วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ ปีที่ 1 ฉบับที่ 3 พฤษภาคม – มิถุนายน 2566
Journal of Engineering and Industrial Technology, Kalasin University, Vol. 1 No. 3 May – June 2023

บทความวิจัย (Research Article)

การลดคุณลักษณะสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์โดยการใช้รูปแบบข้อมูลแนวตั้ง

อัจฉรา ชุมพล[1,*], มงคล แสนสุข[2]

[1] สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์
[2]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร

*ผู้ประสานงานบทความต้นฉบับ: atchara.ch@ksu.ac.th
(รับบทความ: 1 มิถุนายน 2566; แก้ไขบทความ: 17 มิถุนายน 2566; ตอบรับบทความ; 26 มิถุนายน 2566)

บทคัดย่อ

งานวิจัยนี้ ทำการพัฒนาขั้นตอนวิธีในการลดคุณลักษณะ โดยไม่ก่อให้เกิดการสูญเสียคุณลักษณะที่ส่งผลต่อการจำแนก เพื่อนำไปสู่การลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความถูกต้องการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ โดยทำการเปรียบเทียบการลดคุณลักษณะโดยการใช้รูปแบบข้อมูลแนวตั้งกับการลดคุณลักษณะด้วยวิธีการไคสแคว์ จำแนกความคิดเห็นด้วยวิธีการนาอีฟเบย์ (Naïve Bayes) ข้อมูลที่ใช้ในการวิจัยรวบรวมจาก Stanford Twitter Sentiment Data ผลการวิจัยพบว่า วิธีการลดคุณลักษณะโดยการใช้รูปแบบข้อมูลแนวตั้งมีประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้องในการจำแนก เท่ากับ 72.64%

คำสำคัญ: การลดคุณลักษณะ การจำแนกความคิดเห็น เครือข่ายสังคมออนไลน์ รูปแบบข้อมูลแนวตั้ง

การอ้างอิงบทความ: อัจฉรา ชุมพล และ มงคล แสนสุข, "การลดคุณลักษณะสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์โดยการใช้รูปแบบข้อมูลแนวตั้ง," *วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์,* vol. 1, no. 3, pp. 38-45, 2566.

วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ ปีที่ 1 ฉบับที่ 3 พฤษภาคม – มิถุนายน 2566
Journal of Engineering and Industrial Technology, Kalasin University, Vol. 1 No. 3 May – June 2023

บทความวิจัย (Research Article)

Features Elimination for Opinion Classification on Social Networks using Vertical Data Format

Atchara Choompol[1,*] and Mongkol Saensuk[2]

[1] Department of Computer Engineering, Faculty of Engineering and Industrial Technology, Kalasin University
[2] Department of Business Computer, Faculty of Management Science, Sakon Nakhon Rajabhat University

* Corresponding Author: atchara.ch@ksu.ac.th
(Received: June 1, 2023; Revised: June 17, 2023; Accepted: June 26, 2023)

Abstract

This research has developed an algorithm for reducing features without affecting classification, leading to a decrease processing time and enhancing classification accuracy. The reduction of features using the vertical data model method was contrasted with the reduction of features using the chi-square method. The Naïve Bayes method was used to classify opinions. The data used in the research was collected from Stanford Twitter Sentiment Data. The most effective feature reduction method employs a vertical data model provides an average efficiency of 72.75%.

Keywords: Features Elimination, Opinion Classification, Social Networks, Vertical Data Format

Please cite this article as: A. Choompol and M. Saensuk, "Features Elimination for Opinion Classification on Social Networks using Vertical Data Format," *Journal of Engineering and Industrial Technology, Kalasin University,* vol. 1, no. 3, pp. 38-45, 2023.

บทความวิจัย (Research Article)

1. บทนำ

ในปัจจุบันมีองค์กรและหน่วยงานต่างๆ นำข้อมูลบนเครือข่ายสังคมออนไลน์มาใช้เพื่อการวิเคราะห์เป็นจำนวนมาก เนื่องจากข้อมูลเหล่านั้นสามารถสะท้อนให้เห็นถึงผลการดำเนินงานในองค์กรหรือหน่วยงานหลากหลายมิติ เช่น องค์กรที่เกี่ยวกับด้านการตลาดได้นำข้อมูลความคิดเห็นของลูกค้าไปวิเคราะห์เพื่อติดตามทัศนคติของผู้บริโภคที่มีต่อสินค้าหรือบริการเพื่อให้ทราบความต้องการที่แท้จริงของผู้บริโภค ซึ่งนำไปสู่การจัดทำแผนการตลาด การจัดโปรโมชั่นให้ถูกใจผู้ใช้บริการในกลุ่มต่างๆ [1] ด้านการเมืองมีการนำข้อมูลบนเครือข่ายสังคมออนไลน์ไปวิเคราะห์เพื่อสำรวจทัศนคติของประชาชนที่มีต่อพรรคหรือนักการเมืองหรือเพื่อทำนายผลการเลือกตั้ง [2] ด้านการศึกษาใช้ติดตามทัศนคติของผู้เรียนเพื่อนำไปปรับปรุงการจัดการเรียนการสอนให้มีประสิทธิภาพยิ่งขึ้น [3] เป็นต้น

เว็บไซต์ทวิตเตอร์ (Twitter) เป็นเครื่องมือโซเชียลมีเดียที่ได้รับความนิยมเป็นอย่างมากในปัจจุบัน มีผู้ลงทะเบียนมากกว่า 300 ล้านคน และทวีตมากกว่า 500 ล้านครั้งต่อวัน [4] ผู้คนสามารถแลกเปลี่ยนข้อมูลได้อย่างอิสระและรวดเร็ว เว็บไซต์ทวิตเตอร์จึงกลายเป็นโซเชียลมีเดียลที่มีข้อมูลสำคัญจำนวนมาก โดยเฉพาะข้อมูลความคิดเห็นเกี่ยวกับเหตุการณ์ สินค้า องค์กรหรือนักการเมือง อย่างไรก็ตามด้วยข้อจำกัดของทวิตเตอร์ซึ่งอนุญาตให้แต่ละโพสต์มีเพียง 140 ตัวอักษร จึงมีการใช้ตัวย่อและคำแสลง เป็นจำนวนมาก ด้วยเหตุนี้การเพิ่มความแม่นยำในการจำแนกความคิดเห็นบนทวิตเตอร์จึงเป็นเรื่องที่ยังต้องพัฒนาให้มีประสิทธิภาพมากขึ้นต่อไป [5] การจำแนกความคิดเห็นด้วยวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นวิธีการหนึ่งที่ได้รับความนิยม ซึ่งก่อนนำไปเข้าสู่กระบวนการจำแนกความคิดเห็น จำเป็นจะต้องทำการแปลงข้อความคิดเห็นให้อยู่ในรูปแบบเวกเตอร์ โดยเวกเตอร์ที่สร้างขึ้นจะถูกจัดอยู่ในรูปแบบเวกเตอร์แนวนอน (Horizontal Vector) ซึ่งแต่ละแถวจะแทน มีจำนวนมิติเท่ากับจำนวนคุณลักษณะ (Feature) ทั้งหมดที่สกัดได้ โดยเวกเตอร์ที่ได้จะถูกจัดอยู่ในรูปแบบเวกเตอร์แนวนอน (Horizontal Vector) ชุดข้อมูลจะถูกจัดเก็บในรูปแบบของแถว แต่ละแถวแทนด้วยเอกสาร และประกอบด้วยค่าน้ำหนักของคำคุณลักษณะที่เป็นตัวแทนของเอกสารทั้งหมด ถ้าไม่ปรากฏคำคุณลักษณะในเอกสารค่าน้ำหนักจะมีค่าเป็น 0 เนื่องจากข้อความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ส่วนมากเป็นประโยคสั้นๆ และค่อนข้างกำกวม (Implicit Sentence) ไม่ได้ระบุถึงคุณลักษณะของสิ่งที่กล่าวไว้ชัดเจน [6] และในการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์จะมีคำคุณลักษณะ (Feature) จำนวนมาก ส่งผลให้ใช้เวลาในการประมวลผลค่อนข้างมาก การนำเสนอข้อมูลในรูปแบบแนวตั้ง (Vertical Data Format) เป็นอีกวิธีการหนึ่งที่ช่วยให้การวิเคราะห์ข้อมูลง่ายและรวดเร็วขึ้น

จากงานวิจัยที่ผ่านมาทำการลดจำนวนคุณลักษณะโดยการจัดอันดับคุณลักษณะ (Feature Ranking) ด้วยวิธีการต่างๆ เช่น การใช้ค่าการเพิ่มของข้อมูล (Information Gain) ค่าสถิติไคสแคว์ (Chi-Squared) และค่าสารสนเทศร่วม (Mutual Information) เป็นต้น [7-9] ซึ่งวิธีการนี้จะทำให้ได้คุณลักษณะที่มีความสำคัญสูง แต่อาจจะทำให้สูญเสียคุณลักษณะที่มีผลต่อการตีความของตัวจำแนก เนื่องจากคุณลักษณะที่มีความสำคัญนั้นอาจจะไม่ได้อยู่ในเอกสารที่ใช้ในการทดสอบ งานวิจัยนี้ผู้วิจัยจึงได้มีแนวคิดในการพัฒนาขั้นตอนวิธีในการลดคุณลักษณะ โดยไม่ก่อให้เกิดการสูญเสียคุณลักษณะที่ส่งผลต่อการจำแนก เพื่อนำไปสู่การลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความถูกต้องการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์

2. แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

การทำเหมืองความคิดเห็น (Opinion Mining) [10] เป็นการประเมินทัศนคติและอารมณ์ของผู้คนที่มีต่อสิ่งต่างๆ เช่น สินค้า บริการ องค์กร ประเด็นหรือเหตุการณ์ที่เกิดขึ้น เป็นต้น การทำเหมืองความคิดเห็นสามารถทำได้โดยการเก็บรวบรวมข้อมูลความคิดเห็นของผู้คนจากข้อความคิดเห็นที่อยู่บนเว็บไซต์ โดยเฉพาะอย่างยิ่งในสื่อสังคมออนไลน์ เช่น Facebook, LinkedIn, Twitter, Flickr และ YouTube เป็นต้น ที่จัดว่าเป็นสื่อที่มีผู้คนได้โพสต์แสดงความคิดเห็นเป็นจำนวนมาก ประเด็นหลักในการทำเหมืองความคิดเห็น คือ จำแนกความคิดเห็นเพื่อให้ทราบถึงความพึงพอใจของกลุ่มบุคคลว่ามีความรู้สึกในเชิงบวกหรือเชิงลบ การทำเหมืองความคิดเห็นได้นำหลักการทำเหมืองข้อความและการประมวลผลภาษาธรรมชาติมาประยุกต์ใช้ การจำแนกความคิดเห็น มี 3 วิธีการหลักที่ได้รับความนิยม ได้แก่ วิธีการใช้คลังคำ (Lexical Based) วิธีการเรียนรู้ของเครื่อง (Machine Learning) และวิธีการผสมผสาน (Hybrid Approach) วิธีการใช้คลังคำ เป็นการจำแนกความคิดเห็นโดยใช้พจนานุกรมคำที่ระบุขั้วความรู้สึกของคำไว้แล้ว ในงานวิจัยด้านการทำเหมืองความคิดเห็น เรียกว่า คำแสดงความรู้สึก (Sentiment Word) คำแสดงความคิดเห็น (Opinion Words) คำระบุขั้วความเห็น (Polar Word) หรือคำที่เป็นความคิดเห็น (Opinion Bearing Words) [11] พจนานุกรมที่ใช้ในการทำเหมืองความคิดเห็นส่วนมากจะมีคำแสดงความรู้สึก 2 ด้าน คือ ความรู้สึกเชิงบวก (Positive) และ ความรู้สึกเชิงลบ (Negative)

การเรียนรู้ของเครื่อง เป็นวิธีการพัฒนากระบวนการเรียนรู้เพื่อให้เครื่องคอมพิวเตอร์สามารถทำงานได้อย่างมีประสิทธิภาพโดยอาศัยหลักการเรียนรู้ของมนุษย์ แบ่งเป็น 2 เทคนิคหลัก [12] คือ การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นเทคนิคการเรียนรู้จากการนำข้อมูลที่มีอยู่ในอดีตมาสร้างสมการหรือรูปแบบของชุดข้อมูลสอน (Training Set) เพื่อการหาคำตอบให้กับข้อมูลชุดใหม่ (Test Set) ตัวอย่างวิธีการเรียนรู้แบบมีผู้สอนที่ได้รับความนิยม เช่น Decision Tree, Naïve Bayes, Support Vector Machines เป็นต้น และเทคนิค คือ การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นเทคนิคที่เน้นการนำข้อมูลที่มีอยู่มาศึกษาเพื่อพิจารณาหาความสัมพันธ์ของข้อมูลเป็นหลัก เทคนิคการเรียนรู้แบบไม่มีผู้สอน ประกอบด้วย การแบ่งกลุ่มข้อมูล (Clustering) และ การหากฎความสัมพันธ์ (Association Rule)

3. วิธีการดำเนินการวิจัย

3.1 การเก็บรวบรวมข้อมูล

ผู้วิจัยรวบรวมข้อมูลความคิดเห็นจาก Stanford Twitter Sentiment Data [13] ที่รวบรวมระหว่างวันที่ 6 เมษายน ถึงวันที่ 25 มิถุนายน ปี ค.ศ. 2009 ประกอบด้วย ข้อความคิดเห็นเชิงบวก (Positive) จำนวน 800,000 ข้อความ และข้อความคิดเห็นเชิงลบ (Negative) จำนวน 800,000 ข้อความ รูปแบบของข้อมูลเป็นไฟล์เอกสารประเภท Comma Separated Value (CSV) ประกอบด้วย 6 ฟิลด์ ได้แก่ 1) ขั้วความคิดเห็น มี 2 ด้าน คือ ความคิดเห็นเชิงลบ และ ความคิดเห็นที่เป็นเชิงบวก 2) ลำดับ มีรูปแบบเป็นตัวเลขจำนวนเต็ม 3) วัน/เวลา 4) ประเด็นที่โพสต์ 5) ชื่อผู้โพสต์ และ 6) ข้อความคิดเห็นตามลำดับ

3.2 การเตรียมข้อมูล (Data Preprocessing)

ข้อมูลที่ใช้ในการวิจัยเป็นข้อความที่รวบรวมจากเว็บไซต์ทวิตเตอร์ มีคุณลักษณะเด่น คือ ความยาวของตัวอักษร ไม่เกิน 140 อักขระ ประกอบด้วยแฮชแท็ก (Hash Tag: #) ที่ใช้อธิบายสถานะที่ต้องการเน้นประเด็นหรือความรู้สึกต่างๆ เป็นพิเศษ ใช้สัญลักษณ์

ข้อความแสดงอารมณ์ (Emoticons) ประกอบด้วยที่อยู่เว็บไซต์ (URL) การแท็กชื่อบุคคลขึ้นต้นด้วยเครื่องหมาย “@”ข้อความคิดเห็นที่โพสต์เป็นประโยคสั้นๆ ไม่มีโครงสร้างที่แน่นอน มีคำศัพท์สแลง อักขระพิเศษ และคำสะกดผิดค่อนข้างมาก การนำข้อความเข้าสู่กระบวนการจำแนกความคิดเห็น ต้องมีการเตรียมข้อมูลประกอบด้วย 4 ขั้นตอน ได้แก่ กระบวนการทำความสะอาดข้อความ (Cleaning) โดยการลบตัวอักษรซ้ำและแก้ไขคำที่สะกดผิด กระบวนการลบคำหยุดหรือคำที่ไม่มีนัยสำคัญในการจำแนกความคิดเห็น (Remove Stop words) โดยใช้คลังคำศัพท์ที่มีอยู่แล้ว จากนั้นทำการหารากคำศัพท์ (Stemming) โดยใช้คลังคำศัพท์ จากนั้นนำเข้าสู่กระบวนการตัดคำ (Tokenization) งานวิจัยนี้ทำการตัดคำด้วยวิธีการ Unigrams ร่วมกับ Bigrams โดยข้อความที่มีคำปฏิเสธ เช่น no, not จะใช้หลักการตัดคำด้วยวิธีการ Bigrams ส่วนข้อความอื่นๆ จะใช้หลักการตัดคำด้วย Unigrams จากนั้นทำการแทนค่าในเอกสาร (Document Representation) เพื่อให้ได้ข้อมูลที่อยู่ในรูปแบบเวกเตอร์เพื่อนำไปเข้าสู่กระบวนการจำแนกความคิดเห็น งานวิจัยนี้ทำการแทนค่าน้ำหนักในเอกสารด้วยวิธีการ Boolean Weighting (Index Documents by using Boolean Weighting) โดยหากพบคุณลักษณะในเอกสาร จะให้มีค่า เท่ากับ 1 หาก ไม่พบให้มีค่าเท่ากับ 0 ขนาดเวกเตอร์จะมีขนาดเท่ากับจำนวนเอกสาร

3.3 การกำจัดคุณลักษณะที่มีความซ้ำซ้อน

การจำแนกความคิดเห็นโดยทั่วไปจะใช้เวกเตอร์ข้อมูลในรูปแบบเวกเตอร์แนวนอน (Horizontal Vector) ประกอบไปด้วย เซตของเอกสาร $D=\{d_1,d_2,..,d_n\}$ เมื่อ n = จำนวนเอกสารทั้งหมด เซตของคุณลักษณะ $F=\{f_1,f_2,..,f_m\}$ เมื่อ m = จำนวนคุณลักษณะทั้งหมด และเซตของคลาส $C=\{c_1,c_2\}$ แต่ละเอกสาร $d\in D$ จะนำเสนอในรูปแบบ $(f_1,f_2,..f_k,c_i\}$ เมื่อ $f\in F$ และ $c\in C$ ดังตารางที่ 1

ตารางที่ 1 รูปแบบเวกเตอร์แนวนอน (Horizontal Vector)

Document id	f_1	f_2	...	f_m	Class
d_1	w_{11}	w_{12}	...	w_{1m}	c_1
d_2	w_{21}	w_{22}	...	w_{2m}	c_1
...	...	...	...	...	...
d_n	w_{n1}	w_{n2}		w_{nm}	c_2

ตารางที่ 2 ตัวอย่างข้อมูลที่อยู่ในรูปแบบเวกเตอร์แนวนอน

Doc. ID	Feature (F)						Class
	f_1	f_2	f_3	f_4	f_5	f_6	
d_1	1	0	1	0	1	0	c_1
d_2	0	1	0	1	0	0	c_1
d_3	0	1	1	1	0	1	c_1
d_4	0	0	1	0	0	0	c_2
d_5	1	0	0	0	1	0	c_2

จากตารางที่ 2 เขียนแทนในรูปแบบเซตของเอกสารดังนี้

$$d_1=\{f_1,f_3,f_5,c_1\}$$
$$d_2=\{f_2,f_4,c_1\}$$
$$d_3=\{f_2,f_3,f_4,f_6,c_1\}$$
$$d_4=\{f_3,c_2\}$$
$$d_5=\{f_1,f_5,c_2\}$$

บทความวิจัย (Research Article)

งานวิจัยนี้ทำแปลงเวกเตอร์ที่อยู่ในรูปแบบแนวนอน เป็นข้อมูลแนวตั้ง (Vertical Data) เพื่อให้ง่ายและรวดเร็วในการคำนวณตัดคุณลักษณะที่ไม่จำเป็นออกไป กำหนดให้ $t(f_i)$ คือ เซตของเอกสารที่มีคุณลักษณะ f_i เช่น $t(f_1)=\{d_1,d_3\}$ แสดงว่าปรากฏคุณลักษณะ f_1 ในเอกสาร d_1 และ d_3 และ $t(c_i)$ คือ เซตของเอกสารที่อยู่ในคลาส c_i ข้อมูลแนวตั้งแต่ละแถวจะประกอบไปด้วย f_i และ $t(f_i)$ ดังตารางที่ 3

ตารางที่ 3 รูปแบบข้อมูลแนวตั้ง (Vertical Data)

Transaction	Set of Documents
f_1	$\{d_1,d_5\}$
f_2	$\{d_2,d_3\}$
f_3	$\{d_1,d_3,d_4\}$
f_4	$\{d_2,d_3\}$
f_5	$\{d_1,d_5\}$
f_6	$\{d_3\}$

เมื่อทำการแปลงข้อมูลให้อยู่ในรูปแบบแนวตั้งแล้ว ผู้วิจัยทำการตัดคุณลักษณะเพื่อลดค่าศูนย์โดยใช้แนวคิด ดังต่อไปนี้

ถ้า $t(f_i) = t(f_j)$ โดยที่ $f_i, f_j \in F$ แสดงว่า f_i และ f_j ปรากฏอยู่ในเอกสารชุดเดียวกัน ให้ตัดตัวใดตัวหนึ่งออกได้ โดยการจะตัดตัวใดออกจะพิจารณาจากหน้าที่ของคำ (Part of Speech) จากตารางที่ 3 พบว่า f_1, f_5 เป็นคุณลักษณะที่ปรากฏในเอกสารเดียวกัน คือ d_1 และ d_5 และ f_2, f_4 ปรากฏในเอกสารเดียวกัน คือ d_2 และ d_3 จากนั้นผู้วิจัยจะทำการตรวจสอบหน้าที่ของคำคุณลักษณะเหล่านั้น โดยใช้ POS Tagger เพื่อเลือกคุณลักษณะที่ทำหน้าที่เป็นคำกริยาไว้ แล้วตัดคุณลักษณะที่เหลือออก ซึ่งจากตัวอย่างพบว่า คุณลักษณะที่ทำหน้าที่เป็นคำกริยา คือ f_1 และ f_2 ดังนั้น f_5 และ f_4 จะถูกตัดออก เพราะถือว่าเป็นคุณลักษณะที่ซ้ำกับ f_1 และ f_2 และเมื่อลบออกแล้วจะเห็นว่าจำนวนคุณลักษณะลดลง แต่เอกสารยังคงปรากฏคุณลักษณะอยู่ จึงไม่ทำให้สูญเสียข้อมูลที่ส่งผลต่อตัวจำแนก เมื่อผ่านกระบวนการลดคุณลักษณะที่ซ้ำแล้ว จะคงเหลือทรานเซคชั่นของคุณลักษณะ ดังนี้

$$t(f_1) = \{d_1, d_5\}$$
$$t(f_2) = \{d_2, d_3\}$$
$$t(f_3) = \{d_1, d_3, d_4\}$$
$$t(f_6) = \{d_3\}$$

ข้อที่ 2. ถ้า $t(f_i) \cap t(f_j) = t(f_k)$ โดยที่ $f_i, f_j, f_k \in F$ แสดงว่า ทุกเอกสารที่ปรากฏ f_k จะปรากฏ f_i และ f_j ด้วยเสมอ ดังนั้น $t(f_i)$ และ $t(f_j)$ จึงเป็นตัวแทน $t(f_k)$ สามารถตัด f_k ออกได้ เช่น ทรานเซคชั่นที่เหลือหลังจากผ่านกระบวนการในข้อที่ 1 พบว่า เมื่อ $t(f_2) \cap t(f_3)$ จะได้เท่ากับ $t(f_6)$ แสดงว่า f_2 และ f_3 เป็นตัวแทนของ f_6 ได้ ผู้วิจัยจึงตัด f_6 ออก จะคงเหลือทรานเซคชั่นของคุณลักษณะ ดังนี้

$$t(f_1) = \{d_1, d_5\}$$
$$t(f_2) = \{d_2, d_3\}$$
$$t(f_3) = \{d_1, d_3, d_4\}$$

เมื่อผ่านการบวนการตัดคุณลักษณะที่ซ้ำออกแล้ว จะต้องทำการแปลงข้อมูลในแนวตั้งกลับมาเป็นเวกเตอร์แนวนอนก่อนนำไปเข้าสู่กระบวนการจำแนก ซึ่งจะเห็นว่าเมื่อแปลงข้อมูลกลับมาอยู่ในรูปแบบเวกเตอร์แนวนอน มีจำนวนคุณลักษณะลดลง และจำนวนค่าศูนย์ลดลง แต่ทุกเอกสารยังคงปรากฏคุณลักษณะ ดังตารางที่ 4

ตารางที่ 4 ตัวอย่างข้อมูลหลังผ่านกระบวนการลดคุณลักษณะ

Document (D)	Feature (F)			Class
	f_1	f_2	f_3	
d_1	1	0	1	c_1
d_2	0	1	0	c_1
d_3	0	1	1	c_1
d_4	0	0	1	c_2
d_5	1	0	0	c_2

4. ผลการทดลองและอภิปรายผล

การวัดประสิทธิภาพการจำแนกความคิดเห็น ผู้วิจัยทำการสุ่มข้อมูลความคิด จำนวน 8,000 ความคิดเห็น ซึ่งประกอบด้วยความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบจำนวนเท่ากัน เปรียบเทียบการเลือกคุณลักษณะด้วยวิธีการตัดคุณลักษณะที่ซ้ำซ้อนดังที่ได้นำเสนอ กับการไม่เลือกคุณลักษณะ และการเลือกคุณลักษณะด้วยวิธีการวิธีการใช้ค่าสถิติไคสแคว์ แบ่งข้อมูลชุดสอน ข้อมูลชุดทดสอบ ด้วยวิธีการ 10-Fold Cross Validation ใช้ตัวจำแนกและทดสอบโดยวิธีการนาอีฟเบย์ เพื่อทดสอบผลการคัดเลือกคุณลักษณะ ผลการประเมินประสิทธิภาพการจำแนกความคิดเห็น พบว่า การจำแนกความคิดเห็นโดยการตัดคุณลักษณะที่ซ้ำซ้อน มีประสิทธิภาพความถูกต้องสูงกว่าการไม่เลือกคุณลักษณะและวิธีการใช้ค่าสถิติไคสแคว์ (Chi-square) ดังตารางที่ 5

ตารางที่ 5 ผลการทดลอง

วิธีการ	ค่าความถูกต้อง
ไม่เลือกคุณลักษณะ	70.68
เลือกคุณลักษณะโดยใช้ค่าสถิติไคสแคว์ (Chi-square)	71.76
เลือกคุณลักษณะโดยวิธีที่นำเสนอ	72.75

5. อภิปรายผลการประเมินประสิทธิภาพ

งานวิจัยนี้ผู้วิจัยจึงได้มีแนวคิดในการพัฒนาขั้นตอนวิธีในการลดคุณลักษณะ โดยไม่ก่อให้เกิดการสูญเสียคุณลักษณะที่ส่งผลต่อการจำแนก เพื่อนำไปสู่การลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความถูกต้องการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ จากผลการประเมินประสิทธิภาพการลดคุณลักษณะโดยการใช้รูปแบบข้อมูลแนวตั้ง พบว่า มีประสิทธิภาพดีกว่าการไม่เลือกคุณลักษณะ และวิธีการคัดเลือกคุณลักษณะด้วยวิธีการเลือกคุณลักษณะด้วยวิธีการใช้ค่าสถิติไคสแคว์ (Chi-square) แสดงให้เห็นว่าการเลือกตัดคุณลักษณะที่ซ้ำซ้อน มีผลต่อประสิทธิภาพการจำแนก และการตัดคุณสมบัติที่ซ้ำซ้อนอย่างมีประสิทธิภาพ จะช่วยให้การจำแนกมีประสิทธิภาพเพิ่มขึ้น

จากผลการวิจัยนี้สามารถนำวิธีการลดคุณลักษณะโดยการใช้รูปแบบข้อมูลแนวตั้งไปประยุกต์ใช้กับการตรวจสอบหรือการจำแนกประเภทข้อความอื่นได้ เช่น การตรวจสอบคัดกรองเอกสาร การจัดกลุ่มข้อความคิดเห็นบนเครือข่ายสังคมออนไลน์อื่นๆ การจำแนกความคิดเห็นของการให้บริการต่างๆ เป็นต้น

6. เอกสารอ้างอิง

[1] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, and J. Caro,

"Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in *Information, Intelligence, Systems and Applications (IISA),Fourth International Conference*, 2013, pp. 1-6.

[2] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference*, 2014, pp. 1-8.

[3] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior,* vol. 31, pp. 527-541, 2014.

[4] S. Aslam. "Twitter Statistics [omincoreagency.com]." https://www.omnicoreagency.com/twitter-statistics/ (accessed 1 March 2020).

[5] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for Twitter sentiment analysis," in *Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on the World Wide Web (WWW'12)*, Lyon, France, 2012.

[6] A. A. G. M. Karamibekr, "Sentiment Analysis of Social Issues," in *International Conference on Social Informatics*, Canada, 2012, pp. 215-221.

[7] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Inf. Process. Manage,* vol. 48, pp. 741-754, 2012.

[8] Q. Song, J. Ni, and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE Trans. on Knowl. and Data Eng.,* vol. 25, pp. 1-14, 2013.

[9] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," in *the Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[10] J. C. Hall, "A Linguistic Model for Improving Sentiment Analysis Systems," Master of ScienceThesis, North Dakota State University, Fargo, North Dakota, 2014.

[11] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[12] เอกสิทธิ์ พัชรวงศ์ศักดา, *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น*, 1 ed. กรุงเทพฯ:: เอเชีย ดิจิตอลการพิมพ์, 2557.

[13] A. Go, L. Huang, and R. Bhayani, *Twitter sentiment analysis*. 2009.