็บรวบรวมจากคลังคำศัพท์ศาลฎีกาและแหล่งอื่นๆ เช่น ไอ้สัตว์ ไอ้เหี้ย กระจอก

จากคุณลักษณะทั้ง 8 ข้อ สามารถจำแนกตามองค์ประกอบของประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาทได้ว่า ข้อที่ 1 คือ ผู้กระทำ ข้อที่ 2 และ 3 คือ ผู้ถูกกระทำ ข้อที่ 4, 5 และ 6 คือ การใส่ความ ข้อที่ 7 คือ การยืนยันข้อเท็จจริง และข้อที่ 8 คือ คำที่ไม่ทำให้เสื่อมเสียชื่อเสียง โดยคุณลักษณะทั้ง 8 ที่กล่าวมาข้างต้น คือ คุณลักษณะแบบทวินาม กล่าวคือ เมื่อพบรูปแบบดังกล่าว จะให้คุณลักษณะแทนด้วย 1 หากไม่พบให้แทนด้วย 0 หากพิจารณาตัวอย่างประโยค “มันโกงเข้ามาเป็นผบ.” จะพบรูปแบบที่ 2 และ 4 ซึ่งเป็นคำสรรพนามบุคคลที่สาม “มัน” และคำกริยา “โกง” ตามลำดับ จึงได้คุณลักษณะที่ใช้จำแนกเป็น [0,1,0,1,0,0,0,0]

3.2.3 โครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกัน (Dependency Structure)

จากการสำรวจคำพิพากษาศาลฎีกาจะปรากฏอยู่ในรูปของ ประธาน กริยา หรือ ประธาน กริยา กรรม ผู้วิจัยจึงใช้โครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันเพื่อตรวจสอบคำศัพท์ภายในประโยคว่ามีความเกี่ยวข้องกันหรือไม่ดังรูปที่ 3.1



รูปที่ 3.1 ตัวอย่างรูปโครงสร้างต้นไม้ย่อยของประโยคที่พิจารณา

จากรูปที่ 3.1 ด้านซ้ายคือรูปแบบโครงสร้างต้นไม้ย่อยของประโยคที่ประกอบไปด้วย ประธาน (A) กริยา (B) และกรรม (C) ในขณะที่รูปด้านขวาคือรูปแบบโครงสร้างต้นไม้ย่อยที่พิจารณาเพียง ประธาน (A) และ กริยา (B) โดยที่ กริยา คือ คำกริยาที่ถูกเก็บรวบรวมไว้ในคลังคำศัพท์จากหัวข้อที่ 3.2.2 ประธาน คือ คำสรรพนามบุรุษที่สอง สรรพนามบุรุษที่สาม หรือชื่อเฉพาะ กรรม คือ คำนามที่ถูกเก็บรวบรวมไว้ในคลังคำศัพท์จากหัวข้อที่ 3.2.2 ชื่อเฉพาะ หรือคำสรรพนามบุรุษที่หนึ่ง จากโครงสร้างต้นไม้ย่อยดังกล่าว ผู้วิจัยสร้าง 8 คุณลักษณะดังนี้

1) โครงสร้างต้นไม้ย่อย NV_1 คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของชื่อเฉพาะ (ประธาน) และ กริยาที่กรรมไม่มีผล เช่น ติ๋มขโมย

2) โครงสร้างต้นไม้ย่อย NV_1O คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของ ชื่อเฉพาะ (ประธาน) และ กริยาที่กรรมไม่มีผล แต่มีกรรม เช่น ติ๋มขโมยเงิน

3) โครงสร้างต้นไม้ย่อย PV_1 คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของคำสรรพนามบุรุษที่สอง หรือสรรพนามบุรุษที่สาม (ประธาน) และ กริยาที่กรรมไม่มีผล เช่น เธอขโมย

4) โครงสร้างต้นไม้ย่อย PV_1O คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของคำสรรพนามบุรุษที่สอง หรือสรรพนามบุรุษที่สาม (ประธาน) และ กริยาที่กรรมไม่มีผล แต่มีกรรม เช่น เธอขโมยเงิน

5) โครงสร้างต้นไม้ย่อย NV_2O คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของ ชื่อเฉพาะ (ประธาน) และ กริยาที่กรรมมีผล และมีกรรม เช่น ติ๋มเป็นเมียน้อย

6) โครงสร้างต้นไม้ย่อย PV_2O คือ โครงสร้างต้นไม้ย่อยที่ประกอบด้วยความสัมพันธ์ของคำสรรพนามบุรุษที่สอง หรือสรรพนามบุรุษที่สาม (ประธาน) และ กริยาที่กรรมมีผล และมีกรรม เช่น เธอเป็นเมียน้อย

7) โครงสร้างต้นไม้ย่อย V_1O คือ โครงสร้างต้นไม้ที่ไม่มีประธาน แต่มีกริยาที่กรรมไม่มีผล แต่มีกรรม เช่น โกงเงิน

8) โครงสร้างต้อนไม้ย่อย V_2O คือ โครงสร้างต้นไม้ย่อยที่ไม่มีประธาน แต่มีกริยาที่กรรมมีผล และมีกรรม เช่น เป็นชู้

คุณลักษณะที่ได้ในหัวข้อนี้ มีความแตกต่างจากคุณลักษณะในข้อที่ 3.2.2 โดยที่คุณลักษณะในข้อที่ 3.2.2 ดูเพียงการปรากฏของคำ ในขณะที่คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันนั้นสนใจความสัมพันธ์ของคำในประโยค และเป็นตัวบ่งชี้ว่าคำศัพท์ที่พบตามที่ระบุไว้ในคุณลักษณะข้อที่ 3.2.2 นั้นมีความสัมพันธ์กันจริงๆ สำหรับข้อ 7 และข้อ 8 ผู้วิจัยได้ปรับปรุงการพิจารณาโครงสร้างเพื่อให้สอดคล้องกับลักษณะของการแสดงความคิดเห็นของผู้ใช้สื่อสังคมออนไลน์ เมื่อผู้แสดงความคิดเห็นไม่ได้กล่าวถึงผู้เสียหาย แต่บุคคลที่สาม หรือผู้ที่เห็น หรืออ่านความคิดเห็นนั้นสามารถเชื่อมโยงได้ว่าใครคือผู้เสียหายโดยอ้างอิงจากข้อมูลในโพสต์ของเฟซบุ๊ก

ตัวอย่างในการสร้างโครงสร้างต้นไม้ไวยากรณ์จากประโยคในตารางที่ 3.1 สามารถหาความน่าจะเป็นจากขั้นตอนวิธีในหัวข้อที่ 2.7.1 ความน่าจะเป็นที่คำจะเป็นปมราก และหัวข้อที่ 2.7.2 ความน่าจะเป็นของความสัมพันธ์ของคำแสดงดังตารางที่ 3.3 และ 3.4

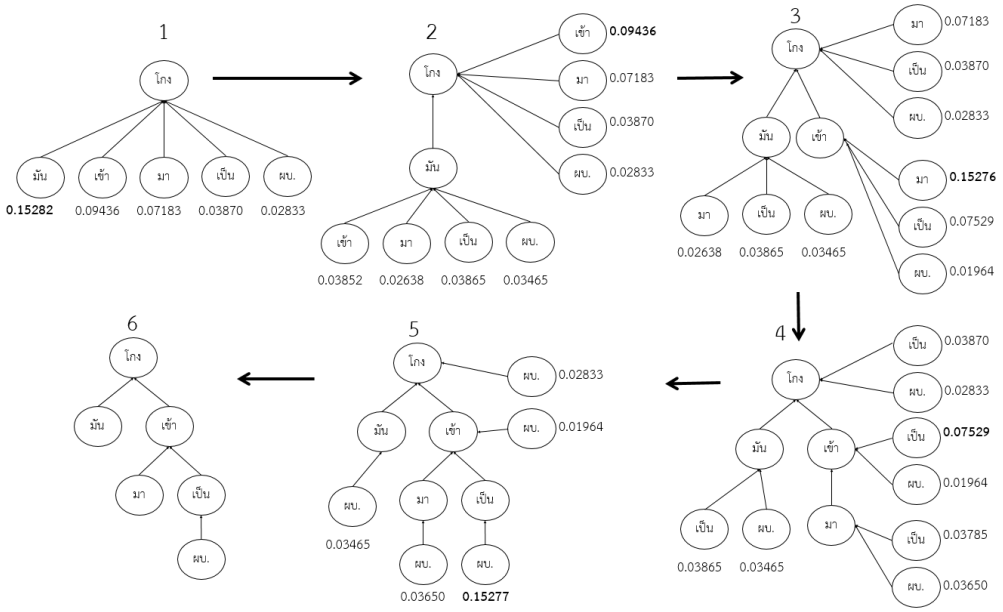
ตารางที่ 3.3 ความน่าจะเป็นที่คำจะเป็นปมราก

คำ	ความน่าจะเป็น
มัน	0.0169
โกง	0.6871
เข้า	0.0175
มา	0.0133
เป็น	0.0104
ผบ.	0.0306

ตารางที่ 3.4 ความน่าจะเป็นของความสัมพันธ์ของคำ

	มัน	โกง	เข้า	มา	เป็น	ผบ.
มัน	0.00000	0.15282	0.15278	0.03872	0.03106	0.03818
โกง	0.03870	0.00000	0.03871	0.03794	0.03872	0.03865
เข้า	0.03852	0.09436	0.00000	0.03752	0.03258	0.03871
มา	0.02638	0.07183	0.15276	0.00000	0.07490	0.03194
เป็น	0.03865	0.03870	0.07529	0.03785	0.00000	0.03320
ผบ.	0.03465	0.02833	0.01964	0.03650	0.15277	0.00000

จากตารางที่ 3.3 คำที่มีความน่าจะเป็นมากที่สุดคือคำว่า “โกง” ด้วยความน่าจะเป็นเท่ากับ 0.6871 ซึ่งการสร้างโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันจะเริ่มต้นจากปมรากคำว่า “โกง” จากนั้นค้นหาโครงสร้างต้นไม้ที่มีความน่าจะเป็นมากที่สุดจากประโยค โดยที่แถวในตารางที่ 3.4 แสดงถึงความน่าจะเป็นที่คำในแถวจะมีความสัมพันธ์กับคำในคอลัมน์ เช่น คำว่า “มัน” มีความน่าจะเป็นที่จะมีความสัมพันธ์กับคำว่า “โกง” เท่ากับ 0.15282 การสร้างโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันทำได้ดังรูปที่ 3.2



รูปที่ 3.2 ตัวอย่างการสร้างโครงสร้างต้นไม้ของคำในประโยค

จากโครงสร้างต้นไม้รูปที่ 3.2 หมายเลข 6 ซึ่งเป็นโครงสร้างที่สมบูรณ์แล้ว พบโครงสร้างย่อย PV_1 คือ ความสัมพันธ์ของคำว่า “มัน” และ “โกง” ซึ่งคำว่า “โกง” คือ คำกริยาที่เก็บรวบรวมในขั้นตอนที่ 3.2.2 และคำว่าโกง คือ คำกริยาที่กรรมไม่มีผล เพราะฉะนั้นคุณลักษณะของโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันของข้อความนี้จึงเป็น [0,0,1,0,0,0,0,0] เนื่องจากรูปแบบที่ปรากฏมีเพียงโครงสร้างย่อย PV_1 ในตำแหน่งที่ 3 ของคุณลักษณะจึงแทนด้วย 1 ในขณะที่ไม่พบโครงสร้างย่อยอื่นจึงถูกแทนด้วย 0 ซึ่งในงานวิจัยนี้ผู้วิจัยได้ประยุกต์ใช้การสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันโดยอาศัยชุดข้อมูล Universal dependency [3] เป็นข้อมูลสอน

จากวิธีการสร้างคุณลักษณะที่กล่าวมาข้างต้นทั้งหมด 4 วิธี คุณลักษณะทั้งหมดที่ใช้ในการจำแนกมีทั้งหมด 47,533 คุณลักษณะ (คอลัมน์) ได้แก่ คุณลักษณะ n-grams (หัวข้อที่ 3.2.1) แบ่งเป็นระดับคำ 35,015 คอลัมน์ และ ระดับตัวอักษร 12,502 คอลัมน์ คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (หัวข้อที่ 3.2.2) 8 คอลัมน์ และ คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกัน (หัวข้อที่ 3.3.3) 8 คอลัมน์ดังตารางที่ 3.5

ตารางที่ 3.5 ตัวอย่างคุณลักษณะที่ใช้ในการจำแนกข้อความ

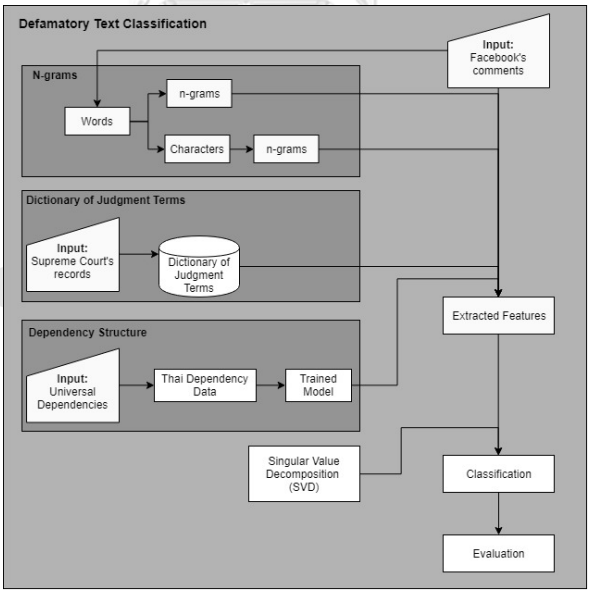
Text	n-grams													Dictionary of Judgment Terms								Dependency Structure								Class
	Words							Chars																						
	มัน	โกง	เข้า	มา	เป็น	ผบ.	...	ม	ั	น	โ	ก	ง	เ	ข	...	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	
มันโกงเข้ามาเป็นผบ.	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1

บทที่ 4
การทดลองและผลการทดลอง

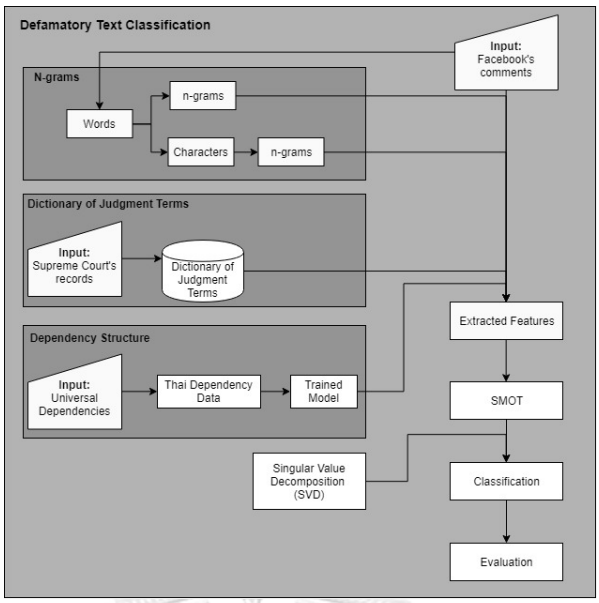
จากแนวคิดการทดลองในบทที่ 3 ผู้วิจัยได้ออกแบบการทดลองเป็น 2 แบบ คือ การทดลองกับชุดข้อความที่ยังไม่แก้ไขปัญหาความไม่สมดุลของจำนวนประเภทข้อความ และการทดลองกับชุดข้อความที่แก้ไขปัญหาความไม่สมดุลของจำนวนประเภทข้อความด้วยขั้นตอนวิธี SMOT โดยที่การทดลองทั้งสองแบบใช้การเรียนรู้ของเครื่อง และคุณลักษณะที่ใช้ในการจำแนกเหมือนกัน

จากรูปที่ 4.1 ข้อความความคิดเห็นที่ผู้วิจัยเก็บรวบรวม จะถูกนำไปสังเคราะห์คุณลักษณะตามที่ได้กำหนดไว้ จากนั้นความคิดเห็นจะถูกจำแนกประเภทด้วยการเรียนรู้ของเครื่อง

จากรูปที่ 4.2 ข้อความความคิดเห็นที่ผู้วิจัยเก็บรวบรวม จะถูกนำไปสังเคราะห์คุณลักษณะตามที่ได้กำหนดไว้ จากนั้นข้อความประเภทเข้าข่ายหมิ่นประมาทจะถูกเพิ่มจำนวนด้วยขั้นตอนวิธี SMOT ให้มีจำนวนเท่ากับข้อความไม่เข้าข่ายหมิ่นประมาท ได้จำนวนเท่ากับ 1,047 ข้อความเข้าข่ายหมิ่นประมาท และ 1,047 ข้อความไม่เข้าขายหมิ่นประมาท รวมจำนวนข้อความทั้งหมดเท่ากับ 2,094 ข้อความ ชุดข้อความใหม่จะถูกจำแนกประเภทด้วยการเรียนรู้ของเครื่อง



รูปที่ 4.1 ภาพรวมของการทดลองโดยยังไม่แก้ไขปัญหาความไม่สมดุลของจำนวนประเภทข้อความ



รูปที่ 4.2 ภาพรวมของการทดลองโดยแก้ไขปัญหาความไม่สมดุลของจำนวนประเภทข้อความด้วย SMOT

จากคุณลักษณะ n-grams ระดับคำ และระดับตัวอักษรที่กล่าวถึงในบทที่ 3 ข้อที่ 3.2.1 นั้น มีคุณลักษณะจำนวนมาก ทำให้ผู้วิจัยใช้ SVD ลดจำนวนของคุณลักษณะ n-grams ที่ใช้ในการทดลอง เพื่อเพิ่มความเร็วในการจำแนกข้อความ และทดสอบประสิทธิภาพของการจำแนกหากจำนวนของคุณลักษณะของ n-grams มีจำนวนลดลง

คุณลักษณะที่ใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทประกอบด้วย n-grams ระดับคำ ได้แก่ 1-grams 2-grams และ 3-grams n-grams ระดับตัวอักษร ได้แก่ 2-grams 3-grams และ 4-grams คุณลักษณะจากคลังคำศัพท์ศาลฎีกา 8 รูปแบบ และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ 8 รูปแบบ ในการทดลอง ผู้วิจัยได้ทดลองจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการผสมผสานคุณลักษณะ n-grams ทั้งระดับคำและระดับตัวอักษร ดังนี้

1. Word 1-gram
2. Word 2-gram
3. Word 3-gram
4. Char 2-gram
5. Char 3-gram
6. Char 4-gram

7. All grams (Char 2-gram + 3-gram + 4-gram + Word 1-gram + Word 2-gram + Word 3-gram)

คุณลักษณะทั้ง 7 ข้อที่กล่าวมาข้างต้นจะถูกนำไปใช้ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทร่วมกับคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ เพื่อเปรียบเทียบประสิทธิภาพ

4.1 การวัดประสิทธิภาพ

ประสิทธิภาพของการจำแนกข้อความเข้าข่ายหมิ่นประมาท จะถูกวัดด้วยค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และ คะแนนเอฟวัน (F1) การคำนวณค่าเหล่านี้อาศัยการคำนวณดังตารางที่ 4.1

ตารางที่ 4.1 การนับจำนวนข้อความเพื่อใช้ในการวัดประสิทธิภาพ

		ประเภทของข้อความที่โปรแกรมทำนาย	
		เข้าข่าย	ไม่เข้าข่าย
ประเภทของข้อความ	เข้าข่าย	TP	FN
	ไม่เข้าข่าย	FP	TN

จากตารางที่ 4.1 TP คือ จำนวนของข้อความเข้าข่ายหมิ่นประมาท ที่โปรแกรมทำนายว่าเข้าข่ายหมิ่นประมาท, FN คือ จำนวนของข้อความที่เข้าข่ายหมิ่นประมาท ที่โปรแกรมทำนายว่าไม่เข้าข่าย, FP คือ ข้อความไม่เข้าข่ายหมิ่นประมาท ที่โปรแกรมทำนายว่าเข้าข่ายหมิ่นประมาท และ TN คือ ข้อความไม่เข้าข่ายหมิ่นประมาท ที่โปรแกรมทำนายว่าไม่เข้าข่ายหมิ่นประมาท ซึ่งการคำนวณค่าความแม่น (ACC) ค่าความเที่ยง (PRE) ค่าเรียกคืน (REC) และคะแนนเอฟวัน (F_1) คำนวณตามสมการต่อไปนี้

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} \tag{19}$$

$$PRE = \frac{TP}{TP+FP} \tag{20}$$

$$REC = \frac{TP}{TP+FN} \tag{21}$$

$$F_1 = \frac{2(PRE*REC)}{PRE+REC} \quad (22)$$

4.2 ข้อมูลที่ใช้ในการทดลอง

การทดลองในงานวิจัยนี้ ผู้วิจัยเก็บข้อมูลความคิดเห็นจากเฟซบุ๊กเพจ และกลุ่มบนเฟซบุ๊กได้ข้อความทั้งหมด 1,340 ข้อความ ซึ่งแยกเป็นข้อความเข้าข่ายหมิ่นประมาท 293 ข้อความ และข้อความไม่เข้าข่ายหมิ่นประมาท 1,047 ข้อความ โดยผู้วิจัยได้คัดกรองความคิดเห็นที่ขาดใจความสำคัญออกไป เช่น ความคิดเห็นที่เป็นเพียงตัวเลขแสดงถึงการหัวเราะ “5555” ความคิดเห็นที่ใช้เพียงสติ๊กเกอร์ หรือความคิดเห็นที่เป็นเพียงการแท็กเพื่อนตนเอง เป็นต้น แล้วส่งความคิดเห็นที่ผ่านการคัดกรองให้ผู้เชี่ยวชาญด้านกฎหมาย 3 คน จำแนกความคิดเห็นว่าเข้าข่ายหมิ่นประมาทหรือไม่ และผู้เชี่ยวชาญ 3 คน ด้านภาษาไทยในการระบุถึงชนิดของประโยคว่าเป็นประโยคชนิดใด ดังตัวอย่างในตารางที่ 4.2 ข้อมูลนี้ถูกนำไปใช้เป็นข้อมูลสอนดังตารางที่ 4.3

ตารางที่ 4.2 ตัวอย่างของข้อความที่ใช้ในการทดลอง

ข้อความ	ประเภทข้อความ	ชนิดของประโยค	เหตุผล
ศรีฆ่าน้อง ฟ้องนาย ขายเพื่อน	เข้าข่ายหมิ่นประมาท	ประโยคบอกเล่า	
มึงเป็นทหารกบฏ	เข้าข่ายหมิ่นประมาท	ประโยคบอกเล่า	
ไม่ต้องแก้ตัว กูเข้าใจว่าพวกมึงขายชาติ ถูกแล้ว	เข้าข่ายหมิ่นประมาท	ประโยคบอกเล่า	
ไอ้โชคมึงยังไม่ตายอีกรึ	**ไม่**เข้าข่ายหมิ่นประมาท	ประโยคคำถาม	ข้อความสาปแช่ง
ขอให้มึงตายไม่ดีไอ้เปรมชัย	**ไม่**เข้าข่ายหมิ่นประมาท	ประโยคขอร้อง	ข้อความสาปแช่ง
เอามันติดคุกเลย อีสัส	**ไม่**เข้าข่ายหมิ่นประมาท	ประโยคคำสั่ง	ข้อความสาปแช่ง
แสดงว่า ทหารชั่วมากเลยใช่ไหม	**ไม่**เข้าข่ายหมิ่นประมาท	ประโยคคำถาม	ไม่ได้เจาะจงตัวทหาร

ตารางที่ 4.3 ตัวอย่างข้อมูลสอน

Text	n-grams													Dictionary of Judgment Terms								Dependency Structure								Class
	Words										Chars																			
	ศรี	ฆ่า	น้อง	ฟ้อง	นาย	ขาย	เพื่อน	มึง	เป็น	...	ศร	รี	...	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	
ศรีฆ่าน้อง ฟ้องนาย ขายเพื่อน	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1
มึงเป็นทหารกบฏ	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	1	1
ไม่ต้องแก้ตัว กูเข้าใจว่าพวกมึงขายชาติ ถูกแล้ว	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	1
ไอ้โชคมึงยังไม่ตายอีกรึ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0
ขอให้มึงตายไม่ดีไอ้เปรมชัย	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0	0	0	0	0	0	0	0	0
เอามันติดคุกเลยอีสัส	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	5	0	0	0	0	0	1	0	0	1	0
แสดง ว่าทหาร ชั่วมากเลย ใช่ไหม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0

4.3 การทดลอง

ในการทดลองผู้วิจัยได้ใช้การเรียนรู้ของเครื่องทั้งหมด 3 ชนิด ได้แก่ SVM, Logistic Regression และ Multi-layer Perceptron ในการจำแนกประเภทข้อความเข้าข่ายหมิ่นประมาท และเปรียบเทียบประสิทธิภาพของการจำแนกข้อความ ซึ่งตัวแปรเสริมของการเรียนรู้ของเครื่องทั้ง 3 ชนิดเท่ากับค่าเริ่มต้นที่ Scikit-learn ได้ตั้งเอาไว้ ข้อความที่ใช้ในการทดลองจะถูกแบ่งเป็นข้อความสำหรับสอนและข้อความสำหรับทดสอบด้วยวิธี k-fold cross validation โดยกำหนดให้ k มีค่าเท่ากับ 10

คุณลักษณะจากคลังคำศัพท์ศาลฎีกา 8 รูปแบบ (term) และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ 8 รูปแบบ (dep) คือ คุณลักษณะที่ผู้วิจัยได้ออกแบบ ซึ่งในการทดลองจะใช้สองคุณลักษณะนี้ด้วยกันเสมอ (term+dep) และทดลองสับเปลี่ยนกับ n-grams (gram) เพื่อเปรียบเทียบประสิทธิภาพของคุณลักษณะ โดยที่คุณลักษณะ n-grams จะถูกลดขนาดด้วย SVD ทีละ 1,000 และแถวบนสุดของตารางแสดงถึงการใช้คุณลักษณะ n-grams โดยไม่มีการลดจำนวนของคุณลักษณะ

4.3.1 การทดลอง term+dep

ตารางที่ 4.4 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด ซึ่งจะเห็นได้ว่า Multi-layer Perceptron มีประสิทธิภาพดีที่สุดเมื่อเทียบกับ SVM และ Logistic Regression นอกจากนี้ SVM มีค่าความเที่ยงมากกว่า Logistic Regression ในขณะที่ Logistic Regression มีค่าเรียกคืนมากกว่า SVM เมื่อยังไม่ใช้ขั้นตอนวิธี SMOT ในการเพิ่มจำนวนข้อความเข้าข่ายหมิ่นประมาท แต่เมื่อเพิ่มจำนวนข้อความแล้ว Logistic Regression และ SVM มีประสิทธิภาพเท่ากัน

ตารางที่ 4.4 ผลการทดลองใช้คุณลักษณะ term+dep

	SVM				Logistic Regression				Multi-layer Perceptron			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	0.78	0.59	0.11	0.19	0.78	0.55	0.16	0.25	0.79	**0.65**	**0.18**	**0.28**
SMOT	0.62	0.63	0.64	0.63	0.62	0.63	0.64	0.63	0.67	**0.65**	**0.74**	**0.69**

4.3.2 การทดลองเปรียบเทียบ term+dep+word1-gram และ word1-gram

ตารางที่ 4.5 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 1-gram ระดับคำ (w1gram) เปรียบเทียบกับการใช้คุณลักษณะ 1-gram ระดับคำ (w1gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืนเพิ่มขึ้นเมื่อเทียบกับการใช้ 1-gram ระดับคำเพียงอย่างเดียว นอกจากนี้จำนวนของคุณลักษณะ 1-gram ระดับคำมีผลต่อค่าความเที่ยงและค่าเรียกคืน โดยที่การลดลงของจำนวนคุณลักษณะ 1-gram ระดับคำจะทำให้ค่าความเที่ยงเพิ่มขึ้น แต่ค่าเรียกคืนลดลง เมื่อไม่ใช้ SMOT แต่เมื่อใช้ SMOT ค่าความเที่ยงและค่าเรียกคืนลดลงเรื่อยๆตามจำนวนของ 1-gram ระดับคำ

การทดลองโดยไม่ใช้ SMOT SVM สามารถจำแนกข้อความได้ดีที่สุดเมื่อไม่ลดจำนวนของคุณลักษณะ 1-gram ระดับคำ F_1 เท่ากับ 0.41 แต่เมื่อคุณลักษณะ 1-gram ระดับคำมีจำนวนลดลง Multi-layer perceptron สามารถจำแนกข้อความได้ดีกว่า F_1 เท่ากับ 0.29 ในขณะที่ Logistic Regression มีค่าความเที่ยงมากที่สุด

การทดลองโดยใช้ SMOT Multi-layer perceptron มีความสามารถในการจำแนกได้ดีที่สุดเมื่อลด และไม่ลดจำนวนของคุณลักษณะ 1-gram ระดับคำ

ตารางที่ 4.5 ผลการทดลองใช้คุณลักษณะ term+dep+word1-gram และ word1-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+w1gram				w1gram				term+dep+w1gram				w1gram				term+dep+w1gram				w1gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	3143	0.76	0.46	**0.37**	**0.41**	0.75	0.43	0.31	0.36	0.79	**0.58**	0.28	0.38	0.79	0.59	0.26	0.36	0.76	0.43	0.28	0.34	0.74	0.36	0.24	0.29
	3000	0.74	0.41	0.36	0.38	0.75	0.44	0.33	0.38	0.79	0.58	0.28	0.38	0.79	0.59	0.26	0.36	0.76	0.42	0.29	0.34	0.75	0.39	0.29	0.33
	2000	0.75	0.44	0.37	0.40	0.75	0.45	0.37	0.41	0.79	0.58	0.29	0.39	0.78	0.58	0.27	0.37	0.76	0.44	0.29	0.35	0.74	0.40	0.27	0.32
	1000	0.74	0.43	0.36	0.39	0.73	0.40	0.34	0.37	0.79	0.58	0.31	0.40	0.79	0.59	0.28	0.38	0.74	0.41	0.33	0.37	0.74	0.41	0.32	0.36
	500	0.76	0.46	0.33	0.38	0.76	0.48	0.34	0.40	0.79	0.55	0.29	0.38	0.78	0.58	0.29	0.39	0.75	0.42	0.33	0.37	0.75	0.44	0.31	0.36
	50	0.78	0.59	0.14	0.23	0.77	0.36	0.02	0.04	0.79	**0.62**	0.16	0.25	0.77	0.15	0.01	0.02	0.74	0.36	**0.24**	**0.29**	0.73	0.33	0.20	0.25
SMOT	3143	0.91	0.86	0.99	0.92	0.91	0.87	0.98	0.92	0.89	0.86	0.95	0.90	0.90	0.87	0.94	0.90	0.92	**0.87**	**0.99**	**0.93**	0.91	0.86	0.99	0.92
	3000	0.87	0.81	0.98	0.89	0.85	0.78	0.98	0.87	0.87	0.83	0.92	0.87	0.85	0.82	0.90	0.86	0.88	0.81	0.98	0.89	0.84	0.77	0.98	0.86
	2000	0.87	0.80	0.98	0.88	0.83	0.76	0.97	0.85	0.86	0.81	0.93	0.87	0.82	0.79	0.88	0.83	0.87	0.81	0.98	0.89	0.84	0.77	0.98	0.86
	1000	0.84	0.78	0.95	0.86	0.80	0.74	0.94	0.83	0.80	0.76	0.88	0.82	0.76	0.73	0.85	0.79	0.87	0.81	0.98	0.89	0.84	0.77	0.98	0.86
	500	0.78	0.75	0.84	0.79	0.75	0.72	0.81	0.76	0.76	0.74	0.81	0.77	0.73	0.71	0.78	0.74	0.89	0.83	0.98	0.90	0.86	0.80	0.98	0.88
	50	0.65	0.66	0.62	0.64	0.60	0.62	0.57	0.59	0.64	0.66	0.61	0.63	0.59	0.60	0.55	0.57	0.88	**0.84**	**0.95**	**0.89**	0.86	0.81	0.93	0.87

4.3.3 การทดลองเปรียบเทียบ term+dep+word2-gram และ word2-gram

ตารางที่ 4.6 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 2-gram ระดับคำ (w2gram) เปรียบเทียบกับการใช้คุณลักษณะ 2-gram ระดับคำ (w2gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ 2-gram ระดับคำเพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT จำนวนของคุณลักษณะ 2-gram ระดับคำระหว่าง 4,000 ถึง 1,000 ทำให้การเรียนรู้ของเครื่องทั้ง 3 ชนิด มีประสิทธิภาพในการจำแนกข้อความที่ดี ดังนี้ 1) ที่จำนวน 4,000 Multi-layer perceptron จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.33 2) ที่จำนวน 2,000 ถึง 1,000 Logistic Regression จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.31 3) ที่จำนวน 1,000 SVM จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.35 นอกจากนี้เมื่อใช้ SMOT ค่าความเที่ยง และค่าเรียกคืนจะลดลงเมื่อขนาดของมิติ 2-gram ระดับคำลดลง

การทดลองโดยไม่ใช้ SMOT SVM สามารถจำแนกข้อความได้ดีที่สุดเมื่อไม่มีการลดจำนวนคุณลักษณะ 2-gram ระดับคำด้วย F1 เท่ากับ 0.26 ในขณะที่ Logistic Regression มีค่าความเที่ยงสูงที่สุด เมื่อจำนวนของ 2-gram ระดับคำลดลง Multi-layer Perceptron มีความสามารถในการจำแนกข้อความดีที่สุด

การทดลองโดยใช้ SMOT Logistic Regression สามารถจำแนกข้อความได้ดีที่สุด เมื่อไม่มีการลดจำนวนคุณลักษณะ 2-gram ระดับคำ เมื่อจำนวนคุณลักษณะ 2-gram ระดับคำมีจำนวนลดลง Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุด

ตารางที่ 4.6 ผลการทดลองใช้คุณลักษณะ term+dep+word2-gram และ word2-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+w2gram				w2gram				term+dep+w2gram				w2gram				term+dep+w2gram				w2gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	13952	0.79	0.59	**0.17**	**0.26**	0.78	0.55	0.09	0.15	0.80	**0.71**	0.15	0.25	0.78	0.52	0.03	0.06	0.78	0.50	0.15	0.23	0.78	**0.**62	0.07	0.13
	13000	0.79	0.57	0.21	0.31	0.78	0.47	0.10	0.16	0.80	0.69	0.14	0.23	0.78	0.47	0.04	0.07	0.75	0.42	0.22	0.29	0.77	0.43	0.11	0.18
	12000	0.78	0.52	0.24	0.33	0.77	0.41	0.16	0.23	0.80	0.66	0.18	0.28	0.78	0.55	0.06	0.11	0.70	0.31	0.33	0.32	0.76	0.37	0.20	0.26
	11000	0.76	0.43	0.27	0.33	0.76	0.41	0.19	0.26	0.79	0.63	0.20	0.30	0.78	0.59	0.07	0.13	0.70	0.33	0.34	0.33	0.73	0.31	0.22	0.26
	10000	0.76	0.44	0.24	0.31	0.76	0.41	0.19	0.26	0.79	0.61	0.20	0.30	0.78	0.61	0.07	0.13	0.72	0.35	0.30	0.32	0.75	0.36	0.18	0.24
	9000	0.77	0.47	0.22	0.30	0.77	0.44	0.19	0.27	0.79	0.61	0.17	0.27	0.78	0.56	0.07	0.12	0.73	0.34	0.24	0.28	0.75	0.37	0.17	0.23
	8000	0.77	0.49	0.20	0.28	0.77	0.42	0.16	0.23	0.79	0.60	0.16	0.25	0.78	0.56	0.07	0.12	0.75	0.39	0.21	0.27	0.76	0.41	0.16	0.23
	7000	0.78	0.50	0.20	0.29	0.77	0.45	0.18	0.26	0.79	0.61	0.16	0.25	0.78	0.61	0.07	0.13	0.75	0.39	0.21	0.27	0.76	0.39	0.15	0.22
	6000	0.77	0.47	0.20	0.28	0.77	0.47	0.20	0.28	0.79	0.59	0.14	0.23	0.78	0.62	0.07	0.13	0.74	0.35	0.20	0.25	0.76	0.39	0.18	0.25
	5000	0.76	0.46	0.23	0.31	0.76	0.44	0.20	0.28	0.79	0.61	0.18	0.28	0.79	0.66	0.09	0.16	0.74	0.37	0.26	0.31	0.74	0.37	0.22	0.28
	4000	0.76	0.48	0.24	0.32	0.76	0.47	0.23	0.31	0.79	0.61	0.20	0.30	0.78	0.60	0.10	0.17	0.75	0.40	0.28	0.33	0.75	0.41	0.24	0.30
	3000	0.76	0.46	0.25	0.32	0.75	0.40	0.20	0.27	0.79	0.58	0.20	0.30	0.78	0.59	0.11	0.19	0.73	0.35	0.25	0.29	0.72	0.34	0.23	0.27
	2000	0.76	0.47	0.27	0.34	0.75	0.39	0.19	0.26	0.79	0.61	0.21	0.31	0.78	0.49	0.10	0.17	0.71	0.31	0.23	0.26	0.73	0.33	0.23	0.27
	1000	0.77	0.51	0.27	0.35	0.76	0.46	0.18	0.26	0.78	0.56	0.21	0.31	0.77	0.52	0.10	0.17	0.71	0.33	0.26	0.29	0.71	0.32	0.23	0.27
	500	0.76	0.45	0.19	0.27	0.75	0.33	0.06	0.10	0.78	0.53	0.19	0.28	0.77	0.48	0.04	0.07	0.72	0.31	0.26	0.28	0.72	0.29	0.19	0.23
	50	0.78	0.59	0.12	0.20	0.78	0.25	0.01	0.02	0.78	0.56	0.16	0.25	0.78	0.10	0.00	0.00	0.79	**0.62**	**0.20**	**0.30**	0.78	0.31	0.02	0.04
SMOT	13952	0.85	0.78	0.98	0.87	0.74	0.67	0.98	0.80	0.88	**0.83**	**0.98**	**0.90**	0.89	0.87	0.93	0.90	0.84	0.77	0.99	0.87	0.78	0.70	0.99	0.82
	13000	0.83	0.76	0.99	0.86	0.70	0.63	0.99	0.77	0.86	0.79	0.98	0.87	0.78	0.70	0.98	0.82	0.83	0.75	0.99	0.85	0.75	0.67	0.99	0.80
	12000	0.77	0.69	0.99	0.81	0.70	0.63	0.99	0.77	0.82	0.75	0.97	0.85	0.78	0.70	0.98	0.82	0.76	0.67	0.99	0.80	0.74	0.66	0.99	0.79
	11000	0.78	0.70	0.99	0.82	0.70	0.63	0.99	0.77	0.82	0.75	0.97	0.85	0.77	0.69	0.98	0.81	0.79	0.71	0.99	0.83	0.73	0.65	0.99	0.78
	10000	0.81	0.73	0.98	0.84	0.73	0.66	0.98	0.79	0.83	0.78	0.93	0.85	0.79	0.71	0.98	0.82	0.81	0.73	0.99	0.84	0.76	0.69	0.97	0.81
	9000	0.83	0.76	0.98	0.86	0.77	0.69	0.99	0.81	0.84	0.80	0.91	0.85	0.81	0.75	0.93	0.83	0.84	0.77	0.97	0.86	0.80	0.75	0.93	0.83
	8000	0.84	0.78	0.95	0.86	0.79	0.71	0.99	0.83	0.84	0.81	0.89	0.85	0.84	0.82	0.87	0.84	0.83	0.77	0.96	0.85	0.81	0.76	0.92	0.83
	7000	0.86	0.81	0.94	0.87	0.81	0.75	0.95	0.84	0.84	0.82	0.88	0.85	0.84	0.83	0.85	0.84	0.85	0.80	0.95	0.87	0.83	0.78	0.91	0.84
	6000	0.83	0.78	0.92	0.84	0.81	0.74	0.95	0.83	0.82	0.79	0.87	0.83	0.82	0.82	0.82	0.82	0.85	0.80	0.93	0.86	0.82	0.79	0.89	0.84
	5000	0.81	0.77	0.90	0.83	0.78	0.71	0.93	0.81	0.80	0.77	0.84	0.80	0.79	0.80	0.79	0.79	0.83	0.78	0.92	0.84	0.81	0.78	0.86	0.82
	4000	0.80	0.77	0.88	0.82	0.78	0.75	0.86	0.80	0.78	0.76	0.82	0.79	0.78	0.79	0.76	0.77	0.84	0.79	0.93	0.85	0.80	0.78	0.85	0.81
	3000	0.78	0.75	0.85	0.80	0.77	0.76	0.78	0.77	0.76	0.75	0.79	0.77	0.74	0.76	0.70	0.73	0.84	0.79	0.93	0.85	0.79	0.78	0.82	0.80
	2000	0.75	0.72	0.82	0.77	0.73	0.75	0.71	0.73	0.73	0.72	0.76	0.74	0.71	0.75	0.63	0.68	0.83	0.78	0.93	0.85	0.76	0.77	0.75	0.76
	1000	0.70	0.70	0.73	0.71	0.67	0.70	0.58	0.63	0.69	0.69	0.69	0.69	0.65	0.70	0.53	0.60	0.83	0.78	0.93	0.85	0.79	0.73	0.92	0.81
	500	0.69	0.69	0.71	0.70	0.60	0.63	0.50	0.56	0.67	0.68	0.68	0.68	0.59	0.63	0.45	0.53	0.85	0.80	0.93	0.86	0.82	0.76	0.93	0.84
	50	0.64	0.66	0.61	0.63	0.56	0.60	0.40	0.48	0.64	0.66	0.62	0.64	0.57	0.62	0.39	0.48	0.78	**0.76**	**0.82**	**0.79**	0.62	0.64	0.55	0.59

4.3.4 การทดลองเปรียบเทียบ term+dep+word3-gram และ word3-gram

ตารางที่ 4.7 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 3-gram ระดับคำ (w3gram) เปรียบเทียบกับการใช้คุณลักษณะ 3-gram ระดับคำ (w3gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ 3-gram ระดับคำเพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT จำนวนของคุณลักษณะ 3-gram ระดับคำเท่ากับ 14,000 ทำให้การเรียนรู้ของเครื่องทั้ง 2 ชนิด มีประสิทธิภาพในการจำแนกข้อความที่ดี ดังนี้ 1) Logistic Regression จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.30 2) SVM จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.38 นอกจากนี้เมื่อใช้ SMOT ค่าความเทียง และค่าเรียกคืนจะลดลงเมื่อ ขนาดของ 3-gram ระดับคำลดลง

การทดลองโดยไม่ใช้ SMOT Multi-layer Perceptron มีความสามารถในการจำแนกข้อความได้ดีที่สุดเมื่อไม่มีการลดขนาดคุณลักษณะ 3-gram ระดับคำด้วย F_1 เท่ากับ 0.37 ในขณะที่ Logistic Regression มีค่าความเที่ยงสูงที่สุด เมื่อจำนวนของ 3-gram ระดับคำลดลง Multi-layer Perceptron มีความสามารถในการจำแนกข้อความดีที่สุด

การทดลองโดยใช้ SMOT Logistic Regression ที่ใช้ร่วมกับคุณลักษณะ 3-gram สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.89 แต่ SVM มีค่าเรียกคืนสูงที่สุด เมื่อจำนวนของคุณลักษณะ 3-gram มีจำนวนลดลง Multi-layer Perceptron ที่ใช้ทั้ง 3 คุณลักษณะสามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.70

ตารางที่ 4.7 ผลการทดลองใช้คุณลักษณะ term+dep+word3-gram และ word3-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+w3gram				w3gram				term+dep+w3gram				w3gram				term+dep+w3gram				w3gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	17920	0.79	0.65	0.15	0.24	0.78	0.00	0.00	0.00	0.79	**0.67**	0.10	0.17	0.78	0.00	0.00	0.00	0.74	0.41	**0.34**	**0.37**	0.78	0.20	0.01	0.02
	17000	0.78	0.51	0.11	0.18	0.78	0.35	0.02	0.04	0.79	0.72	0.09	0.16	0.78	0.00	0.00	0.00	0.73	0.34	0.25	0.29	0.78	0.27	0.02	0.04
	16000	0.79	0.58	0.16	0.25	0.78	0.39	0.04	0.07	0.79	0.65	0.09	0.16	0.78	0.00	0.00	0.00	0.71	0.30	0.28	0.29	0.77	0.40	0.06	0.10
	15000	0.77	0.49	0.25	0.33	0.77	0.31	0.05	0.09	0.78	0.57	0.14	0.22	0.78	0.10	0.00	0.00	0.74	0.38	0.27	0.32	0.76	0.37	0.09	0.14
	14000	0.77	0.50	0.31	0.38	0.76	0.33	0.08	0.13	0.79	0.62	0.20	0.30	0.78	0.10	0.01	0.02	0.71	0.36	0.33	0.34	0.75	0.35	0.11	0.17

ตารางที่ 4.7 ผลการทดลองใช้คุณลักษณะ term+dep+word3-gram และ word3-gram (ต่อ)

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+w3gram				w3gram				term+dep+w3gram				w3gram				term+dep+w3gram				w3gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
	13000	0.77	0.48	0.26	0.34	0.76	0.37	0.07	0.12	0.79	0.62	0.19	0.29	0.78	0.20	0.01	0.02	0.74	0.39	0.32	0.35	0.75	0.35	0.11	0.17
	12000	0.77	0.45	0.17	0.25	0.76	0.38	0.07	0.12	0.79	0.62	0.15	0.24	0.78	0.20	0.01	0.02	0.76	0.45	0.21	0.29	0.75	0.32	0.09	0.14
	11000	0.77	0.43	0.13	0.20	0.76	0.37	0.06	0.10	0.78	0.55	0.10	0.17	0.78	0.20	0.01	0.02	0.76	0.42	0.16	0.23	0.76	0.38	0.09	0.15
	10000	0.77	0.48	0.10	0.17	0.77	0.40	0.06	0.10	0.79	0.60	0.10	0.17	0.78	0.20	0.01	0.02	0.77	0.43	0.15	0.22	0.76	0.37	0.08	0.13
	9000	0.77	0.51	0.10	0.17	0.77	0.44	0.06	0.11	0.79	0.61	0.10	0.17	0.78	0.10	0.00	0.00	0.77	0.44	0.14	0.21	0.77	0.43	0.09	0.15
	8000	0.77	0.46	0.11	0.18	0.77	0.46	0.06	0.11	0.78	0.55	0.10	0.17	0.78	0.10	0.01	0.02	0.76	0.42	0.16	0.23	0.76	0.40	0.08	0.13
	7000	0.78	0.53	0.12	0.20	0.78	0.51	0.08	0.14	0.79	0.58	0.12	0.20	0.78	0.10	0.01	0.02	0.77	0.49	0.16	0.24	0.77	0.47	0.10	0.16
	6000	0.78	0.56	0.13	0.21	0.77	0.47	0.07	0.12	0.79	0.58	0.11	0.18	0.78	0.20	0.01	0.02	0.77	0.47	0.16	0.24	0.77	0.39	0.09	0.15
	5000	0.78	0.60	0.14	0.23	0.77	0.49	0.09	0.15	0.79	0.63	0.14	0.23	0.78	0.35	0.02	0.04	0.78	0.54	0.21	0.30	0.77	0.46	0.11	0.18
	4000	0.78	0.55	0.16	0.25	0.77	0.37	0.07	0.12	0.79	0.65	0.15	0.24	0.78	0.30	0.01	0.02	0.78	0.53	0.22	0.31	0.77	0.38	0.09	0.15
	3000	0.78	0.62	0.15	0.24	0.77	0.27	0.06	0.10	0.79	0.65	0.14	0.23	0.78	0.35	0.01	0.02	0.78	0.58	0.20	0.30	0.77	0.32	0.09	0.14
	2000	0.78	0.66	0.14	0.23	0.77	0.30	0.06	0.10	0.79	0.63	0.13	0.22	0.78	0.22	0.01	0.02	0.78	0.59	0.21	0.31	0.77	0.39	0.07	0.12
	1000	0.78	0.60	0.12	0.20	0.78	0.24	0.03	0.05	0.78	0.54	0.14	0.22	0.78	0.14	0.01	0.02	0.78	0.58	0.19	0.29	0.77	0.32	0.03	0.05
	500	0.78	0.51	0.12	0.19	0.78	0.10	0.00	0.00	0.78	0.55	0.16	0.25	0.78	0.00	0.00	0.00	0.78	0.59	0.23	0.33	0.77	0.00	0.00	0.00
	50	0.79	0.59	0.12	0.20	0.78	0.00	0.00	0.00	0.78	0.55	0.16	0.25	0.78	0.00	0.00	0.00	0.79	**0.67**	**0.17**	**0.27**	0.78	0.00	0.00	0.00
SMOT	17920	0.83	0.75	0.99	0.85	0.54	0.52	**1.00**	0.68	0.87	0.80	0.98	0.88	0.86	**0.84**	0.94	**0.89**	0.74	0.66	0.99	0.79	0.65	0.60	0.99	0.75
	17000	0.74	0.67	0.98	0.80	0.56	0.53	1.00	0.69	0.81	0.76	0.93	0.84	0.86	0.83	0.93	0.88	0.78	0.71	0.98	0.82	0.82	0.79	0.95	0.86
	16000	0.73	0.66	0.98	0.79	0.57	0.54	1.00	0.70	0.81	0.77	0.90	0.83	0.72	0.69	0.93	0.79	0.77	0.70	0.98	0.82	0.62	0.57	0.99	0.72
	15000	0.70	0.63	0.98	0.77	0.57	0.54	1.00	0.70	0.79	0.75	0.89	0.81	0.62	0.57	0.99	0.72	0.76	0.69	0.97	0.81	0.62	0.57	0.99	0.72
	14000	0.69	0.63	0.97	0.76	0.57	0.54	1.00	0.70	0.78	0.74	0.88	0.80	0.61	0.56	0.99	0.72	0.74	0.67	0.97	0.79	0.61	0.56	0.99	0.72
	13000	0.74	0.68	0.93	0.79	0.58	0.54	0.99	0.70	0.78	0.74	0.87	0.80	0.62	0.58	0.95	0.72	0.78	0.72	0.94	0.82	0.62	0.57	0.99	0.72
	12000	0.79	0.75	0.89	0.81	0.63	0.60	0.94	0.73	0.78	0.76	0.84	0.80	0.80	0.83	0.79	0.81	0.82	0.79	0.89	0.84	0.80	0.81	0.81	0.81
	11000	0.81	0.79	0.85	0.82	0.82	0.88	0.75	0.81	0.79	0.78	0.82	0.80	0.82	0.89	0.72	0.80	0.84	0.82	0.88	0.85	0.82	0.86	0.77	0.81
	10000	0.81	0.81	0.82	0.81	0.81	0.89	0.71	0.79	0.79	0.80	0.79	0.79	0.80	0.91	0.68	0.78	0.83	0.82	0.85	0.83	0.82	0.87	0.75	0.81
	9000	0.81	0.81	0.82	0.81	0.81	0.90	0.70	0.79	0.80	0.81	0.78	0.79	0.80	0.92	0.67	0.78	0.84	0.84	0.85	0.84	0.82	0.87	0.75	0.81
	8000	0.81	0.82	0.81	0.81	0.80	0.90	0.68	0.77	0.79	0.81	0.78	0.79	0.79	0.92	0.64	0.75	0.85	0.84	0.86	0.85	0.82	0.87	0.74	0.80
	7000	0.80	0.81	0.79	0.80	0.78	0.91	0.62	0.74	0.79	0.80	0.77	0.78	0.77	0.92	0.59	0.72	0.84	0.84	0.84	0.84	0.79	0.87	0.68	0.76
	6000	0.77	0.80	0.72	0.76	0.74	0.90	0.55	0.68	0.75	0.78	0.71	0.74	0.74	0.92	0.52	0.66	0.81	0.82	0.79	0.80	0.75	0.86	0.59	0.70
	5000	0.76	0.80	0.71	0.75	0.72	0.89	0.50	0.64	0.73	0.75	0.71	0.73	0.71	0.91	0.48	0.63	0.79	0.82	0.75	0.78	0.72	0.86	0.54	0.66
	4000	0.74	0.77	0.70	0.73	0.69	0.88	0.44	0.59	0.73	0.75	0.71	0.73	0.68	0.88	0.43	0.58	0.78	0.79	0.76	0.77	0.69	0.86	0.47	0.61
	3000	0.71	0.74	0.66	0.70	0.66	0.87	0.38	0.53	0.70	0.72	0.68	0.70	0.65	0.87	0.37	0.52	0.76	0.78	0.73	0.75	0.67	0.86	0.41	0.56
	2000	0.69	0.72	0.66	0.69	0.63	0.86	0.30	0.44	0.68	0.70	0.65	0.67	0.62	0.86	0.29	0.43	0.75	0.76	0.73	0.74	0.63	0.86	0.32	0.47
	1000	0.66	0.68	0.62	0.65	0.54	0.70	0.13	0.22	0.65	0.67	0.63	0.65	0.54	0.70	0.13	0.22	0.72	0.73	0.71	0.72	0.53	0.62	0.38	0.47
	500	0.64	0.65	0.61	0.63	0.50	0.51	0.87	0.64	0.63	0.64	0.62	0.63	0.50	0.50	0.87	0.64	0.72	0.72	0.72	0.72	0.52	0.55	0.70	0.62
	50	0.64	0.65	0.61	0.63	0.49	0.50	0.88	0.64	0.63	0.65	0.61	0.63	0.50	0.50	**0.97**	0.66	0.70	**0.71**	0.70	**0.70**	0.51	0.52	0.71	0.60

4.3.5 การทดลองเปรียบเทียบ term+dep+char2-gram และ char2-gram

ตารางที่ 4.8 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 2-gram ระดับตัวอักษร (c2gram) เปรียบเทียบกับการใช้คุณลักษณะ 2-gram ระดับตัวอักษร (c2gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ 2-gram ระดับตัวอักษรเพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT เมื่อจำนวนของคุณลักษณะ 2-gram ระดับตัวอักษรระหว่าง 1,000 ถึง 500 ทำให้การเรียนรู้ของเครื่องทั้ง 3 ชนิด มีประสิทธิภาพในการจำแนกข้อความดีที่สุด ดังนี้ 1) ที่จำนวน 1,000 และ 500 Logistic Regression จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.44 2) ที่จำนวน 1,000 SVM จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.41 3) ที่จำนวน 1,000 Multi-layer Perceptron จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.42 นอกจากนี้เมื่อใช้ SMOT ค่าความเที่ยง และค่าเรียกคืนจะลดลงเมื่อ ขนาดของ 2-gram ระดับตัวอักษรลดลง

การทดลองโดยไม่ใช้ SMOT Logistic Regression สามารถจำแนกข้อความได้ดีที่สุดเมื่อไม่มีการลดขนาดคุณลักษณะ 2-gram ระดับตัวอักษรด้วย F_1 เท่ากับ 0.43 แต่ SVM มีค่าเรียกคืนสูงที่สุด เมื่อจำนวนของคุณลักษณะ 2-gram ระดับคำลดลง Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.37 ในขณะที่ SVM ที่ใช้ร่วมกับคุณลักษณะ 2-gram ระดับคำมีค่าความเที่ยงมากที่สุด

การทดลองโดยใช้ SMOT Multi-layer Perceptron ที่ใช้ร่วมกับคุณลักษณะ 2-gram ระดับตัวอักษรมีความสามารถในการจำแนกข้อความได้ดีที่สุดเมื่อไม่มีการลดจำนวนของคุณลักษณะ 2-gram ระดับคำด้วย F_1 เท่ากับ 0.91 แต่เมื่อจำนวนของคุณลักษณะ 2-gram ระดับคำลดลง Multi-layer Perceptron ที่ใช้คุณลักษณะทั้ง 3 ชนิดสามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.91

ตารางที่ 4.8 ผลการทดลองใช้คุณลักษณะ term+dep+char2-gram และ char2-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+c2gram				c2gram				term+dep+c2gram				c2gram				term+dep+c2gram				c2gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	1566	0.72	0.39	**0.39**	0.39	0.71	0.39	0.38	0.38	0.76	**0.50**	0.37	**0.43**	0.76	0.49	0.35	0.41	0.76	0.49	0.32	0.39	0.74	0.40	0.27	0.32
	1000	0.72	0.41	0.42	0.41	0.72	0.42	0.39	0.40	0.76	0.51	0.38	0.44	0.76	0.49	0.34	0.40	0.76	0.49	0.36	0.42	0.75	0.45	0.30	0.36
	500	0.72	0.39	0.40	0.39	0.72	0.39	0.37	0.38	0.77	0.52	0.38	0.44	0.76	0.50	0.36	0.42	0.77	0.50	0.35	0.41	0.76	0.48	0.33	0.39
	50	0.79	0.59	0.21	0.31	0.79	**0.66**	0.13	0.22	0.80	0.60	0.23	0.33	0.78	0.57	0.11	0.18	0.76	0.44	**0.32**	**0.37**	0.73	0.36	0.27	0.31

ตารางที่ 4.8 ผลการทดลองใช้คุณลักษณะ term+dep+char2-gram และ char2-gram (ต่อ)

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+c2gram				c2gram				term+dep+c2gram				c2gram				term+dep+c2gram				c2gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
SMOT	1566	0.88	0.82	0.98	0.89	0.88	0.82	0.98	0.89	0.87	0.83	0.95	0.89	0.87	0.83	0.95	0.89	0.89	0.84	0.98	0.90	0.90	**0.85**	**0.98**	**0.91**
	1000	0.88	0.82	0.98	0.89	0.88	0.82	0.97	0.89	0.87	0.82	0.95	0.88	0.87	0.82	0.94	0.88	0.90	0.85	0.98	0.91	0.89	0.84	0.97	0.90
	500	0.84	0.80	0.94	0.86	0.84	0.79	0.92	0.85	0.83	0.80	0.89	0.84	0.83	0.80	0.89	0.84	0.91	0.86	0.98	0.92	0.91	0.86	0.98	0.92
	50	0.70	0.71	0.68	0.69	0.67	0.68	0.65	0.66	0.68	0.69	0.66	0.67	0.65	0.66	0.63	0.64	0.90	**0.85**	**0.97**	**0.91**	0.89	0.83	0.97	0.89

4.3.6 การทดลองเปรียบเทียบ term+dep+char3-gram และ char3-gram

ตารางที่ 4.9 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 3-gram ระดับตัวอักษร (c3gram) เปรียบเทียบกับการใช้คุณลักษณะ 3-gram ระดับตัวอักษร (c3gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ 3-gram ระดับตัวอักษรเพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT เมื่อจำนวนของคุณลักษณะ 3-gram ระดับตัวอักษรระหว่าง 3,000 ถึง 1,000 ทำให้การเรียนรู้ของเครื่องทั้ง 3 ชนิด มีประสิทธิภาพในการจำแนกข้อความดีที่สุด ดังนี้ 1) ที่จำนวน 3,000 Logistic Regression และ SVM จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.45 และ 0.41 ตามลำดับ 2) ที่จำนวน 1,000 Multi-layer Perceptron จำแนกข้อความได้ดีด้วย F_1 เท่ากับ 0.42 นอกจากนี้เมื่อใช้ SMOT ค่าความเที่ยง และค่าเรียกคืนจะลดลงเมื่อ ขนาดของมิติ 3-gram ระดับตัวอักษรลดลง

การทดลองโดยไม่ใช้ SMOT Logistic Regression สามารถจำแนกข้อความได้ดีที่สุดเมื่อจำนวนของคุณลักษณะ 3-gram ระดับตัวอักษรไม่ลดลงด้วย F_1 เท่ากับ 0.43 ในขณะที่ SVM มีค่าเรียกคืนสูงที่สุด แต่เมื่อจำนวนของคุณลักษณะ 3-gram ระดับคำมีจำนวนลดลง Multi-layer Perceptron มีความสามารถในการจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.32 ในขณะที่ SVM มีความเที่ยงสูงที่สุด

การทดลองโดยใช้ SMOT Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดเมื่อขนาดของคุณลักษณะ 3-gram ไม่ลดลง และเมื่อจำนวนคุณลักษณะน้อยที่สุดด้วย F_1 เท่ากับ 0.92 และ 0.89 ตามลำดับ

ตารางที่ 4.9 ผลการทดลองใช้คุณลักษณะ term+dep+char3-gram และ char3-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+c3gram				c3gram				term+dep+c3gram				c3gram				term+dep+c3gram				c3gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	5143	0.75	0.46	**0.38**	0.42	0.73	0.43	0.35	0.39	0.78	**0.58**	0.34	**0.43**	0.77	0.56	0.30	0.39	0.75	0.41	0.27	0.33	0.73	0.33	0.24	0.28
	5000	0.74	0.43	0.37	0.40	0.73	0.42	0.36	0.39	0.78	0.58	0.34	0.43	0.77	0.56	0.31	0.40	0.75	0.41	0.28	0.33	0.73	0.33	0.22	0.26
	4000	0.73	0.43	0.37	0.40	0.73	0.41	0.34	0.37	0.78	0.57	0.34	0.43	0.77	0.56	0.32	0.41	0.76	0.46	0.31	0.37	0.73	0.34	0.22	0.27
	3000	0.74	0.45	0.38	0.41	0.73	0.43	0.34	0.38	0.78	0.59	0.36	0.45	0.77	0.55	0.31	0.40	0.76	0.45	0.29	0.35	0.74	0.39	0.25	0.30
	2000	0.73	0.42	0.37	0.39	0.73	0.42	0.38	0.40	0.78	0.56	0.36	0.44	0.77	0.56	0.33	0.42	0.77	0.49	0.33	0.39	0.75	0.41	0.29	0.34
	1000	0.73	0.42	0.37	0.39	0.74	0.44	0.38	0.41	0.78	0.55	0.34	0.42	0.77	0.56	0.33	0.42	0.76	0.49	0.37	0.42	0.76	0.49	0.35	0.41
	500	0.76	0.48	0.34	0.40	0.75	0.44	0.29	0.35	0.79	0.58	0.33	0.42	0.77	0.57	0.26	0.36	0.77	0.50	0.31	0.38	0.76	0.46	0.27	0.34
	50	0.80	**0.61**	0.19	0.29	0.78	0.54	0.05	0.09	0.79	0.55	0.18	0.27	0.78	0.52	0.03	0.06	0.73	0.38	**0.27**	**0.32**	0.73	0.35	0.21	0.26
SMOT	5143	0.90	0.85	0.98	0.91	0.90	0.85	0.98	0.91	0.89	0.85	0.96	0.90	0.89	0.85	0.96	0.90	0.91	**0.86**	**0.98**	**0.92**	0.90	0.85	0.98	0.91
	5000	0.90	0.84	0.99	0.91	0.90	0.84	0.98	0.90	0.89	0.85	0.96	0.90	0.88	0.84	0.95	0.89	0.91	0.86	0.98	0.92	0.90	0.84	0.98	0.90
	4000	0.90	0.84	0.99	0.91	0.89	0.84	0.98	0.90	0.88	0.84	0.96	0.90	0.88	0.84	0.95	0.89	0.91	0.86	0.98	0.92	0.90	0.85	0.98	0.91
	3000	0.89	0.83	0.99	0.90	0.89	0.83	0.98	0.90	0.88	0.83	0.96	0.89	0.87	0.83	0.95	0.89	0.90	0.85	0.99	0.91	0.90	0.84	0.98	0.90
	2000	0.89	0.83	0.98	0.90	0.88	0.82	0.98	0.89	0.87	0.83	0.95	0.89	0.87	0.82	0.94	0.88	0.90	0.85	0.99	0.91	0.89	0.84	0.98	0.90
	1000	0.85	0.79	0.96	0.87	0.85	0.79	0.96	0.87	0.81	0.78	0.87	0.82	0.80	0.77	0.87	0.82	0.90	0.85	0.99	0.91	0.90	0.85	0.98	0.91
	500	0.79	0.76	0.85	0.80	0.77	0.74	0.84	0.79	0.77	0.75	0.82	0.78	0.76	0.73	0.82	0.77	0.90	0.85	0.98	0.91	0.90	0.85	0.98	0.91
	50	0.68	0.70	0.64	0.67	0.62	0.64	0.58	0.61	0.65	0.67	0.62	0.64	0.60	0.62	0.57	0.59	0.88	**0.83**	**0.96**	**0.89**	0.87	0.82	0.95	0.88

4.3.7 การทดลองเปรียบเทียบ term+dep+char4-gram และ char4-gram

ตารางที่ 4.10 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ 4-gram ระดับตัวอักษร (c4gram) เปรียบเทียบกับการใช้คุณลักษณะ 4-gram ระดับตัวอักษร (c4gram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ 4-gram ระดับตัวอักษรเพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT เมื่อจำนวนของคุณลักษณะ 4-gram ระดับตัวอักษรที่จำนวนเท่ากับ 5,793 4,000 และ 1,000 ทำให้ SVM สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.43 ที่จำนวน 5,000 2,000 และ 1,000 Logistic Regression สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.42 และที่จำนวน 1,000 Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.39

การทดลองโดยไม่ใช้ SMOT SVM มีความสามารถในการจำแนกดีที่สุดเมื่อไม่มีการลดจำนวนคุณลักษณะ 4-gram ในขณะที่ Logistic Regression มีค่าความเที่ยงมากที่สุด แต่เมื่อจำนวน

คุณลักษณะ 4-gram มีจำนวนลดลง Multi-layer Perceptron มีความสามารถในการจำแนกดีที่สุดด้วย F_1 เท่ากับ 0.29 ในขณะที่ SVM มีค่าความเที่ยงมากที่สุด

การทดลองโดยใช้ SMOT Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดเมื่อขนาดของคุณลักษณะ 4-gram ไม่ลดลง และเมื่อจำนวนคุณลักษณะน้อยที่สุดด้วย F_1 เท่ากับ 0.92 และ 0.88 ตามลำดับ

ตารางที่ 4.10 ผลการทดลองใช้คุณลักษณะ term+dep+char4-gram และ char4-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+c4gram				c4gram				term+dep+c4gram				c4gram				term+dep+c4gram				c4gram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	5793	0.76	0.49	**0.39**	**0.43**	0.74	0.43	0.32	0.37	0.79	**0.57**	0.30	0.39	0.78	0.58	0.25	0.35	0.75	0.43	0.31	0.36	0.73	0.35	0.27	0.30
	5000	0.75	0.45	0.40	0.42	0.73	0.38	0.31	0.34	0.79	0.58	0.33	0.42	0.78	0.55	0.26	0.35	0.74	0.40	0.31	0.35	0.71	0.33	0.28	0.30
	4000	0.76	0.47	0.40	0.43	0.75	0.43	0.31	0.36	0.79	0.59	0.32	0.41	0.78	0.57	0.25	0.35	0.76	0.43	0.30	0.35	0.74	0.40	0.30	0.34
	3000	0.76	0.47	0.38	0.42	0.74	0.39	0.29	0.33	0.78	0.58	0.32	0.41	0.78	0.56	0.23	0.33	0.75	0.42	0.29	0.34	0.73	0.34	0.28	0.31
	2000	0.76	0.48	0.38	0.42	0.75	0.43	0.32	0.37	0.79	0.59	0.33	0.42	0.78	0.56	0.26	0.36	0.74	0.39	0.29	0.33	0.73	0.36	0.30	0.33
	1000	0.77	0.50	0.38	0.43	0.75	0.45	0.31	0.37	0.79	0.58	0.33	0.42	0.78	0.56	0.26	0.36	0.74	0.42	0.36	0.39	0.71	0.37	0.33	0.35
	500	0.78	0.54	0.34	0.42	0.76	0.48	0.26	0.34	0.79	0.60	0.31	0.41	0.78	0.63	0.23	0.34	0.75	0.43	0.31	0.36	0.73	0.38	0.32	0.35
	50	0.79	**0.62**	0.13	0.21	0.78	0.20	0.01	0.02	0.78	0.61	0.14	0.23	0.78	0.10	0.00	0.00	0.75	0.44	**0.22**	**0.29**	0.75	0.41	0.14	0.21
SMOT	5793	0.91	0.85	0.99	0.91	0.90	0.84	0.98	0.90	0.89	0.85	0.96	0.90	0.89	0.85	0.96	0.90	0.91	**0.86**	**0.99**	**0.92**	0.90	0.85	0.98	0.91
	5000	0.87	0.80	0.98	0.88	0.84	0.78	0.95	0.86	0.86	0.81	0.93	0.87	0.83	0.79	0.91	0.85	0.87	0.81	0.99	0.89	0.84	0.77	0.97	0.86
	4000	0.88	0.82	0.97	0.89	0.84	0.78	0.96	0.86	0.86	0.82	0.93	0.87	0.81	0.78	0.88	0.83	0.88	0.82	0.99	0.90	0.84	0.77	0.97	0.86
	3000	0.87	0.81	0.98	0.89	0.82	0.75	0.97	0.85	0.84	0.80	0.92	0.86	0.79	0.76	0.87	0.81	0.88	0.82	0.99	0.90	0.85	0.78	0.97	0.86
	2000	0.84	0.80	0.93	0.86	0.79	0.73	0.90	0.81	0.82	0.79	0.87	0.83	0.76	0.74	0.82	0.78	0.88	0.82	0.98	0.89	0.84	0.79	0.95	0.86
	1000	0.77	0.74	0.84	0.79	0.73	0.71	0.78	0.74	0.75	0.74	0.79	0.76	0.71	0.71	0.73	0.72	0.89	0.83	0.98	0.90	0.86	0.80	0.96	0.87
	500	0.76	0.74	0.82	0.78	0.73	0.71	0.77	0.74	0.73	0.72	0.77	0.74	0.70	0.70	0.71	0.70	0.89	0.84	0.98	0.90	0.87	0.82	0.96	0.88
	50	0.66	0.68	0.62	0.65	0.59	0.60	0.54	0.57	0.65	0.67	0.61	0.64	0.58	0.60	0.49	0.54	0.87	**0.82**	**0.96**	**0.88**	0.84	0.80	0.90	0.85

4.3.8 การทดลองเปรียบเทียบ term+dep+all-gram และ all-gram

ตารางที่ 4.11 แสดงถึงการทดลองโดยใช้คุณลักษณะจากคลังคำศัพท์ศาลฎีกา (term) คุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ (dep) และ all-gram (allgram) ประกอบด้วย n-grams ทั้งหมด คือ 1-gram 2-gram 3-gram ระดับคำ และ 2-gram 3-gram 4-gram ระดับตัวอักษร เปรียบเทียบกับการใช้คุณลักษณะ all-gram (allgram) อย่างเดียว ในการจำแนกข้อความเข้าข่ายหมิ่นประมาทด้วยการเรียนรู้ของเครื่องทั้ง 3 ชนิด

จากการทดลองจะเห็นได้ว่าคุณลักษณะจากคลังคำศัพท์ศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ ทำให้ค่าเรียกคืน และค่าความเที่ยงเพิ่มขึ้นเมื่อเทียบกับการใช้ all-gram เพียงอย่างเดียว นอกจากนี้เมื่อใช้ SMOT เมื่อจำนวนของคุณลักษณะ all-gram ที่จำนวนเท่ากับ

43,000 ทำให้ SVM สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.43 ที่จำนวน 47,000 46,000 44,000 42,000 และ 41,000 Logistic Regression สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.43 และที่จำนวน 13,000 Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดด้วย F_1 เท่ากับ 0.35

การทดลองโดยไม่ใช้ SMOT Logistic Regression มีความสามารถในการจำแนกดีที่สุดเมื่อไม่มีการลดจำนวนคุณลักษณะ all-gram ด้วย F_1 เท่ากับ 0.42 ในขณะที่ SVM มีค่าเรียกคืนมากที่สุด แต่เมื่อจำนวนคุณลักษณะ all-gram มีจำนวนลดลง Multi-layer Perceptron มีความสามารถในการจำแนกดีที่สุดด้วย F_1 เท่ากับ 0.27 ในขณะที่ SVM มีค่าความเที่ยงมากที่สุด

การทดลองโดยใช้ SMOT Multi-layer Perceptron สามารถจำแนกข้อความได้ดีที่สุดเมื่อขนาดของคุณลักษณะ all-gram ไม่ลดลง และเมื่อจำนวนคุณลักษณะน้อยที่สุดด้วย F_1 เท่ากับ 0.95 และ 0.88 ตามลำดับ

ตารางที่ 4.11 การทดลองเปรียบเทียบ term+dep+all-gram และ all-gram

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+allgram				allgram				term+dep+allgram				allgram				term+dep+allgram				allgram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
No SMOT	47517	0.77	0.52	**0.33**	0.40	0.76	0.50	0.31	0.38	0.79	**0.63**	0.31	**0.42**	0.79	0.63	0.29	0.40	0.75	0.41	0.14	0.21	0.75	0.39	0.14	0.21
	47000	0.77	0.53	0.33	0.41	0.76	0.49	0.30	0.37	0.79	0.64	0.32	0.43	0.79	0.64	0.29	0.40	0.76	0.41	0.16	0.23	0.75	0.40	0.14	0.21
	46000	0.77	0.53	0.34	0.41	0.76	0.49	0.30	0.37	0.79	0.64	0.32	0.43	0.79	0.64	0.29	0.40	0.77	0.46	0.16	0.24	0.75	0.37	0.13	0.19
	45000	0.77	0.52	0.33	0.40	0.76	0.50	0.30	0.38	0.79	0.65	0.31	0.42	0.79	0.64	0.29	0.40	0.77	0.48	0.17	0.25	0.75	0.38	0.13	0.19
	44000	0.77	0.52	0.33	0.40	0.77	0.52	0.32	0.40	0.79	0.64	0.32	0.43	0.79	0.64	0.30	0.41	0.77	0.50	0.17	0.25	0.76	0.42	0.15	0.22
	43000	0.77	0.55	0.35	0.43	0.76	0.52	0.32	0.40	0.79	0.64	0.31	0.42	0.79	0.64	0.29	0.40	0.77	0.48	0.17	0.25	0.75	0.38	0.14	0.20
	42000	0.77	0.53	0.34	0.41	0.76	0.51	0.30	0.38	0.80	0.65	0.32	0.43	0.79	0.63	0.29	0.40	0.77	0.49	0.17	0.25	0.75	0.37	0.12	0.18
	41000	0.77	0.53	0.33	0.41	0.76	0.53	0.31	0.39	0.80	0.66	0.32	0.43	0.79	0.64	0.29	0.40	0.77	0.51	0.17	0.26	0.75	0.34	0.10	0.15
	40000	0.78	0.58	0.31	0.40	0.78	0.58	0.28	0.38	0.80	0.66	0.27	0.38	0.80	0.67	0.24	0.35	0.77	0.50	0.13	0.21	0.77	0.46	0.12	0.19
	39000	0.78	0.58	0.31	0.40	0.78	0.59	0.28	0.38	0.80	0.66	0.27	0.38	0.79	0.67	0.24	0.35	0.77	0.47	0.13	0.20	0.77	0.44	0.11	0.18
	38000	0.78	0.58	0.33	0.42	0.78	0.59	0.28	0.38	0.80	0.66	0.27	0.38	0.80	0.68	0.24	0.35	0.78	0.51	0.15	0.23	0.76	0.43	0.10	0.16
	37000	0.78	0.57	0.32	0.41	0.79	0.61	0.29	0.39	0.80	0.67	0.29	0.40	0.80	0.69	0.24	0.36	0.77	0.51	0.15	0.23	0.76	0.39	0.09	0.15
	36000	0.78	0.56	0.31	0.40	0.79	0.59	0.27	0.37	0.80	0.66	0.28	0.39	0.80	0.71	0.24	0.36	0.77	0.50	0.16	0.24	0.77	0.43	0.12	0.19
	35000	0.79	0.60	0.22	0.32	0.79	0.60	0.18	0.28	0.80	0.69	0.21	0.32	0.79	0.64	0.12	0.20	0.78	0.61	0.11	0.19	0.78	0.54	0.08	0.14
	34000	0.79	0.59	0.23	0.33	0.79	0.56	0.20	0.29	0.80	0.69	0.21	0.32	0.79	0.64	0.12	0.20	0.78	0.59	0.13	0.21	0.78	0.54	0.06	0.11
	33000	0.79	0.57	0.23	0.33	0.79	0.60	0.21	0.31	0.80	0.69	0.20	0.31	0.79	0.63	0.12	0.20	0.79	0.67	0.13	0.22	0.78	0.44	0.04	0.07
	32000	0.79	0.59	0.14	0.23	0.78	0.57	0.10	0.17	0.79	0.68	0.12	0.20	0.78	0.34	0.03	0.06	0.78	0.54	0.08	0.14	0.78	0.37	0.02	0.04
	31000	0.79	0.62	0.12	0.20	0.78	0.47	0.04	0.07	0.80	0.69	0.10	0.17	0.78	0.40	0.01	0.02	0.77	0.49	0.14	0.22	0.78	0.41	0.03	0.06

ตารางที่ 4.11 การทดลองเปรียบเทียบ term+dep+all-gram และ all-gram (ต่อ)

	n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
		term+dep+allgram				allgram				term+dep+allgram				allgram				term+dep+allgram				allgram			
		ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
	30000	0.79	0.66	0.14	0.23	0.78	0.53	0.04	0.07	0.80	0.77	0.11	0.19	0.78	0.50	0.02	0.04	0.78	0.51	0.14	0.22	0.78	0.52	0.04	0.07
	29000	0.79	0.60	0.18	0.28	0.78	0.59	0.06	0.11	0.80	0.77	0.12	0.21	0.79	0.50	0.02	0.04	0.76	0.38	0.19	0.25	0.78	0.49	0.05	0.09
	28000	0.79	0.59	0.16	0.25	0.78	0.61	0.05	0.09	0.80	0.77	0.12	0.21	0.79	0.50	0.02	0.04	0.77	0.46	0.14	0.21	0.78	0.52	0.05	0.09
	27000	0.79	0.60	0.13	0.21	0.78	0.60	0.05	0.09	0.80	0.77	0.11	0.19	0.79	0.50	0.02	0.04	0.77	0.51	0.16	0.24	0.78	0.53	0.04	0.07
	26000	0.79	0.61	0.13	0.21	0.79	0.63	0.05	0.09	0.80	0.77	0.11	0.19	0.79	0.50	0.02	0.04	0.78	0.53	0.13	0.21	0.78	0.40	0.03	0.06
	25000	0.79	0.61	0.13	0.21	0.78	0.61	0.05	0.09	0.80	0.77	0.11	0.19	0.79	0.50	0.02	0.04	0.78	0.56	0.12	0.20	0.78	0.46	0.03	0.06
	24000	0.79	0.63	0.12	0.20	0.78	0.62	0.06	0.11	0.80	0.68	0.11	0.19	0.78	0.45	0.02	0.04	0.77	0.48	0.16	0.24	0.78	0.36	0.03	0.06
	23000	0.79	0.59	0.11	0.19	0.78	0.52	0.05	0.09	0.79	0.65	0.11	0.19	0.79	0.50	0.02	0.04	0.76	0.47	0.17	0.25	0.78	0.47	0.03	0.06
	22000	0.79	0.64	0.13	0.22	0.78	0.61	0.06	0.11	0.80	0.65	0.12	0.20	0.78	0.45	0.02	0.04	0.76	0.44	0.20	0.28	0.78	0.35	0.03	0.06
	21000	0.79	0.62	0.13	0.21	0.78	0.52	0.05	0.09	0.79	0.60	0.11	0.19	0.78	0.43	0.02	0.04	0.77	0.46	0.21	0.29	0.78	0.27	0.03	0.05
	20000	0.79	0.58	0.14	0.23	0.78	0.54	0.05	0.09	0.80	0.72	0.13	0.22	0.78	0.35	0.01	0.02	0.76	0.45	0.22	0.30	0.78	0.36	0.03	0.06
	19000	0.78	0.53	0.14	0.22	0.78	0.46	0.05	0.09	0.79	0.66	0.12	0.20	0.78	0.27	0.01	0.02	0.75	0.41	0.23	0.29	0.78	0.30	0.03	0.05
	18000	0.79	0.67	0.15	0.25	0.78	0.00	0.00	0.00	0.79	0.70	0.10	0.18	0.78	0.00	0.00	0.00	0.72	0.40	0.28	0.33	0.78	0.30	0.01	0.02
	17000	0.78	0.51	0.11	0.18	0.78	0.35	0.02	0.04	0.79	0.72	0.09	0.16	0.78	0.00	0.00	0.00	0.73	0.33	0.24	0.28	0.78	0.35	0.03	0.06
	16000	0.79	0.58	0.16	0.25	0.78	0.39	0.04	0.07	0.79	0.65	0.09	0.16	0.78	0.00	0.00	0.00	0.70	0.33	0.32	0.32	0.77	0.36	0.06	0.10
	15000	0.77	0.49	0.25	0.33	0.77	0.31	0.05	0.09	0.78	0.57	0.14	0.22	0.78	0.10	0.00	0.00	0.73	0.35	0.27	0.30	0.76	0.37	0.09	0.14
	14000	0.77	0.50	0.31	0.38	0.76	0.33	0.08	0.13	0.79	0.62	0.20	0.30	0.78	0.10	0.01	0.02	0.72	0.36	0.33	0.34	0.75	0.33	0.11	0.17
	13000	0.77	0.48	0.26	0.34	0.76	0.37	0.07	0.12	0.79	0.62	0.19	0.29	0.78	0.20	0.01	0.02	0.74	0.41	0.31	0.35	0.75	0.35	0.11	0.17
	12000	0.77	0.45	0.17	0.25	0.76	0.38	0.07	0.12	0.79	0.62	0.15	0.24	0.78	0.20	0.01	0.02	0.75	0.40	0.24	0.30	0.76	0.39	0.10	0.16
	11000	0.77	0.43	0.13	0.20	0.76	0.37	0.06	0.10	0.78	0.55	0.10	0.17	0.78	0.20	0.01	0.02	0.76	0.42	0.16	0.23	0.76	0.41	0.10	0.16
	10000	0.77	0.48	0.10	0.17	0.77	0.40	0.06	0.10	0.79	0.60	0.10	0.17	0.78	0.20	0.01	0.02	0.76	0.42	0.15	0.22	0.76	0.37	0.07	0.12
	9000	0.77	0.51	0.10	0.17	0.77	0.44	0.06	0.11	0.79	0.61	0.10	0.17	0.78	0.10	0.00	0.00	0.77	0.45	0.14	0.21	0.77	0.39	0.08	0.13
	8000	0.77	0.46	0.11	0.18	0.77	0.46	0.06	0.11	0.78	0.55	0.10	0.17	0.78	0.10	0.01	0.02	0.77	0.47	0.15	0.23	0.76	0.41	0.08	0.13
	7000	0.78	0.53	0.12	0.20	0.78	0.51	0.08	0.14	0.79	0.58	0.12	0.20	0.78	0.10	0.01	0.02	0.78	0.52	0.15	0.23	0.77	0.46	0.09	0.15
	6000	0.78	0.56	0.13	0.21	0.77	0.47	0.07	0.12	0.79	0.58	0.11	0.18	0.78	0.20	0.01	0.02	0.77	0.50	0.17	0.25	0.77	0.38	0.09	0.15
	5000	0.78	0.60	0.14	0.23	0.77	0.49	0.09	0.15	0.79	0.63	0.14	0.23	0.78	0.35	0.02	0.04	0.77	0.52	0.21	0.30	0.76	0.40	0.10	0.16
	4000	0.78	0.55	0.16	0.25	0.77	0.37	0.07	0.12	0.79	0.65	0.15	0.24	0.78	0.30	0.01	0.02	0.78	0.58	0.24	0.34	0.77	0.40	0.10	0.16
	3000	0.78	0.62	0.15	0.24	0.77	0.27	0.06	0.10	0.79	0.65	0.14	0.23	0.78	0.35	0.01	0.02	0.77	0.56	0.20	0.29	0.77	0.34	0.09	0.14
	2000	0.78	0.66	0.14	0.23	0.77	0.30	0.06	0.10	0.79	0.63	0.13	0.22	0.78	0.22	0.01	0.02	0.78	0.57	0.23	0.33	0.77	0.39	0.07	0.12
	1000	0.78	0.53	0.12	0.20	0.77	0.33	0.03	0.06	0.78	0.54	0.15	0.23	0.78	0.13	0.01	0.02	0.74	0.43	0.31	0.36	0.68	0.24	0.20	0.22
	500	0.78	0.53	0.13	0.21	0.78	0.10	0.00	0.00	0.78	0.54	0.16	0.25	0.78	0.00	0.00	0.00	0.73	0.36	0.26	0.30	0.72	0.34	0.22	0.27
	50	0.79	**0.60**	0.12	0.20	0.78	0.00	0.00	0.00	0.78	0.56	0.16	0.25	0.78	0.00	0.00	0.00	0.79	0.57	**0.18**	**0.27**	0.78	0.00	0.00	0.00
SMOT	47517	0.93	0.89	0.99	0.94	0.93	0.89	0.99	0.94	0.92	0.89	0.97	0.93	0.92	0.89	0.97	0.93	0.95	**0.93**	**0.98**	**0.95**	0.95	0.93	0.98	0.95
	47000	0.93	0.89	0.99	0.94	0.93	0.89	0.99	0.94	0.92	0.89	0.97	0.93	0.92	0.89	0.97	0.93	0.95	0.92	0.98	0.95	0.95	0.93	0.98	0.95
	46000	0.93	0.89	0.99	0.94	0.93	0.89	0.99	0.94	0.92	0.89	0.97	0.93	0.92	0.89	0.97	0.93	0.95	0.93	0.98	0.95	0.95	0.93	0.98	0.95
	45000	0.93	0.89	0.99	0.94	0.93	0.89	0.99	0.94	0.92	0.89	0.97	0.93	0.92	0.89	0.97	0.93	0.95	0.93	0.98	0.95	0.95	0.93	0.98	0.95
	44000	0.93	0.88	0.99	0.93	0.93	0.89	0.99	0.94	0.92	0.89	0.97	0.93	0.92	0.89	0.97	0.93	0.95	0.93	0.98	0.95	0.95	0.93	0.98	0.95

ตารางที่ 4.11 การทดลองเปรียบเทียบ term+dep+all-gram และ all-gram (ต่อ)

n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
	term+dep+allgram				allgram				term+dep+allgram				allgram				term+dep+allgram				allgram			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
43000	0.93	0.89	0.99	0.94	0.93	0.89	0.99	0.94	0.92	0.89	0.98	0.93	0.92	0.89	0.97	0.93	0.95	0.93	0.98	0.95	0.95	0.92	0.98	0.95
42000	0.92	0.88	0.99	0.93	0.93	0.89	0.99	0.94	0.93	0.89	0.98	0.93	0.92	0.89	0.97	0.93	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
41000	0.93	0.89	0.99	0.94	0.93	0.89	0.98	0.93	0.92	0.89	0.98	0.93	0.92	0.90	0.97	0.93	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
40000	0.94	0.90	0.99	0.94	0.94	0.91	0.99	0.95	0.93	0.90	0.97	0.93	0.93	0.91	0.97	0.94	0.94	0.92	0.98	0.95	0.95	0.92	0.98	0.95
39000	0.94	0.90	0.99	0.94	0.94	0.91	0.99	0.95	0.93	0.90	0.97	0.93	0.93	0.91	0.97	0.94	0.94	0.92	0.98	0.95	0.95	0.92	0.98	0.95
38000	0.94	0.90	0.99	0.94	0.94	0.91	0.99	0.95	0.93	0.89	0.97	0.93	0.93	0.91	0.97	0.94	0.95	0.92	0.98	0.95	0.95	0.93	0.98	0.95
37000	0.93	0.90	0.99	0.94	0.93	0.89	0.99	0.94	0.93	0.90	0.97	0.93	0.93	0.91	0.97	0.94	0.95	0.92	0.98	0.95	0.95	0.92	0.98	0.95
36000	0.94	0.91	0.99	0.95	0.93	0.89	0.99	0.94	0.93	0.90	0.97	0.93	0.93	0.91	0.97	0.94	0.95	0.92	0.98	0.95	0.94	0.92	0.98	0.95
35000	0.87	0.80	0.98	0.88	0.93	0.89	0.99	0.94	0.92	0.87	0.98	0.92	0.92	0.91	0.95	0.93	0.95	0.92	0.98	0.95	0.94	0.92	0.98	0.95
34000	0.86	0.80	0.98	0.88	0.93	0.89	0.99	0.94	0.90	0.85	0.98	0.91	0.90	0.85	0.97	0.91	0.95	0.92	0.98	0.95	0.94	0.91	0.98	0.94
33000	0.86	0.79	0.99	0.88	0.93	0.89	0.99	0.94	0.89	0.83	0.98	0.90	0.88	0.82	0.97	0.89	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
32000	0.86	0.80	0.98	0.88	0.93	0.89	0.99	0.94	0.88	0.83	0.97	0.89	0.87	0.81	0.97	0.88	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
31000	0.84	0.77	0.98	0.86	0.92	0.88	0.99	0.93	0.86	0.80	0.98	0.88	0.76	0.68	0.98	0.80	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
30000	0.84	0.77	0.98	0.86	0.92	0.88	0.99	0.93	0.86	0.80	0.98	0.88	0.76	0.68	0.98	0.80	0.94	0.91	0.98	0.94	0.94	0.91	0.98	0.94
29000	0.84	0.76	0.98	0.86	0.92	0.87	0.98	0.92	0.86	0.79	0.98	0.87	0.76	0.68	0.98	0.80	0.93	0.89	0.98	0.93	0.93	0.89	0.98	0.93
28000	0.84	0.77	0.98	0.86	0.91	0.87	0.99	0.93	0.86	0.80	0.98	0.88	0.77	0.69	0.98	0.81	0.93	0.89	0.98	0.93	0.93	0.88	0.98	0.93
27000	0.85	0.78	0.98	0.87	0.91	0.86	0.99	0.92	0.87	0.80	0.98	0.88	0.78	0.71	0.98	0.82	0.92	0.87	0.98	0.92	0.92	0.87	0.98	0.92
26000	0.91	0.86	0.99	0.92	0.70	0.63	0.99	0.77	0.88	0.81	0.98	0.89	0.80	0.72	0.98	0.83	0.91	0.87	0.98	0.92	0.91	0.87	0.98	0.92
25000	0.91	0.87	0.99	0.93	0.70	0.63	0.99	0.77	0.88	0.81	0.98	0.89	0.80	0.73	0.98	0.84	0.92	0.88	0.98	0.93	0.92	0.87	0.98	0.92
24000	0.91	0.87	0.99	0.93	0.70	0.63	0.99	0.77	0.89	0.83	0.98	0.90	0.80	0.73	0.98	0.84	0.92	0.88	0.98	0.93	0.92	0.88	0.98	0.93
23000	0.91	0.87	0.99	0.93	0.69	0.62	0.99	0.76	0.89	0.83	0.98	0.90	0.79	0.71	0.98	0.82	0.93	0.89	0.98	0.93	0.93	0.89	0.98	0.93
22000	0.92	0.87	0.99	0.93	0.69	0.62	0.99	0.76	0.88	0.83	0.98	0.90	0.79	0.71	0.98	0.82	0.93	0.90	0.98	0.94	0.93	0.90	0.98	0.94
21000	0.92	0.88	0.99	0.93	0.68	0.62	0.99	0.76	0.88	0.82	0.97	0.89	0.78	0.71	0.98	0.82	0.93	0.89	0.98	0.93	0.93	0.89	0.98	0.93
20000	0.92	0.87	0.99	0.93	0.67	0.61	0.99	0.75	0.88	0.82	0.98	0.89	0.77	0.70	0.98	0.82	0.93	0.89	0.98	0.93	0.93	0.89	0.98	0.93
19000	0.86	0.79	0.98	0.87	0.65	0.59	0.99	0.74	0.87	0.81	0.98	0.89	0.75	0.68	0.98	0.80	0.93	0.89	0.98	0.93	0.92	0.88	0.98	0.93
18000	0.83	0.76	0.99	0.86	0.57	0.54	1.00	0.70	0.87	0.81	0.97	0.88	0.73	0.67	0.97	0.79	0.92	0.88	0.98	0.93	0.92	0.87	0.98	0.92
17000	0.74	0.67	0.98	0.80	0.56	0.53	1.00	0.69	0.81	0.75	0.94	0.83	0.86	0.84	0.93	0.88	0.92	0.88	0.98	0.93	0.91	0.87	0.98	0.92
16000	0.71	0.64	0.98	0.77	0.57	0.54	1.00	0.70	0.81	0.76	0.91	0.83	0.72	0.69	0.93	0.79	0.92	0.87	0.98	0.92	0.92	0.87	0.98	0.92
15000	0.71	0.64	0.98	0.77	0.57	0.54	1.00	0.70	0.80	0.76	0.89	0.82	0.62	0.57	0.99	0.72	0.92	0.87	0.98	0.92	0.91	0.86	0.98	0.92
14000	0.68	0.62	0.98	0.76	0.57	0.54	1.00	0.70	0.77	0.74	0.87	0.80	0.61	0.56	0.99	0.72	0.91	0.86	0.98	0.92	0.90	0.85	0.98	0.91
13000	0.74	0.69	0.94	0.80	0.58	0.54	0.99	0.70	0.78	0.74	0.87	0.80	0.62	0.58	0.95	0.72	0.91	0.86	0.98	0.92	0.89	0.84	0.98	0.90
12000	0.78	0.73	0.89	0.80	0.63	0.60	0.94	0.73	0.78	0.76	0.84	0.80	0.80	0.83	0.79	0.81	0.91	0.86	0.98	0.92	0.89	0.84	0.98	0.90
11000	0.81	0.79	0.86	0.82	0.82	0.88	0.75	0.81	0.79	0.78	0.82	0.80	0.81	0.89	0.72	0.80	0.91	0.86	0.98	0.92	0.89	0.84	0.98	0.90
10000	0.81	0.81	0.83	0.82	0.81	0.89	0.71	0.79	0.80	0.80	0.80	0.80	0.81	0.91	0.68	0.78	0.91	0.86	0.98	0.92	0.89	0.84	0.98	0.90
9000	0.81	0.82	0.82	0.82	0.81	0.90	0.69	0.78	0.80	0.81	0.79	0.80	0.80	0.92	0.67	0.78	0.91	0.86	0.98	0.92	0.90	0.84	0.98	0.90
8000	0.81	0.82	0.81	0.81	0.80	0.90	0.68	0.77	0.80	0.81	0.78	0.79	0.79	0.92	0.64	0.75	0.91	0.86	0.98	0.92	0.89	0.84	0.98	0.90
7000	0.80	0.81	0.79	0.80	0.78	0.90	0.62	0.73	0.79	0.80	0.77	0.78	0.77	0.92	0.59	0.72	0.91	0.85	0.99	0.91	0.89	0.84	0.98	0.90

ตารางที่ 4.11 การทดลองเปรียบเทียบ term+dep+all-gram และ all-gram (ต่อ)

n-gram feature dimension	SVM								Logistic Regression								Multi-layer Perceptron							
	term+dep+allgram				allgram				term+dep+allgram				allgram				term+dep+allgram				allgram			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
6000	0.77	0.80	0.73	0.76	0.74	0.90	0.55	0.68	0.76	0.79	0.71	0.75	0.73	0.92	0.52	0.66	0.91	0.86	0.98	0.92	0.90	0.84	0.98	0.90
5000	0.76	0.79	0.72	0.75	0.72	0.89	0.50	0.64	0.74	0.76	0.70	0.73	0.71	0.90	0.48	0.63	0.90	0.86	0.98	0.92	0.90	0.85	0.98	0.91
4000	0.75	0.79	0.70	0.74	0.69	0.88	0.44	0.59	0.72	0.74	0.69	0.71	0.68	0.88	0.43	0.58	0.91	0.86	0.98	0.92	0.90	0.85	0.98	0.91
3000	0.72	0.75	0.67	0.71	0.66	0.87	0.38	0.53	0.70	0.71	0.67	0.69	0.65	0.87	0.37	0.52	0.90	0.85	0.99	0.91	0.90	0.85	0.98	0.91
2000	0.69	0.71	0.66	0.68	0.63	0.86	0.30	0.44	0.68	0.69	0.65	0.67	0.62	0.86	0.29	0.43	0.90	0.85	0.99	0.91	0.89	0.83	0.98	0.90
1000	0.67	0.69	0.65	0.67	0.53	0.54	0.41	0.47	0.66	0.67	0.64	0.65	0.51	0.52	0.32	0.40	0.91	0.86	0.98	0.92	0.90	0.85	0.98	0.91
500	0.66	0.67	0.63	0.65	0.59	0.58	0.65	0.61	0.64	0.65	0.62	0.63	0.57	0.57	0.63	0.60	0.90	0.85	0.98	0.91	0.91	0.86	0.98	0.92
50	0.64	0.65	0.61	0.63	0.49	0.49	0.54	0.51	0.64	0.65	0.62	0.63	0.49	0.49	0.53	0.51	0.87	**0.82**	**0.96**	**0.88**	0.88	0.83	0.95	0.89

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

วิทยานิพนธ์นี้ได้พัฒนาวิธีการสำหรับการจำแนกประเภทข้อความที่เข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์ ทดสอบประสิทธิภาพ และเปรียบเทียบวิธีการสำหรับการจำแนกประเภทข้อความเข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์ จากผลการทดลองที่แสดงไว้ในบทที่ 4 สามารถสรุปผลการวิจัยและข้อเสนอแนะ ดังนี้

5.1 สรุปผลการวิจัย

สำหรับวิธีการจำแนกประเภทข้อความที่เข้าข่ายหมิ่นประมาทบนสื่อสังคมออนไลน์นั้น พบว่าการทดลองโดยใช้คุณลักษณะคลังคำศัพท์จากศาลฎีกา ที่อ้างอิงจากองค์ประกอบของหลักประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาท ร่วมกับคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์มีประสิทธิภาพด้อยกว่า เมื่อเทียบกับการใช้คุณลักษณะ n-grams ระดับคำ และตัวอักษร แต่เมื่อใช้คุณลักษณะทั้งสองร่วมกับ n-grams ทำให้การจำแนกข้อความมีประสิทธิภาพมากขึ้น เทียบกับการใช้เพียงคุณลักษณะ n-grams ซึ่งการจำแนกข้อความด้วยคุณลักษณะคลังคำศัพท์จากศาลฎีกา และคุณลักษณะโครงสร้างต้นไม้ไวยากรณ์ Multi-layer Perceptron มีความสามารถในการจำแนกข้อความได้เข้าข่ายหมิ่นประมาทได้ดีที่สุด

จำนวนของคุณลักษณะ n-grams มีผลต่อประสิทธิภาพของการจำแนกข้อความ เมื่อจำนวนของคุณลักษณะ n-grams โดย n-gram แต่ละชนิดสำหรับแต่ละการเรียนรู้ของเครื่องจะมีจำนวนที่เหมาะสมสำหรับการจำแนกข้อความเข้าข่ายหมิ่นประมาทแตกต่างกัน ในขณะที่ชุดข้อความที่ใช้ขั้นตอนวิธี SMOT ค่าความเที่ยงและค่าเรียกคืนจะลดลงเมื่อจำนวนของคุณลักษณะ n-grams ลดลง

การทดลองการเพิ่มจำนวนข้อความเข้าข่ายหมิ่นประมาทด้วยขั้นตอนวิธี SMOT พบว่าจำนวนของข้อมูลมีผลอย่างมากกับการจำแนกข้อความ ซึ่งแตกต่างกันอย่างเห็นได้ชัดเจนที่ค่าเรียกคืนจากการทดลองกับชุดข้อความที่ไม่ใช้ขั้นตอนวิธี SMOT สังเกตได้ว่าข้อความเข้าข่ายหมิ่นประมาทขึ้นอยู่กับคำๆเดียวภายในประโยค เนื่องจากผลลัพธ์ของการจำแนกข้อความโดยใช้ 1-grams นั้นมีประสิทธิภาพมากกว่า 2-gram และ 3-gram จากค่า F_1 ในขณะที่ความยาวของตัวอักษรไม่มีผลต่อการจำแนกประเภทข้อความมากนัก

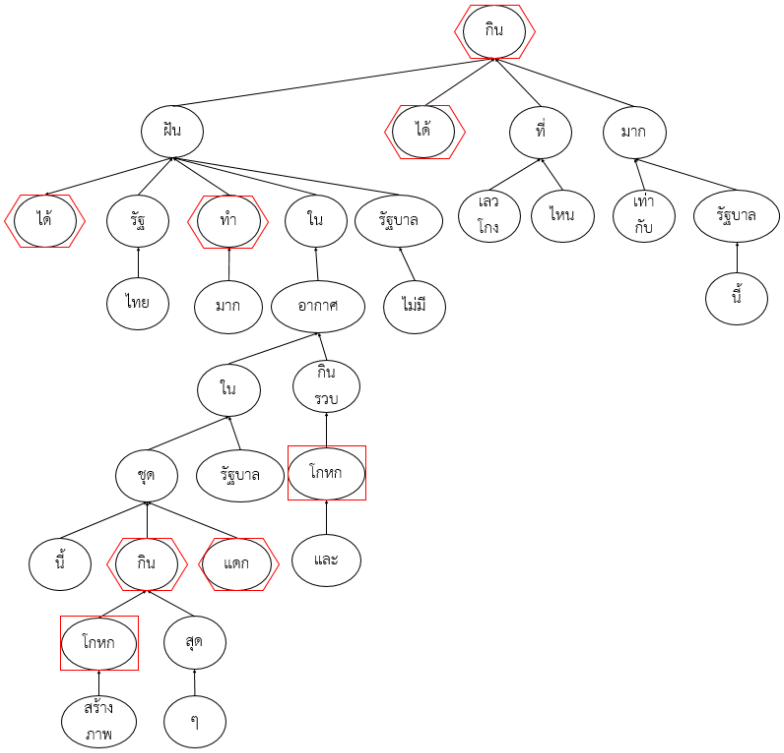
การเรียนรู้ของเครื่องทั้ง 3 ชนิด โดยภาพรวมแล้ว Logistic Regression มีความเที่ยงมากที่สุด ในขณะที่ จำนวนคุณลักษณะ n-grams มีจำนวนน้อยลง Multi-layer Perceptron มี

ประสิทธิภาพดีที่สุดสำหรับข้อมูลที่ไม่ใช้ขั้นตอนวิธี SMOT แต่สำหรับข้อมูลที่ใช้ขั้นตอนวิธี SMOT การเรียนรู้ของเครื่องทั้ง 3 ชนิด ไม่มีความแตกต่างกันเท่าไหร่นัก

5.2 อภิปรายผลการทดลอง

การตัดคำ คือ ขั้นตอนสำคัญก่อนการจำแนกข้อความ ซึ่งการจำแนกข้อความโดยใช้คุณลักษณะคลังคำศัพท์จากศาลฎีกา และโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันนั้น ใช้คำในการตรวจหาคุณลักษณะจากประโยค หากการตัดคำเกิดข้อผิดพลาด จะทำให้การสร้างคุณลักษณะเกิดข้อผิดพลาด ส่งผลต่อประสิทธิภาพในการจำแนกข้อความด้วยเช่นกัน ดังตัวอย่างประโยค “รัฐ|ไทย|ไม่มี|รัฐบาล|ไหน|ที่|**เลวโกง**|กิน|ได้|มาก|เท่ากับ|รัฐบาล|นี้|กินรวบ|โกหก|และ|ทำ|ฝัน|ใน|อากาศ|ได้|มากใน|รัฐบาล|ชุด|นี้|แดก|กิน|โกหก|สร้างภาพ|สุด|ๆ” จากตัวอย่างประโยค การตัดคำในประโยคแทนด้วยตัวอักษร “|” ซึ่งในคำที่ 7 คำว่า “เลวโกง” เกิดความผิดพลาดในการตัดคำ แทนที่จะถูกเป็น 2 คำ คือ “เลว” และ “โกง” กลับตัดได้คำเดียวคือ “เลวโกง” ทำให้การสร้างคุณลักษณะพลาดคำว่า “โกง” ไป ส่งผลให้คุณลักษณะที่ใช้ในการจำแนกข้อความมีความผิดพลาด

การสร้างโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันนั้น เป็นการสร้างโครงสร้างต้นไม้จากค่าความน่าจะเป็น ซึ่งหาได้จากขั้นตอนวิธี SVM จากตัวอย่างข้อความ “รัฐ|ไทย|ไม่มี|รัฐบาล|ไหน|ที่|เลวโกง|กิน|ได้|มาก|เท่ากับ|รัฐบาล|นี้|กินรวบ|โกหก|และ|ทำ|ฝัน|ใน|อากาศ|ได้|มาก|ใน|รัฐบาล|ชุด|นี้|แดก|กิน|โกหก|สร้างภาพ|สุด|ๆ” หากพิจารณาแล้ว ผู้ถูกกระทำจากประโยคนี้คือ “รัฐบาล” ในขณะที่โครงสร้างต้นไม้ย่อยที่พิจารณาจากคำกริยาแสดงดังรูปที่ 5.1



รูปที่ 5.1 โครงสร้างต้นไม้ประโยค
“รัฐ|ไทย|ไม่มี|รัฐบาล|ไหน|ที่|เลวโกง|กิน|ได้|มาก|เท่ากับ|รัฐบาล|นี้|กินรวบ|โกหก|และ|ทำ|ฝัน|ใน|อากาศ|ได้|มาก|ใน|รัฐบาล|ชุด|นี้|แดก|กิน|โกหก|สร้างภาพ|สุด|ๆ”

จากรูปที่ 5.1 ประโยคเป็นประโยคเข้าข่ายหมิ่นประมาทแต่ถูกจำแนกเป็นข้อความไม่เข้าข่ายหมิ่นประมาท ผู้ถูกกระทำ “รัฐบาล” ในประโยคนี้ทำหน้าที่เป็นประธาน สี่เหลี่ยมและหกเหลี่ยมสีแดงในรูปแสดงถึงจุดเริ่มต้นของโครงสร้างย่อยที่จะพิจารณา แต่การตรวจหาชื่อเฉพาะตรวจพบเพียงแค่คำว่า “ไทย” แต่ไม่พบคำว่า “รัฐ” หรือ “รัฐบาล” ทำให้การตรวจหาโครงสร้างต้นไม้ย่อยไม่พบประธานจึงเกิดข้อผิดพลาดขึ้น นอกจากนี้คำที่อยู่ในรูปสี่เหลี่ยมคือคำกริยาที่กรรมไม่มีผล โหนดลำดับถัดมากลับไม่พบประธาน “รัฐบาล” หรือ “รัฐ” เลย แสดงให้เห็นถึงการสร้างโครงสร้างต้นไม้ที่ผิดพลาด คำที่อยู่ในรูปหกเหลี่ยมคือคำกริยาที่กรรมมีผล แต่ไม่พบกรรมอยู่ในโหนดลำดับถัดมาเลย ในขณะที่ผู้ที่อ่านความคิดเห็นเข้าใจความหมายของคำกริยานี้มีความหมายว่ารัฐบาลทุจริต

นอกจากนี้ประโยค “เดี๋ยวๆๆๆๆ ถามจริงๆ มึงได้เท่าไหร่วะคดีนี้” เป็นประโยคเข้าข่ายหมิ่นประมาท แต่คำว่า “ได้” ซึ่งเป็นคำกริยาที่กรรมมีผลโดยอาจจะเป็น “ได้เงิน” “ได้สินบน” แต่กลับไม่มีกรรมปรากฏอยู่ในประโยคเลยนอกจากคำว่า “เท่าไหร่” ทำให้จำแนกข้อความผิดพลาดเป็นข้อความไม่เข้าข่ายหมิ่นประมาท

การใช้ n-grams ทำให้ประสิทธิภาพในการจำแนกข้อความเพิ่มมากขึ้น แต่เมื่อจำนวนมิติของ n-grams ลดลง จะเห็นได้ชัดว่า ค่าความเที่ยงเพิ่มมากขึ้น แต่ค่าเรียกคืนลดลง ทั้งนี้การลดจำนวนมิติด้วย SVD เปรียบเสมือนการตัดคอลัมน์ที่ใช้ในการจำแนกข้อความซึ่งเปรียบได้ว่าเป็นค่ารบกวน (Noise) ออกไป สังเกตได้จาก n-grams แต่ละชนิดจะมีจำนวนมิติที่เหมาะสมสำหรับแต่ละขั้นตอนวิธี ในจำนวนมิติที่เหมาะสม n-grams สามารถทำงานได้มีประสิทธิภาพมากที่สุดอ้างอิงจากค่า F_1 แต่เมื่อจำนวนคอลัมน์ลดลงเกินจำนวนที่เหมาะสม ทำให้ประสิทธิภาพในการจำแนกข้อความลดลงเนื่องจากการหายไปของคอลัมน์ที่สำคัญส่งผลให้ค่าเรียกคืนมีจำนวนลดลง ในขณะที่เมื่อจำนวนมิติของ n-grams ลดลง ค่าความเที่ยงเพิ่มมากขึ้น บ่งบอกถึงจำนวนแกรม grams ที่ใช้ในการจำแนกมีน้อย และประสิทธิภาพในการจำแนกข้อความด้วย grams นั้นๆมีประสิทธิภาพดี เนื่องจากค่าเรียกคืนที่น้อยหมายถึงจำนวนประโยคที่จำแนกว่าเข้าข่ายหมิ่นประมาทมีน้อยไปด้วย รวมทั้งข้อมูลส่วนใหญ่ขาดความหลากหลายในเรื่องคำ เช่น คำว่า “โกง” สามารถพบเห็นได้ในหลายข้อความที่เป็นข้อความเข้าข่ายหมิ่นประมาท “หน้ามันก็บอกยี่ห้อแล้วว่าไอ้ขี้โกง” “มันโกงเข้ามาเป็นผบ.” “รัฐบาลโจรชุดนี้โกงมากที่สุดโกงทุกโครงการ” เป็นต้น ขั้นตอนวิธีจึงอาจจะจำแนกโดยใช้เพียงแค่คำว่า “โกง” เช่น หากข้อความทั้งหมดที่ประกอบด้วยคำว่า “โกง” มีจำนวนเท่ากับ 10 ข้อความ และ 7 ใน 10 เป็นข้อความเข้าข่ายหมิ่นประมาท การจำแนกอาจบอกว่าข้อความทั้งหมดนั้นเป็นข้อความเข้าข่ายหมิ่นประมาท ทำให้ค่าความเที่ยงมีค่าเท่ากับ 0.7

การใช้ SMOT เพื่อแก้ไขปัญหาความไม่สมดุลของข้อความทำให้ประสิทธิภาพเพิ่มขึ้น โดยที่ค่าเรียกคืนมีค่าเท่ากับ 1 ในบางการทดลอง เป็นเพราะว่า SMOT เพิ่มจำนวนข้อความแบบเชิงเส้น ในขณะที่ขั้นตอนวิธีที่ใช้ในการทดลอง SVM, Logistic regression และ Multi-layer perceptron เป็นการจำแนกข้อมูลแบบเชิงเส้น ทำให้ประสิทธิภาพในการจำแนกดีขึ้นมาก ซึ่งหากเปลี่ยนขั้นตอนวิธีในการจำแนก SMOT อาจช่วยให้ประสิทธิภาพในการจำแนกดีขึ้น แต่อาจจะไม่ดีเท่าทั้ง 3 ขั้นตอนวิธีที่กล่าวมา

5.3 ปัญหาและอุปสรรคในการดำเนินงาน

1) ข้อมูลที่ใช้ในการวิจัยนี้เก็บรวบรวมจากเฟซบุ๊กผ่าน GraphAPI การเก็บข้อมูลรอบแรกสามารถเก็บข้อมูลได้โดยไม่มีอุปสรรค แต่การเก็บข้อมูลรอบที่สองเกิดอุปสรรคระหว่างการเก็บข้อมูล เนื่องจากเฟซบุ๊กเพิ่มความระมัดระวังในการเข้าถึงข้อมูลทำให้ผู้วิจัยต้องขออนุญาตเฟซบุ๊กเพื่อเข้าถึงข้อมูล ซึ่งผู้วิจัยไม่ได้รับอนุญาต

2) ระยะเวลาในการตอบแบบสอบถามของผู้เชี่ยวชาญในการพิจารณาข้อความเข้าข่ายหมิ่นประมาทหรือไม่ และ การพิจารณาประเภทข้อความ มีความล่าช้าเนื่องจากจำนวนข้อความมีมาก

5.4 ข้อเสนอแนะ

การตัดคำส่งผลอย่างมากต่อการสร้างคุณลักษณะที่ใช้ในการจำแนกข้อความ นอกจากนี้ คำศัพท์ในภาษาไทยมีจำนวนมากและมีการสร้างคำศัพท์ใหม่ทุกวัน อาจทำให้การตัดคำเกิดข้อผิดพลาดได้ ผู้วิจัยคิดว่าควรตัดคำให้เป็นคำๆเดียวแล้วจึงนำคำภายในประโยคมาประกอบเป็นคำศัพท์ใหม่ด้วยการหาคำที่ยาวที่สุดเมื่อเทียบกับพจนานุกรม เช่น จากประโยค “ต้น|หญ้า|ทำ|ไม|ไม่|ไป|โรง|เรียน” หลังจากนำคำแต่ละคำมารวมกันแล้วจะได้ประโยคที่ตัดคำแล้วดังนี้ “ต้นหญ้า|ทำไม|ไม่|ไป|โรงเรียน”

การสร้างโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันในงานวิจัยนี้ ผู้วิจัยเลือกใช้การสร้างด้วย SVM ซึ่งยังมีการสร้างโครงสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อด้วยการใช้ไวยากรณ์ภาษา (Context free grammar) ซึ่งอาจช่วยให้โครงสร้างต้นไม้ที่ได้มีความเหมาะสมมากขึ้น

การทดลองนี้เจาะจงไปที่เรื่องของการเมืองในปัจจุบัน ซึ่งจากคุณลักษณะที่ใช้ คำศัพท์ที่มีอาจจะไม่ครอบคลุมหากจะนำไปใช้ในประโยคเรื่องอื่นๆ จากคุณลักษณะที่ใช้ในการทดลองในองค์ประกอบของการยืนยันข้อเท็จจริง ยังใช้ผู้เชี่ยวชาญในการประเมินว่าข้อความเป็นประโยคชนิดใด ได้แก่ ประโยคบอกเล่า ประโยคคำถาม ประโยคปฏิเสธ ประโยคขอร้องหรือชักชวน ประโยคคำสั่ง และประโยคอุทาน ซึ่งหากเปลี่ยนจากการให้ผู้เชี่ยวชาญประเมินประโยคเป็นการจำแนกประโยคจะทำให้การจำแนกข้อความสะดวก และรวดเร็วยิ่งขึ้น

นอกจากนี้ในองค์ประกอบการยืนยันข้อเท็จจริง ตัวอย่างเช่น “มีข่าวลือว่านาย ก. หากินกับตำรวจด้วยการเรียกสินบน” รูปแบบประโยคลักษณะนี้ ศาลได้ตัดสินว่าไม่เป็นการหมิ่นประมาททั้งๆที่ประโยคเป็นประโยคบอกเล่า กล่าวคือ รูปแบบประโยคเป็นรูปแบบที่ผู้กระทำได้ฟังข่าวสาร หรือได้ยินจากบุคคลอื่นแล้วมาบอกต่อในลักษณะส่งข่าว ศาลจึงตัดสินว่าไม่เป็นการหมิ่นประมาท เพราะฉะนั้นสิ่งที่ต้องเพิ่มในส่วนองค์ประกอบการยืนยันข้อเท็จจริง คือ ตรวจหารูปแบบประโยคว่าเป็นประโยคที่ผู้กระทำกล่าวขึ้นมาเอง หรือเป็นข่าวลือ เรื่องเล่าที่ผู้กระทำได้ยินมา ซึ่งอาจช่วยให้การจำแนกข้อความมีประสิทธิภาพมากขึ้น

ขั้นตอนวิธีที่ใช้ในการทดลองนี้ได้แก่ SVM, Logistic regression และ Multi-layer perceptron โดยที่ Multi-layer perceptron เป็นเพียงรูปแบบหนึ่งของโครงข่ายประสาทเทียมเช่นเดียวกันกับ Logistic regression ที่เป็นรูปแบบหนึ่งของการจำแนกด้วยการถดถอย โครงข่ายประสาทเทียมยังมีขั้นตอนวิธีรูปแบบอื่นซึ่งอาจจะสามารถจำแนกข้อความได้มีประสิทธิภาพมากกว่า เช่น Long Short-Term Memory (LSTM) นอกจากนี้ SVM และ Multi-layer perceptron มีค่าพารามิเตอร์ซึ่งหากมีการทดสอบปรับเปลี่ยนค่า อาจจะช่วยเพิ่มประสิทธิภาพในการจำแนกข้อความได้

เพราะฉะนั้นแนวทางในการวิจัยต่อไปควรปรับปรุงการตัดคำ การสร้างต้นไม้ไวยากรณ์แบบขึ้นต่อกันให้มีประสิทธิภาพมาขึ้น หากใช้คุณลักษณะชนิดของประโยคควรหาวิธีจำแนกประโยคโดยไม่พึ่งผู้เชี่ยวชาญให้ได้อย่างมีประสิทธิภาพ การแยกประโยคที่ผู้กระทำได้ฟังข่าวสารหรือได้ยินจากบุคคลอื่นแล้วมาบอกต่อในลักษณะส่งข่าว ออกจากประโยคปกติ จะช่วยให้องค์ประกอบการยืนยันข้อเท็จจริงหนักแน่นยิ่งขึ้น นอกจากนี้การปรับเปลี่ยนขั้นตอนวิธีที่ใช้ในการจำแนกข้อความ และการปรับเปลี่ยนค่าพารามิเตอร์ของแต่ละขั้นตอนวิธีอาจช่วยให้การจำแนกข้อความมีประสิทธิภาพมากขึ้น

งานวิจัยนี้เป็นการประยุกต์ใช้ประมวลกฎหมายอาญา หมวด 3 ความผิดฐานหมิ่นประมาทร่วมกับข้อความบนสื่อสังคมออนไลน์ การหมิ่นประมาทไม่ได้จำกัดเพียงแค่การใช้ข้อความเท่านั้น โดยสามารถหมิ่นประมาทได้ด้วยรูปภาพ และอื่นๆ นอกจากนี้การทดลองยังเป็นการทดลองเพื่อจำแนกข้อความ”เข้าข่าย”หมิ่นประมาทในเบื้องต้นโดยพิจารณาจากเพียงความคิดเห็นเท่านั้น การวิเคราะห์ว่าข้อความเป็นข้อความเข้าข่ายหมิ่นประมาทแท้จริงหรือไม่ต้องดูองค์ประกอบหลายอย่าง เช่น เจตนาของผู้กระทำ เป็นต้น

บรรณานุกรม

1. Ikonomakis, M., S. Kotsiantis, and V. Tampakas, *Text classification using machine learning techniques.* WSEAS transactions on computers, 2005. **4**(8): p. 966-974.
2. Aggarwal, C.C. and C. Zhai, *A survey of text classification algorithms*, in *Mining text data*. 2012, Springer. p. 163-222.
3. Tongchim, S., et al. *A Dependency Parser for Thai*. in *Proceedings of 6th international conference on Language Resources and Evaluation: LREC*. 2008. Morroco.
4. Nivre, J., et al. *Universal Dependencies v1: A Multilingual Treebank Collection*. in *10th edition of the Language Resources and Evaluation Conference: LREC*. 2016. Protoroz.
5. Al-Rfou, R., et al. *Polyglot-NER: Massive multilingual named entity recognition*. in *Proceedings of the 2015 SIAM International Conference on Data Mining*. 2015. SIAM.
6. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.
7. ศุภการ, ส.ก., หลัก และ คำพิพากษา: กฎหมายอาญา. 7 ed. 2560: บริษัท อมรินทร์พริ้นติ้ง แอนด์พับลิชชิ่ง จำกัด (มหาชน).
8. Haykin, S., *Neural networks: a comprehensive foundation*. 1994: Prentice Hall PTR.
9. Larson, R.R., *Introduction to information retrieval.* Journal of the American Society for Information Science and Technology, 2010. **61**(4): p. 852-853.
10. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression*. Vol. 398. 2013: John Wiley & Sons.
11. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
12. Wall, M.E., A. Rechtsteiner, and L.M. Rocha, *Singular value decomposition and principal component analysis*, in *A practical approach to microarray data analysis*. 2003, Springer. p. 91-109.

13. Potisuk, S. *Typed dependency relations for syntactic analysis of Thai sentences*. in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. 2010.
14. Xu, Z. and S. Zhu. *Filtering offensive language in online communities using grammatical relations*. in *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*. 2010.
15. Chen, Y., et al. *Detecting offensive language in social media to protect adolescent online safety*. in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. 2012. IEEE.
16. Van Hee, C., et al. *Automatic detection and prevention of cyberbullying*. in *International Conference on Human and Social Analytics (HUSO 2015)*. 2015. IARIA.
17. Van Hee, C., et al., *Automatic Detection of Cyberbullying in Social Media Text.* arXiv preprint arXiv:1801.05617, 2018.

ภาคผนวก

ในภาคผนวกจะแสดงคำทั้งหมดที่ใช้ในการทดลอง ซึ่งเก็บรวบรวมจากคำพิพากษาศาลฎีกา และเพิ่มคำเหมือนที่มีความหมายเหมือนกันเข้าไปด้วย โดยแบ่งออกเป็น คำนาม คำกริยาที่กรรมไม่มีผล คำกริยาที่กรรมมีผล คำหยาบ และวลี นอกจากนี้ยังมีคำสรรพนามบุรุษที่หนึ่ง คำสรรพนามบุรุษที่สองและคำสรรพนามบุรุษที่สาม

คำนาม

คนพาล, คนสารเลว, คนไม่เอาถ่าน, คนบ้า, คนเสียจริต, ผู้วิกลจริต, คนหัวรุนแรง, พวกหัวรุนแรง, กิ๊ก, เมียเก็บ, แมงดา, คนเกาะผู้หญิงกิน, คนให้ผู้หญิงเลี้ยง, พ่อเล้า, คนคุมซ่อง, นักเลงคุมซ่อง, คนไม่เต็ม, คนไม่เต็มบาท, คนโกง, คนคดโกง, คนหลอกลวง, นักต้มตุ๋น, ขโมย, หัวขโมย, ขโมยขโจร, คนขี้ขโมย, โจรลักขโมย, คนร้าย, ผู้ร้าย, คนนอกกฎหมาย, คนทำผิดกฎหมาย, ผู้กระทำความผิด, ฆาตกร, นักฆ่า, คนเลว, คนชั่ว, ปชช., ประชาชน, คนชนบท, ชาวบ้าน, คนบ้านนอก, คนต่างจังหวัด, คนในหมู่บ้าน, ตัวโกง, นักโทษ, ผู้ต้องโทษ, ผู้ถูกคุมขัง, ซ่อง, หอนางโลม, ซ่องโสเภณี, ซ่องโจร, ที่กบดาน, ที่ซ่อนตัว, รังโจร, รังอาชญากร, อาบอบนวด, สถานอาบอบนวด, คุก, ซังเต, ห้องขัง, เรือนจำ, ของปลอม, ของเก๊, ของเทียม, ของปลอมแปลง, ซา, หลุม, คนหยาบคาย, ศัตรู, ข้าศึก, สปาย, สายลับ, คนทรยศ, ผู้ทรยศ, คนขายชาติ, คนทรยศชาติ, จราจล, การจราจล, หนี้, หนี้สิน, ปีศาจ, วิญญาณร้าย, ผี, ผีสาง, วิญญาณ, อมนุษย์, คนเลวทราม, คนถ่อย, อันธพาล, คนไร้สกุลรุนชาติ, ตุ๊ด, กระเทย, ยาเสพติด, ยา, ไก่, อีตัว, โสเภณี, นางคณิกา, หญิงบริการ, ผู้หญิงหากิน, ผู้หญิงขายตัว, หญิงขายบริการทางเพศ, กุ๊ย, นักเลง, นักเลงโต, โจ๋อันธพาล, นักเลงหัวไม้, ผู้ก่อการร้าย, อีหนู, อนุภรรยา, นางบำเรอ, สาวฮาเร็ม, รัฐ, ชาติ, ประเทศ, รัฐชาติ, ประเทศชาติ, สัญชาติ, คนโง่, ไอ้ควาย, คนสมองทึบ, หัวขี้เลื่อย, คนโง่เขลา, ตัวตลก, คนโง่เซ่อ, คนเซ่อซ่า, คนไร้ค่า, คนไม่เอาไหน, คนน่ารังเกียจ, คำสั่ง, มิสชั่น, โครงการ, ภาระ, หน้าที่, บ้าน, เรือน, บ้านช่อง, ศีลธรรม, จริยธรรม, คนไร้การศึกษา, มาลีฮวนน่า, ต้นกัญชา, ส่วนได้เสีย, ค่า, ประโยชน์, คอมมิวนิสต์, พวกคอมมิวนิสต์, กบฏ, ผู้ก่อกบฏ, ผู้ก่อการกบฏ, เผด็จการ, ลัทธิเบ็ดเสร็จนิยม, จอมเผด็จการ, ผู้อยู่เหนือกฎหมาย, กฎ, ระเบียบ, กฎระเบียบ, ระเบียบข้อบังคับ, กม., กฎหมาย, สมรรถภาพ, แก๊ง, แก๊งค์อันธพาล, แก๊งค์อาชญากรรม, ประชาธิปไตย, สมอง, ปัญญา, ความคิด, เหตุผล, คนหน้าไหว้หลังหลอก, คนเบี้ยวหนี้, คนเจ้าชู้, ชายเจ้าชู้, คนเจ้าชู้ไก่แจ้, คนเจ้าชู้ประตูดิน, คนเถื่อน, คนชั้นต่ำ, คนโง่, คนไร้ค่า, คนไม่เอาไหน, คนน่ารังเกียจ, อันธพาล, นังเลง, จิ๊กโก๋, นังเลงหัวไม้, เด็กแว้น, คนเกเร, คนโรคจิต, คนวิปริต, คนสติไม่ดี, ผีบ้า, คนสิ้นคิด, คนบ้าเลือด, เด็กแก๊ง, เมียน้อย, บ้านเล็ก, บ้านน้อย, เด็กป๋า, เส้นสาย, ฉลากเข้า, ผัวน้อย, ผัวเก็บ, คนเล่นชู้, คนคบชู้, ม่านรูด, ห้องชั่วคราว, คุณไสย, ของดำ, ของมืด, มนต์ดำ, มนดำ, เวทมนดำ, คนวิกรจริต, คนจิตฟั่นเฟือน, คนไม่สมประกอบ, คนขาดสติ, คนเสียสติ, คน

วิปลาส, คนวิกลจริต, คนโกงกิน, คนฉ้อโกง, คนฉ้อฉล, สิปแปดมงกุฎ, โจร, ไอ้ชั่ว, ไอ้สารเลว, ชาติชั่ว, ไอ้คนจรรลัย, คนนอกกฎหมาย, ขี้คุก, คนคุก, ไอ้เลว, แหล่งซ่องสุม, ซ่องกะหรี่, บ้านโคมแดง, แหล่งกบดาน, อนาจาร, คนเงี่ยน, ขี้เงี่ยน, คนหื่นกาม, คนหื่น, คนลามก, ของก๊อป, ของเถื่อน, ของผิดลิขสิทธิ์, สายเขียว, ปุ๊น, เป็ด, พันลำ, จ๊อย, วิ๊ด, คนก้าวร้าว, คนนุนแรง, คนขวานผ่าซาก, คนโผงผาง, ฝั่งตรงข้าม, ไส้ศึก, หนอนบ่อนไส้, ผู้รุกราน, พวกดาวแดง, คนไม่รักดี, พวกเรือนขี้บนหลังคา, พวกไม่รักชาติ, พวกปากว่าตาขยิบ, อาวุธ, ผู้หลบหนี, ผู้ทำผิดกฎหมาย, ผู้ผ่าฝืนกฎหมาย, นิสัยเสีย, นิสัยไม่ดี, สันดานแย่, สันดานเสีย, สันดานหมา, ไอ้คนพ่อแม่ไม่สั่งสอน, บักสันดาน, หนี้นอกระบบ, คนหนีหนี้, คนชักดาบ, นกสองหัว, คนตีสองหน้า, คนจับปลาสองมือ, ผีผนัน, ผู้มีอิทธิพล, คนมีเส้น, เส้นใหญ่, เด็กเส้น, ทาส, ไพร่, สถุน, เบ๊, คุกขี้ไก่, ส่วย, เงินใต้โต๊ะ, เงินสินบน, คนบาป, คนไร้สัจจะ, ไอ้ตุ๊ด, ไอ้เกย์, สารเสพติด, มัน, ปัญญาอ่อน, ปัญญาน้อย, บัวใต้ตม, ตลาดล่าง, ตีน, ส้นตีน, กะหรี่, กระหรี, หรี, หญิงแพศยา, ดอกทอง, เวศยา, คณิกา, หญิงขายบริการ, คนอันธพาล, คนต่ำช้า, คนต่ำตม, คนจัญไร, คนฟั่นเฟื่อน, คนสติวิปลาส, แพศยา, คนแพศยา, อันตพาล, เมีย, ภรรยา, ชู้, เมียลับ, ภรรยาลับ, ผัว, ภริยา, สามี, แฟน, เจ้าชู้, ปัญหา, โฮเต็ล, โรงแรม, ไสยศาสตร์, ไสยศาสตร, ไสยศาส, เงิน, บ้า, คนประสาท, ประสาท, ฟั่นเฟือง, สติฟั่นเฟือง, ประสาทหลอน, วิปริต, บ้านเมือง, ประชาชาติ, แผ่นดิน, ทุจริต, คนทุจริต, คนเลวร้าย, คนชั่วร้าย, งาน, ความรับผิดชอบ, ความเคารพ, ราชการ, เมือง, คนขายบ้านขายเมือง, จรรยาบรรณ, ความซื่อสัตย์สุจริต, ความซื่อสัตย์, ศีลธรรมจรรยา, จิตสำนึก, คนไม่ดี, นู้ด, โป๊, คนร่าน, เนื้อ, กัญชา, ผลประโยชน์, คนเห็นแก่ได้, ความโปร่งใส, มารยาท, คนไร้มารยาท, ศีล, คู่อริ, คู่อาฆาต, ปฏิปักษ์, เสี้ยนศึก, เสี้ยนหนาม, คู่แค้น, คอมมิวนิส, คอมมิวนิด, คอมมิวนิต, กบฏ, ขบถ, ปืน, รัฐบาล, คนทราม, จอมบงการ, ลัทธิเผด็จการ, ผู้บงการ, อำนาจ, ข้อบังคับ, คุณภาพ, คุณสมบัติ, ความสามารถ, พวกโจร, แก๊งค์, คนนิสัยเสีย, นิสัย, คนสองหัว, คนสองแคม, คนเหยียบเรือสองแคม, ที่ดิน, บ่อน, พนัน, อิทธิพล, กฎหมาย, วินัย, ระเบียบวินัย, คนชาติหมา, มาร, ขี้ข้า, คนขี้คุก, ตะราง, ชีวิต, ขี้ตะราง, คอรับชั่น, คอรัปชั่น, คอรับชัน, คอรัปชัน, ฉ้อโกง, ฉ้อฉล, ตัว, คนสันดานเสีย, สันดาน, คนอัปปรี, คนอัปปรีย์, ใต้โต๊ะ, ประชาธิปไต, ความสามัคคี, ยาบ้า, ยาม้า, ยาอี, ยาไอซ์, ไอซ์, เจ้า, จ้าว, หัวคิด, สติ, เท้า, วุฒิภาวะ, อิตัว, ยางอาย, ลูกกะหรี่, ลูกกระหรี่, กระหรี่, คนโง่, คนไร้ค่า, คนไม่เอาไหน, คนน่ารังเกียจ

คำกริยาที่กรรมไม่มีผล

คบชู้สู่ชาย, จี้ปล้น, จี้ชิงทรัพย์, จี้, ข่มขืน, ข่มขืนกระทำชำเรา, ทรยศ, หักหลัง, รีดไถ, ป้ายสี, ใส่ความ, กล่าวร้าย, พูดให้ร้าย, ใส่ร้ายป้ายสี, หลีกเลี่ยง, หลบหนีจากที่คุมขัง, ทำอันตราย, หลอกลวง, คิดไม่ซื่อ, ไม่ซื่อสัตย์, โกหก, โป้ปด, ตอแหล, พูดปด, พูดโกหก, กล่าวเท็จ, ปั้นน้ำเป็นตัว, แอบอ้าง, กบฏ,

ก่อการกบฏ, รุมโทรม, โทรมหญิง, เรียงคิวขึนใจ, เรียงคิวข่มขืน, ยกเค้า, โจรกรรม, ขึ้นบ้าน, ปล้น, ชิงทรัพย์, ปล้นทรัพย์, ปล้นสะดม, ขโมย, ลักขโมย, ลักเล็กขโมยน้อย, ลอบวางเพลิง, ข่มขู่, แบล็กเมล์, เสพ, พี้, ขโมย, ลักทรัพ, ลักทรัพย์, กรรโชกทรัพ, กรรโชกทรัพย์, ต้มตุ๋น, เล่นชู้, ตีท้ายครัว, เล่นหลังบ้าน, คบชู้, สร้างบ้านเล็ก, มีบ้านเล็ก, มีเพศสัมพันธ์, กระทำชำเรา, ขืนใจ, ล่วงละเมิดทางเพศ, เสียตัว, เจาะไข่แดง, ล่าแต้ม, เก็บแต้ม, มีปากเสียง, ลงไม่ลงมือ, ชักชวนในทางที่ผิด, ทุจริต, คดโกง, ละเมิดกฎหมาย, ฝ่าฝืนกฎหมาย, แหกกฎ, ป้ายความผิด, นั่งเทียนเขียนข่าว, สร้างข่าวลือ, ก่อวินาศกรรม, ก่อเหตุอัคคีภัย, วางระเบิด, รุมประชาทัน, พรากชีวิต, ประทุษร้าย, ขู่ฆ่า, ใช้เส้นสาย, ขู่เข็ญ, ล่วงล้ำ, ฝ่าเขตแดน, ละเมิดสิทธิ, รุกราน, โกงกิน, เล่นเส้นเล่นสาย, ทรยศแผ่นดิน, รับสินบน, รับใต้โต๊ะ, ฮั้วประมูล, ผิดคำพูด, เล่นของสูง, แอบอ้างเบื่องสูง, ก่อจลาจล, โกง, ขบถ, กบฏ, ตลบหลัง, เนรคุณ, แทงข้างหลัง, ขู่กระโชก, ติดสินบน, ปลิ้นปลอก, ปลิ้นปล้อน, แหกตา, สับปลับ, หลอกต้ม, ลุมโทรม, ลวนลาม, ปลอม, ปลอมแปลง, ต้ม, ตุกติก, ยุยง, ปลุกระดม, ปลุกปั่น, ปั่นหัว, เป่าหู, ยุ, วิ่งราว, ปล้นจี้, ทำชั่ว, วางเพลิง, ขู่ขวัญ, พาล, แบล็คเมล์, แบลคเมล, แบล็คเมล, แบคเมล์, แบคเมล, แบล็คเม, แบลคเม, แบคเม, คอรับชั่น, คอรัปชั่น, คอรับชัน, คอรัปชัน, วิปริต, ฉ้อโกง

คำกริยาที่กรรมมีผล

หนี, ละทิ้ง, ชอบ, โปรด, โปรดปราน, นอน, เย็ด, ตีหม้อ, ร่วมรัก, มีเซ็กส์, ร่วมประเวณี, มีเพศสัมพันธ์, ตื้อ, ตามจีบ, ตามตื้อ, นอกใจ, สวมเขา, โจมตี, เข้าโจมตี, ชก, อัด, ตุ๊ย, ฟาดหัว, ฟาดศีรษะ, ตีหัว, ตีศีรษะ, อม, ดูด, เลีย, ฉุด, ย่ำยี, ปรนเปรอ, ขาย, ไถ, คุกคาม, เข้ามาคุกคาม, บีบ, กดดัน, บีบคั้น, บีบบังคับ, ลงแขก, สมคบคิด, ฮั้ว, ฉวย, หยิบฉวย, งัด, ดูถูก, ดูหมิ่น, เหยียดหยาม, บิดเบือน, หลอก, ล่อลวง, ก๊อป, ก๊อปปี้, เลียนแบบ, อิจฉา, ตาร้อน, อิจฉาตาร้อน, กุข่าว, ปล่อยข่าว, ประโคมข่าว, ปล่อยข่าวลือ, กุ, จับกุม, จับ, ควบคุม, คว้าตัว, จับตัว, เสือก, นั่ง, เฝ้า, หลับนอน, เป็น, คือ, มี, ทะเลาะ, ปะทะคารม, ไล่ออก, เถลไถล, สวม, ขึ้น, กอด, จูบ, จุบ, จุ๊บ, หอม, ได้, รับ, ได้รับ, เอา, กิน, หยิบ, ยึด, ใช้, ลุ่มหลง, หลง, ซื้อ, กินกริบ, ให้, บำเรอ, ลวงตา, ชิง, หมิ่น, แกล้ง, อำพราง, ทำ, ปลด, เผา, ยิง, เชื่อ, แสวงหา, บ้า, ขัด, บริหาร, วาง, กล่าวหา, ปฏิบัติ, กระทำ, สมคบ, รวมหัว, งุบงิบ, สมรู้, คบ, จ้าง, งาบ, เม้ม, ขาด, ขาดแคลน, ไร้, โอน, แอบ, หอบ, บริจาค, เข้า, ลงโทษ, กลัว, จ้างวาน, ทุบที, ทุบ, เฆี่ยน, หวด, ซัด, คร่า, เตะ, ถีบ, เปิด, ติด, ค้า, อ้าง, บังคับ, แดก, ทิ้ง, แดรก, กก, นอนกก, คลอเคลีย, เคลียเคล้า, โอบกอด, เยด, ร่วมเพศ, ปู้ยี่ปู้ยำ, ทำลาย, แย่ง, เหยด, ซี่, ซั่ม, ถูกจับ, โดนจับ, ฉกฉวย, กอบโกย, ฉกชิง, ช่วงชิง, กระชาก, จู่โจม, ต่อย, รุมสกัม, ตบ, วิวาท, ปะทะปะทั่ง, ตะลุมบอน, กัดกัน, ทำร้ายกัน, ทำร้าย, หยิบเอา, ทุบตี, ตี, ตบตา, ลวงหลอก, หลงผิด, หน้ามืดตามัว, ไม่ลืมหูลิมตา, ปรนนิบัติ, มัวเมา, เสแสร้ง, แสร้งทำ, รังแก, กดขี่ข่มเหง, ข่มเหงรังแก, กลั่น

แกล้ง, ข่มแหง, สังหาร, ฆ่า, ลั่นไก, เหนี่ยวไก, วิ่ง, เผ่น, หลบหนี, ซ่อนตัว, กบดาน, ลักลอบ, ด้นข่าว, กำจัด, สปอย, ก่อไฟ, เผาทำลาย, ละเมิด, รุกล้ำ, ริศยา, คดิคด, ยักยอก, ฉ้อฉล, กินเศษกินเลย, กินนอกกินใน, กะล่อน, คบคิด, สุมหัว, สมรู้ร่วมคิด, ระเบิด, ฉก, ลัก, ระราน, ข่มเหง, ข่มเหงคะเนงร้าย, ขู่, บุกรุก, ละลาบละล้วง, ล่วงเกิน, จาบจ้วง, ปล้ำ, มั่วสุม, มุสา, บุก, ขี้จุ๊, ลอกเลียนแบบ, ก็อป, กอป, ก๊อบ, ก็อบ, กอบ, ก็อปปี้, กอปปี้, ก๊อบปี้, ก็อบปี้, กอบปี้, โหน, แย่งชิง, ริษยา, อิจฉาริษยา, ฉ้อ, ล้างผลาญ, กุเรื่อง, หลอย, จิ๊ก, ใส่ไฟ

วลี หรือ กลุ่มคำ

เอาดีใส่ตัวเอาชั่วใส่คนอื่น, เอาดีใส่ตัว, เอาชั่วใส่คนอื่น, ลิ้นสองแฉก, นกสองหัว, ไร้สมอง, ไม่มีสมอง, เหยียบเรือสองแคม, มวยล้ม, ล้มมวย

คำหยาบ

อีเหี้ย, อี, อิ, เหี้ย, เชี่ย, เฮี้ย, ควย, สัส, สัด, สิส, สีส, สิด, สีด, สัต, สัตว์, ควาย, ห่า, กระจอก, หี, เย็ด, แม่ง, แม้ง, แตด, หำ, หน้าหี, หัวควย, ดอก, เฮงซวย, ระยำ, เบื๊อก, เบื๊อก, โคตรพ่อ, โคตรแม่, ปอบ, ปรอบ, ปอป, ปอรป, เปรต, อ้วน, จัญไร, จันไร, ควย, เฮงซวย

คำสรรพนามบุรุษที่หนึ่ง

ผม, ฉัน, ดิฉัน, กระผม, ข้าพเจ้า, กู, เรา, ข้าพระพุทธเจ้า, อาตมา, หม่อมฉัน, เกล้ากระหม่อม, ข้า

คำสรรพนามบุรุษที่สอง และคำสรรพนามบุรุษที่สาม

มึง, ท่าน, คุณ, นาย, โจทก์, โจทก์ร่วม, ลูกสะใภ้, เจ้, เจ๊, ไอ้นี่, ไอ้นั่น, คนนี้, คนนั้น, ไอนี่, ไอ่นี่, ไอนั่น, ไอ่นั่น, ลุง, ป้า, ไอ้แก่, อีแก่, อิแก่, ไอแก่, ไอ่แก่, แก, ป้าแก, น้า, อา, เขา, เค้า, พี่, นาง, ตา, ยาย, เธอ, ไอ, ไอ้, ไอ่, เอ็ง, มัน, ไอ้เหี้ยนี่, ไอเหี้ยนี่, ไอ่เหี้ยนี่

ประวัติผู้เขียน

ชื่อ-สกุล	รัชกฤต อารีราษฎร์
วัน เดือน ปี เกิด	5 กรกฎาคม 2537
สถานที่เกิด	ขอนแก่น
วุฒิการศึกษา	ปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2559
ที่อยู่ปัจจุบัน	143/10 ถนนถีนานนท์ ตำบลตลาด อำเภอเมือง จังหวัดมหาสารคาม 44000