การแบ่งกลุ่มข้อความ SMS ตามลักษณะการให้บริการ

Service-Oriented Classifying of SMS Message

ชัยพร เขมะภาตะพันธ์
สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
กรุงเทพมหานคร, ประเทศไทย
chaiyaporn@dpu.ac.th

ภูรดา นนทวาสี
สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
กรุงเทพมหานคร, ประเทศไทย
napatsanun.n@cattelecom.com

บทคัดย่อ -- งานวิจัยนี้นำเสนอวิธีการแบ่งประเภทของข้อความ SMS โดยใช้วิธีการ Naïve Bayesian โดยพิจารณาจากเนื้อหาของข้อความ SMS เพื่อแก้ปัญหาความคับคั่งของ SMSC เมื่อมีปริมาณผู้ส่งเป็นจำนวนมาก โดยมีประโยชน์ต่อระบบ SMSC ในการสร้างลำดับการส่งใหม่ตามระดับความสำคัญของข้อความ SMS ที่กำหนดขึ้น ทำให้ลดอัตราเสี่ยงที่ระบบจะเกิดการ overload และการไม่สามารถให้บริการได้ ผลการศึกษายังสามารถใช้เป็นพื้นฐานเพื่อพัฒนาระบบคัดแยกระดับความสำคัญข้อความ SMS ที่จะนำไปใช้งานเชิงพาณิชย์ สำหรับผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยต่อไปในอนาคต จากผลการทดสอบแสดงให้เห็นว่า วิธีการคัดแบ่งประเภทของข้อความที่นำเสนอใช้เวลาในการทำงานน้อยกว่าการคัดกรองแบบเดิมถึง 6% นอกจากนี้ยังมีความถูกต้องในการทำงานสูงกว่า 13.59%

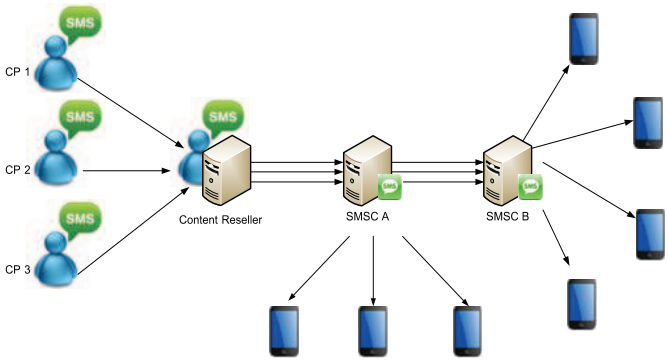
คำสำคัญ : ข้อความ SMS, การกรอง, การคัดแบ่ง, เนฟเบย์เซียน

Abstract – This research proposed a technique for classification of SMS messages by using Naïve Bayesian algorithm and the behavior of the content in an SMS message. The proposed system will be useful for SMSC by rearranging the order of sending an SMS according to the priority of an SMS type that is defined. This aims to solve the congestion at the SMSC when many subscribers are trying to send their SMSs. The system reduces the risk of overload and denial of service. The studied results can also be used as a basis to develop the commercial system that is used to classify type of SMS messages for mobile service operators in Thailand in the future. The results show that the proposed system performs 6% faster and 13.59% higher accuracy than the conventional one.

Keyword - SMS Message; Filtering; Classification; Naïve Bayesian

I. บทนำ

บริการเสริมที่มีผู้ใช้เป็นจำนวนมากที่สุดบริการหนึ่งของระบบโทรศัพท์เคลื่อนที่ในปัจจุบันคือ บริการส่งข้อความสั้นหรือ Short message service (SMS) ซึ่งถือเป็นช่องทางหนึ่งที่มีประสิทธิภาพสามารถเข้าถึงผู้ใช้งานได้รวดเร็วและตรงกลุ่มเป้าหมาย จึงถูกนำมาใช้อย่างแพร่หลายในปัจจุบัน เนื้อหาของข้อความมีหลายประเภท ซึ่งข้อความแต่ละข้อความมีเนื้อหาและระดับความสำคัญของข้อความแตกต่างกันขึ้นอยู่กับรูปแบบและวัตถุประสงค์การใช้งานในกิจกรรมนั้นๆ เช่น งานด้านการเงิน,การศึกษา รวมถึงเป็นสื่อโฆษณาประชาสัมพันธ์ต่างๆ ทำให้มีผู้หวังผลทางการค้าจำนวนมาก หรือ Reseller ที่ซื้อบริการส่งข้อความ SMS จากผู้ให้บริการโทรศัพท์เคลื่อนที่ เพื่อนำไปสร้างเป็นบริการประเภทต่างๆต่อไปอีก เพื่อนำไปขายช่วง เช่น บริการ Bulk SMS ที่เน้นให้บริการส่ง SMS เป็นจำนวนมาก ตามแต่จุดประสงค์ของลูกค้าที่มาซื้อบริการต่ออีกที ซึ่งแสดงรายละเอียดดังรูปที่ 1



รูปที่ 1 : *การให้บริการส่งข้อความจำนวนมากจากผู้ให้บริการ* SMS รายย่อยหรือ Reseller

จากรูปที่ 1 แสดงให้เห็นว่าปัจจุบันมีการส่ง SMS จำนวนมากจาก Content Provider (CP) ที่ซื้อบริการ SMS ต่อจาก Content Reseller ซึ่ง Reseller จะซื้อบริการ SMS จากผู้ให้บริการโทรศัพท์เคลื่อนที่ในจำนวนมาก ซึ่งในรูปซื้อบริการส่ง SMS จากผู้ให้บริการโทรศัพท์เคลื่อนที่ A ทำให้ได้ราคาต่อหน่วยข้อความ SMS ถูกกว่าปกติมาก และนำไปขายเป็นบริการให้กับ CP โดย CP จะส่งข้อความผ่านระบบของ Reseller จากนั้น Reseller จะทำการส่งข้อความดังกล่าวไปยัง Short Message Service Centre (SMSC) A เพื่อให้ SMSC A ทำการส่งต่อข้อความเหล่านั้นให้กับเลขหมายปลายทางที่กำหนดที่อยู่ในเครือข่ายการให้บริการของตนเอง และส่งให้กับ SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่อื่นๆ เมื่อเลขหมาย

ปลายทางไม่ได้อยู่ในเครือข่ายของตนเอง ซึ่งในรูปแสดงด้วย SMSC B ดังนั้น SMSC ทั้งสองจึงไม่สามารถคัดกรองข้อความ โดยใช้เบอร์ผู้ส่ง หรือ A Number ได้ เพราะไม่สามารถระบุตัวตนที่แท้จริงของผู้ส่งได้ ทั้งนี้เป็นเพราะ A Number จะเป็นของ Reseller แทน

งานวิจัยหลายงาน [1]-[7] ที่พยายามนำเสนอวิธีการคัดกรองหรือคัดทิ้งข้อความ SMS โดยการตรวจหาคุณสมบัติของตัวข้อความเพื่อแก้ไขปัญหาข้อความที่สร้างความรบกวนให้กับผู้ใช้บริการโทรศัพท์เคลื่อนที่ แต่งานวิจัยส่วนใหญ่มุ่งเน้นที่จะคัดแยกและหยุดทำการส่งข้อความเหล่านั้น ซึ่งไม่สามารถนำวิธีการเหล่านั้นไปใช้งานจริงในเชิงพาณิชย์ได้ เพราะผู้ให้บริการโทรศัพท์เคลื่อนที่ไม่สามารถหยุดส่งข้อความเหล่านั้นได้ เนื่องจากจะส่งผลกระทบต่อความน่าเชื่อถือของตัวระบบ และกระทบต่อรายได้ที่ผู้ให้บริการโทรศัพท์เคลื่อนที่จะได้รับอีกด้วยอีกด้วย แต่อย่างไรก็ตามข้อความบางประเภทกลับเป็นข้อความที่สร้างความรำคาญหรือมีระดับความสำคัญน้อยกว่าต่อผู้ใช้ในเวลาเดียวกัน จึงจำเป็นต้องมีการปรับปรุงระบบส่งข้อความ SMS เพื่อลดปัญหาดังกล่าว

ดังนั้นในงานวิจัยนี้จึงเสนอวิธีการแบ่งประเภทของข้อความ โดยใช้วิธีการ Naïve Bayesian (NB) [8] และ [9] เข้ามาช่วยในการคัดแยกข้อความประเภทต่างๆ ตามลักษณะการให้บริการ ซึ่งอาจจะพิจารณาได้จากหลายๆ องค์ประกอบด้วยกัน ไม่ว่าจะเป็น ผู้ส่ง เนื้อหาของข้อความ เป็นต้น ทั้งนี้เพื่อให้สามารถจัดลำดับความสำคัญของข้อความได้และทำการส่งข้อความเหล่านั้นตามลำดับความสำคัญก่อนหลัง โดยไม่มีการหยุดส่งข้อความ เพื่อให้สามารถใช้งานได้จริงในเชิงพาณิชย์ ซึ่งระบบดังกล่าวสามารถนำไปประยุกต์ใช้ได้กับ SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่ด้านส่งและด้านรับ (SMSC A และ SMSC B)

ในหัวข้อ II จะกล่าวถึงวิธีการคัดกรองที่ผ่านมาว่าแต่แบบใช้เทคนิควิธีอะไรในการคัดกรองข้อความขยะ หัวข้อถัดไปจะอธิบายถึงเทคนิควิธีที่งานวิจัยได้เลือกใช้ หัวข้อที่ IV จะนำเสนอวิธีการที่ใช้งาน หัวข้อ V นำเสนอผลการทดสอบที่ได้ และสุดท้ายหัวข้อVI จะได้สรุปผลและนำเสนอข้อเสนอแนะต่อไป

II. การคัดกรองข้อความ SMS

งานวิจัยที่นำทฤษฎีต่างๆ มาพัฒนาเพื่อแก้ปัญหาข้อความขยะหรือ Spam Mail หรือ Spam SMS มีดังต่อไปนี้

Content Based SMS Spam Filtering [1] นำเอาเทคนิคของ Bayesian Filtering ที่ใช้สำหรับกรอง Spam Mail มาประยุกต์ใช้ในการตรวจจับ SMS Spam เพื่อดำเนินการบล็อกข้อความเหล่านั้นและช่วยให้ลดปัญหา SMS Spam

Bogofilter [2] ใช้เทคนิคการตรวจจับด้วยทฤษฎี Bayesian หรือการหาความน่าจะเป็นในการคัดกรองข้อความ SMS

Dynamic Markov Compression (DMC) [3] เน้นการกรองข้อความจากข่าวสารที่ถูกบีบอัด

LOHIT Algorithm [4] เสนอการกรองข้อความโดยใช้ Open source spam filter ซึ่งมีหลักการคำนวณทางตรรกศาสตร์ คล้ายกับ Support Vector Machine (SVM) แต่มีความซับซ้อนในการคำนวณน้อยกว่า

การคัดกรองข้อความที่ใช้วิธี SVM โดยการวัดจากคุณสมบัติของตัวข้อความเองนำเสนอโดย [5]

การกรองข้อความ SMS ที่สามารถใช้งานได้ทั้งภาษาไทยและภาษาอังกฤษของบริการส่งข้อความ SMS บนเครือข่ายโทรศัพท์เคลื่อนที่ [6] และ [7] ได้ศึกษาวิธีการกรองแบบ SVM และ NB โดยได้ปรับปรุงการทำ Text Normalization และการใช้วิธีตัดคำแบบผสมด้วยการกรองข้อความ SMS ทั้งภาษาไทย ภาษาอังกฤษและภาษาไทยปนอังกฤษ โดยพบว่าแบบ SVM มีความถูกต้องในการกรองข้อความสูงกวาวิธีการแบบ NB แต่วิธีการกรองแบบ NB ใช้เวลาในการประมวลผลน้อยกว่ามาก

III. ขั้นตอนวิธี Naïve Bayesian

ในการศึกษาวิจัยนี้ได้เลือกขั้นตอนวิธี NB เป็นวิธีในการแบ่งประเภทของข้อความโดยอาศัยเนื้อหาของข้อความเป็นหลัก ซึ่งสามารถได้ทำงานรวดเร็วและไม่มีความซับซ้อนเหมาะสมกับสร้างเพื่อใช้งานได้จริง วิธีการของ NB คือการใช้วิธีการของ Bayes พร้อมสมมติฐานของการเป็นอิสระต่อกันของตัวแปรอิสระทุกตัว การใช้ความน่าจะเป็นในการแก้ปัญหาที่ไม่สามารถใช้หลักสถิติได้ การนำ NB มาใช้กับการคัดกรองประเภทของข้อความ โดยหาค่าความน่าจะเป็นของคำในข้อความที่มีโอกาสอยู่ในประเภทของข้อความนั้นๆ เปรียบเทียบกับค่าความน่าจะเป็นของคำในข้อความประเภทอื่นๆ ซึ่งหากเปรียบเทียบกันแล้วมีค่ามากที่สุดในกลุ่มข้อความใด แสดงว่า

ข้อความดังกล่าวน่าจะจัดอยู่ในกลุ่มข้อความนั้น อธิบายได้ตามสมการที่ (1) ถึง (4) ตามลำดับ

$$P(W_j \mid C_i) = \frac{\text{Frequency of Word in a given Class}}{\text{Frequency of each Class}} \quad (1)$$

$$P(C_i \mid W) = P(C_i) \times \prod_{j=1}^{v} P(w_j \mid C_i) \quad (2)$$

โดยที่ v คือจำนวนของคำที่ใช้ในการคัดแยกข้อความ

n_i คือจำนวนของคำใน class ที่ i

$P(C_i)$ คือค่าความน่าจะเป็นของแต่ละ class หาได้จาก จำนวนของข้อความของแต่ละ class / จำนวนของข้อความทั้งหมด

w_j คือคำในลำดับที่ j ของข้อความ *SMS* ที่ต้องการคัดแยก

ซึ่งผลที่ได้จากการคำนวณหาค่าความน่าจะเป็นตามสมการที่ (2) มีค่าน้อยมาก เนื่องจากจำนวนของคำที่จะพบได้ในขอบเขตของคำที่ใช้เป็นฐานข้อมูลที่มีเป็นพันหรือมากกว่า จึงทำให้มีค่าต่ำมาก ซึ่งสามารถแก้ปัญหาได้ตามสมการที่ (3) ดังนี้

$$P(C_i \mid W) = \log(P(C_i) \times \prod_{j=1}^{v} P(w_j \mid C_i)) \quad (3)$$

ดังนั้นเราสามารถคัดแยกข้อความได้ตามสมการที่ (4) ดังนี้

$$classify(c_1,...,c_n) = \arg\max\, p(W = w) \prod_{i=1}^{n} p(C_i = C_i \mid W = w) \quad (4)$$

IV. การคัดแบ่งประเภทกลุ่มข้อความ SMS

A. การศึกษาและวิเคราะห์

การประมวลผลข้อความ SMS เพื่อแยกข้อความประเภทต่างๆ จำเป็นต้องมีการทำ Text Normalization (TN) และการตัดคำ เพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผล งานวิจัยที่กล่าวถึงวิธีการตัดคำภาษาไทยและภาษาไทยปนอังกฤษที่ผ่านมา [6] และ [7] ใช้การตัดคำกับเอกสารที่ข้อความมีความถูกต้องตามหลักภาษาศาสตร์ คือ วิธีการตัดคำแบบยาวที่สุด (Longest Matching) จะตัดคำโดยตรวจสอบจากตัวอักษรแรกและตัวอักษรถัดไปตามลำดับจนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรม ในขณะที่ลักษณะข้อความ SMS ที่รับส่งของประเทศไทยนั้นพบว่ามีลักษณะของข้อความที่ไม่เป็นไปตามหลักภาษาศาสตร์ ทั้งภาษาไทยและภาษาอังกฤษจำนวนมาก เนื่องจาก SMS เป็นการสื่อสารที่ไม่จำเป็นต้องใช้ภาษาอย่างเป็นทางการ อีกทั้งข้อจำกัดของจำนวนตัวอักษรที่พิมพ์ได้ในข้อความ SMS ทำให้ต้องมีการเพิ่มกระบวนการ TN เพื่อตัดคำหรือตัวอักษรที่ไม่มีความหมายทิ้ง และการแทนคำบางคำด้วยคำที่มีความหมายแทน เพื่อให้รองรับกับลักษณะของข้อความ SMS ที่ใช้ในปัจจุบันมากขึ้น

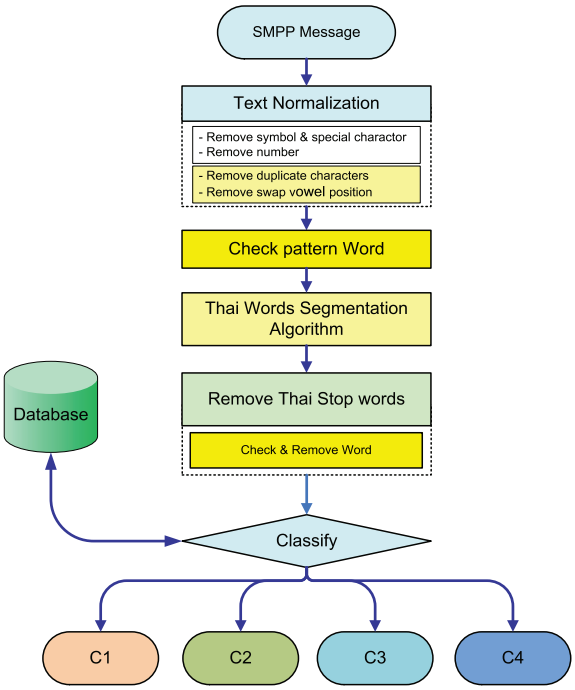
B. การคัดกรองข้อความ SMS แบบเดิม

[6] และ [7] เสนอระบบกรองข้อความ SMS จะถูกติดตั้งที่ SMSC หรือ SMS Gateway โดยเมื่อระบบรับข้อความจากชุมสาย (BTS/MSC) และถอดข้อความจาก SMPP Message ให้อยู่ในรูปของตัวอักษร Ascii แล้วจะผ่านเข้าสู่กระบวนการ TN เพื่อลบตัวอักษรที่ไม่สามารถตัดเป็นคำได้หรือไม่มีความหมายออกไป รวมถึงตัวเลขต่างๆ และตรวจสอบคำแรกและคำสุดท้ายของข้อความ จากนั้นจะลบคำที่จัดอยู่ในประเภท Stop words ออกไปเช่น “.” หรือ “!” หรือ “?” หรือ “”” เป็นต้น และทำการ Mapping คำเข้ากับค่า TFIDF หรือค่า Spamming Rate ตามขั้นตอนวิธีการกรองเพื่อจัดเป็นข้อความปรกติ หรือข้อความสแปม เมื่อได้ผลการกรองเป็นข้อความปรกติ ระบบจะทำการส่งข้อความไปยังผู้รับ หรือหากไม่ผ่านการกรองข้อความ ระบบจะคัดทิ้งข้อความนั้นโดยไม่ส่งต่อ

C. การคัดแบ่งกลุ่มข้อความที่นำเสนอ

การศึกษานี้ได้นำวิธีการกรองแบบเดิมมาปรับปรุงวิธีการเพื่อให้สามารถทำงานได้ตามที่คาดหวังโดยมีกระบวนการแสดงตามรูปที่ 2 โดยมีกระบวนการย่อยที่ปรับเปลี่ยนต่างๆดังนี้

- แต่เดิมนั้น [6] และ [7] จะไม่พิจารณาข้อมูลที่ได้จากเลขหมายหรืออักษรพิเศษ ซึ่งถูกลบทิ้งในกระบวนการ TN แต่การศึกษาวิจัยนี้จะตรวจสอบรูปแบบของกลุ่มคำเฉพาะเช่น เลขหมายหรือเบอร์พิเศษ อีเมล์ หมายเลขบัญชี นำมาใช้ประกอบการคัดแยกเพื่อให้การคัดแยกข้อความ SMS มีความถูกต้องมากขึ้น ด้วยเงื่อนไขของการมีเครื่องหมายหรือคำที่เป็นส่วนประกอบในข้อความ ซึ่งการตรวจสอบกลุ่มคำเฉพาะที่ใช้และหมายเลขพิเศษเหล่านี้ได้ช่วยลดระยะเวลาและเพิ่มประสิทธิภาพในการคัดแบ่งประเภทข้อความ ซึ่งตารางที่ I แสดงตัวอย่างของกลุ่มคำเฉพาะที่สามารถพบเห็นได้ในข้อความ SMS



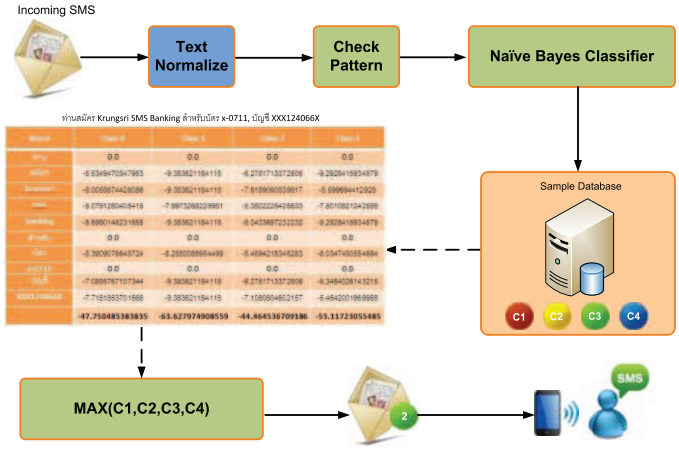
รูปที่ 2 : ขั้นตอนการทำงานของการแบ่งประเภทข้อความที่นำเสนอ

ตารางที่ I. รูปแบบของกลุ่มคำเฉพาะ

กลุ่มคำ	รูปแบบการตรวจสอบ
เลขหมายพิเศษ	- ตรวจคำว่า “โทร” หรือ “กด” และมีเครื่องหมาย # หรือ * แล้วตามด้วยชุดตัวเลข - ตรวจข้อความ “1900” แล้วตามด้วยชุดตัวเลขหลายหลัก - ตรวจคำว่า “โทร” หรือ “กด” และมีชุดตัวเลขแล้วตามด้วย เครื่องหมาย # หรือ เครื่องหมาย *
เบอร์โทรศัพท์	ตรวจคำว่า “โทร” หรือมีชุดตัวเลขที่*มีรูปแบบ*เข้าข่าย เช่น 02-1044828
อีเมล์	ตรวจกลุ่มตัวอักษรและกลุ่มตัวเลขที่มีรูปแบบเป็น E-mail Address เช่น knot_099505411@hotmail.com
หมายเลขบัญชี	ตรวจกลุ่มตัวอักษร X ที่มีชุดตัวเลขผสมอยู่ เช่น 2767XXXX
URL	ตรวจกลุ่มคำที่มีรูปแบบเป็น URL เช่น www.sanook.com
DateTime	ตรวจกลุ่มคำที่มีรูปแบบวันเวลา เช่น DD/MM/YYYY
User & Password	ตรวจคำว่า “username | password ” หรือมีชุดตัวเลขหรือตัวอักษร ไม่เกิน 4-8 หลัก

- ลดขั้นตอนการตรวจสอบคำแรกและคำสุดท้ายของข้อความ เนื่องจากการส่งข้อความ SMS ที่ต้องส่งข้อความ 2 ข้อความติดกันมีโอกาสเกิดน้อยมากเมื่อเทียบกับจำนวนข้อความทั้งหมด และด้วยข้อจำกัดของขนาดของข้อความทำให้ผู้ส่งส่วนใหญ่เลือกที่ส่งข้อความโดยมีความสมบูรณ์ของเนื้อหาอยู่ภายในข้อความเดียว
- การคัดแบ่งประเภทข้อความออกเป็น 4 ประเภทเพื่อให้เหมาะสมกับลักษณะการให้บริการข้อความ SMS ทั้งนี้โดยคำแนะนำจากผู้ให้บริการโทรศัพท์เคลื่อนที่

จากรูปที่ 2 เมื่อรับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message ให้อยู่ในรูปของ text แล้ว จะทำกระบวนการ TN เพื่อลบตัวอักษรที่ไม่สามารถตัดเป็นคำได้ออกไป และตรวจสอบรูปแบบของคำเฉพาะ จากนั้นจะลบคำที่จัดอยู่ในประเภท Stop words ออกไป และทำการแบ่งประเภทของข้อความตามอัลกอริทึมเพื่อสรุปผลของข้อความว่าจัดเป็นข้อความประเภทใดตามรูปที่ 3



รูปที่ 3 : การแบ่งประเภทของข้อความที่นำเสนอ

D. การคัดแบ่งประเภทของข้อความ

จากการเก็บข้อมูลตัวอย่างใน SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่ CDMA เป็นระยะเวลา 3 เดือน ตั้งแต่ กรกฎคม – กันยายน พ.ศ. 2553 พบว่าผู้ใช้บริการ SMS ในปริมาณมากนั้นมีหลายกลุ่ม โดยสามารถจำแนกกลุ่มของผู้ใช้บริการออกเป็น 4 ลักษณะ ดังนี้

1. ผู้ให้บริการต่างๆ เช่น ผู้ให้บริการโทรศัพท์มือถือ
2. ธุรกิจโฆษณาต่างๆ เช่น ดูดวง โหลดเพลง
3. องค์กรภาครัฐ เช่น โรงพยาบาล หน่วยงานของรัฐ
4. บุคคลทั่วไป

ซึ่งแต่ละกลุ่มผู้ใช้บริการจะมีวัตถุประสงค์ของการใช้บริการรับ-ส่งข้อความแตกต่างกันออกไปตามลักษณะของธุรกิจหรือบริการนั้นๆ ซึ่งสามารถแบ่งประเภทของข้อความออกเป็นประเภท 4 ได้ดังนี้

1. ข้อความโฆษณาประชาสัมพันธ์ (Advertising) เป็นข้อความเสนอขายสินค้าหรือบริการต่างๆ ทั้งนี้จัดเป็นประเภท C1
2. ข้อความทั่วไป (Normal) เป็นข้อความทั่วไปที่มีการส่งในชีวิตประจำวันและส่งโดยผู้ใช้งานทั่วไป ทั้งนี้จัดเป็นประเภท C2
3. ข้อความด้านการบริการ (Service) ส่วนใหญ่เป็นข้อความทางด้านการแพทย์,การศึกษา,การใช้งานระบบต่างๆ ทั้งนี้จัดเป็นประเภท C3
4. ข้อความแจ้งเตือนหรือข้อความทางการเงินที่ต้องรับรู้ (Alter or Finance) ที่อาจส่งผลต่อการดำเนินธุรกิจหรือการใช้ชีวิตประจำวันต่างๆ เช่น การเงิน แจ้งเตือนภัยต่างๆ ทั้งนี้จัดเป็นประเภท C4

ตารางที่ II แสดงตัวอย่างของข้อความ SMS ที่จัดอยู่ในประเภทต่างๆ ตามที่ได้กำหนดไว้ทั้ง 4 ประเภท

ตารางที่ II. ตัวอย่างข้อความประเภทต่างๆ

SMS	ประเภท
สิทธิ์พิเศษเฉพาะคุณ โหลดริงโทนฟรีถึง 31 มี.ค.53 กด #1572	โฆษณา (C1)
พักแล้วนะ!บาย!	ทั่วไป (C2)
2-5 missed call from 08646XXX@ 30/07/2010 17:44	บริการ (C3)
โอนเงินเข้า KBank 2767XXXX ผ่าน K-ATM 1,800 บ.	แจ้งเตือน (C4)

จากตารางที่ III. แสดงตัวอย่างการจัดแบ่งประเภทข้อความ ซึ่งจะเห็นว่าตารางการคำนวณนั้นได้ตัดคำที่ไม่มีผลต่อการคำนวณออกให้เหลือเฉพาะคำที่คาดว่าจะมีผลต่อการแยกประเภทข้อความเท่านั้น และคำที่ระบบได้ตัดออกไปจะเป็นคำที่พบว่าใช้บ่อยในข้อความทั่วไป เช่น คำว่า “ฉัน” “เธอ” “ท่าน” เป็นต้น ทั้งนี้ผลของการจัดแบ่งจะได้เป็นประเภท C3 หรือข้อความประเภทบริการ ซึ่งให้ค่าน้ำหนักของการคำนวณมากที่สุด

ตารางที่ III. ตัวอย่างข้อความ “ท่านสมัคร Krungsri SMS Banking สำหรับบัตร X-0711, บัญชี XXX124066X” ที่มีการคำนวณหาประเภทของข้อความ

Word	C1	C2	C3	C4
ท่าน	0.0	0.0	0.0	0.0
สมัคร	-5.5349	-9.3836	-6.2751	-9.2928
krungsri	-8.0058	-9.3836	-7.6189	-8.5996
sms	-5.0791	-7.9973	-5.3502	-7.5010
banking	-8.6990	-9.3836	-6.0433	-9.2928
สำหรับ	0.0	0.0	0.0	0.0
บัตร	-5.3909	-8.2850	-5.4594	-6.0347
x-0711	0.0	0.0	0.0	0.0
บัญชี	-7.0895	-9.3836	-6.2751	-6.3484
XXX124066X	-7.7181	-9.3836	-7.1080	-5.4642
SUMMARY	**-47.7504**	**-63.6279**	**-44.4645**	**-53.1172**

V. ผลการทดสอบ

การทดสอบกระทำกับชุดข้อมูลจำนวน 2 ชุด ได้แก่ ชุดข้อมูลสำหรับฝึกสอน จำนวน 5,198 ข้อความ โดยแบ่งเป็นข้อความปรกติ จำนวน 1,247 ข้อความ ข้อความบริการจำนวน 1,369 ข้อความ ข้อความแจ้งเตือนจำนวน 1,067 ข้อความ ข้อความโฆษณาประชาสัมพันธ์ จำนวน 1,515 ข้อความ และชุดข้อความ SMS ใหม่สำหรับทดสอบจำนวน 1,369 ข้อความ โดยแบ่งเป็นข้อความปรกติ จำนวน 359 ข้อความ ข้อความบริการจำนวน 328 ข้อความ ข้อความแจ้งเตือนจำนวน 256 ข้อความ ข้อความโฆษณาประชาสัมพันธ์ จำนวน 426 ข้อความ

อนึ่งข้อความประเภทโฆษณาประชาสัมพันธ์ C1 จะถูกกำหนดให้เป็นข้อความสแปมที่ใช้ในการทดสอบตามวิธีการคัดกรองแบบเดิม [6] เพื่อใช้ในการเปรียบเทียบประสิทธิภาพการทำงาน ทั้งนี้การทดสอบเปรียบเทียบได้กระทำบนอุปกรณ์ชุดเดียวกัน ใช้การเขียนโปรแกรมที่ใช้งานฟังก์ชันร่วมกัน และใช้ข้อมูลทดสอบเหมือนกันและกำหนดให้ประเภทของข้อความที่ใช้ทดสอบถูกคัดแยกโดยมนุษย์ไว้ก่อนหน้าแล้ว ซึ่งถือว่ามีความถูกต้องทั้งหมด

ผลการทดสอบแสดงให้เห็นตามตารางที่ IV และ V พบว่าการคัดแบ่งประเภทของข้อความด้วยวิธีการที่นำเสนอใช้เวลาในการแบ่งประเภทของข้อความน้อยกว่าเล็กน้อยกล่าวคือวิธีการที่นำเสนอใช้เวลาประมวลผลเฉลี่ยประมาณ 49.11 msec ในขณะที่วิธีการคัดกรองแบบเดิมนั้นใช้เวลาประมวลผลเฉลี่ย 52.10 msec ทั้งนี้เพราะการทำงานตามวิธีการที่เสนอใหม่ลดความซับซ้อนของขั้นตอนลงใน

ส่วนของกระบวนการการตรวจสอบคำแรกและคำสุดท้ายของข้อความ ซึ่งสามารถลดปริมาณงานที่ต้องทำงานได้ระดับหนึ่ง

ตารางที่ IV. ผลการทดสอบประสิทธิภาพทางเวลา

เปรียบเทียบเวลาที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	เวลาเฉลี่ยที่ใช้ในการประมวลผลต่อข้อความ (millisec)
การคัดกรองแบบเดิม	52.10
แบบที่นำเสนอ	49.11

ตารางที่ V. ผลการทดสอบประสิทธิภาพทางความถูกต้อง

เปรียบเทียบความถูกต้องที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	ความถูกต้อง (%)
การคัดกรองแบบเดิม	84.2951
แบบที่นำเสนอ	97.8816

อย่างไรก็ตามเพื่อพิจารณาถึงความถูกต้องพบว่าวิธีการที่นำเสนอสามารถคัดแบ่งประเภทข้อความได้ถูกต้องมากถึง 97.88% ในขณะที่วิธีการคัดกรองแบบเดิมมีความถูกต้องเพียง 84.29% ทั้งนี้เนื่องจากการเพิ่มการตรวจสอบรูปแบบของคำเฉพาะที่เป็นเลขหมาย หรือข้อความพิเศษต่างๆเข้าไป ซึ่งในการคัดกรองแบบเดิมไม่ได้นำมาใช้ในการประมวลผล ซึ่งกลุ่มเลขหมายหรือข้อความพิเศษดังกล่าวพบมากในข้อความประเภท C1, C3 และ C4 ทำให้เมื่อประมวลผลพบกลุ่มข้อความดังกล่าวจึงมีผลอย่างมากในการคัดแบ่งประเภทได้อย่างถูกต้อง

VI. สรุปผลการวิจัยและข้อเสนอแนะ

จากผลการทดสอบแสดงให้เห็นว่า วิธีการแบ่งประเภทข้อความแบบที่นำเสนอมีความแม่นยำมากกว่าวิธีการคัดกรองแบบเดิมและใช้เวลาในการประมวลผลน้อยกว่า ซึ่งเป็นผลมาจากการตัดขั้นตอนการตรวจสอบคำแรกและคำสุดท้ายสามารถลดเวลาในการทำงานลงได้เล็กน้อยประมาณ 6% เมื่อเทียบกับวิธีการคัดกรองแบบเดิม และการปรับปรุงวิธีตรวจสอบรูปแบบของคำเฉพาะ สามารถช่วยให้กำหนดกลุ่มของข้อความได้ง่ายขึ้น ลดคำผิดในฐานข้อมูลลงได้อย่างมาก ช่วยเพิ่มความถูกต้องให้กับวิธีการที่นำเสนอแบบใหม่ได้มากกว่าถึง 13.59% เมื่อเทียบกับวิธีการคัดกรองแบบเดิม นอกจากนี้ผลการศึกษานี้มุ่งเน้นให้เป็นพื้นฐานการแบ่งประเภทและจัดลำดับความสำคัญของข้อความ SMS ในประเทศไทยที่ไม่มีการคัดทิ้งหรือหยุดส่งข้อความ SMS ทำให้สามารถนำไปใช้งานจริงเชิงพาณิชย์ต่อไป

VII. กิตติกรรมประกาศ

คณะผู้วิจัยต้องขอขอบพระคุณ ฝ่ายพัฒนาผลิตภัณฑ์สื่อสารไร้สาย บริษัท กสท โทรคมนาคม จำกัด ที่ได้อนุเคราะห์ให้ข้อมูลการส่ง SMS ในครั้งนี้

VIII. เอกสารอ้างอิง

[1] *José María Gómez Hidalgo, Guillermo CajigasBringas and Enrique PuertasSánz "Content Based SMS Spam Filtering"* In Proceedings of the 2006 ACM Symposium on Document Engineering (Ams- terdam, The Netherlands, October 10 - 13, 2006)

[2] *Gordon V. Cormack, José María Gómez Hidalgo and Enrique Puertas Sánz "Spam Filtering for Short Messages"* Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.2007

[3] Bratko, Cormack, Filipl˘c, Lynam and Zupan, *"Spam Filtering Using Statistical Data Compression Model",* Journal of Machine Learning Research 7 (2006) ,2673-2698,2006

[4] *Siddharth Dixit, Sandeep Gupta and Chinya V. Ravishankar " LOHIT: AN ONLINE DETECTION & CONTROL SYSTEM FORCELLULAR SMS SPAMProceeding"* Proceedings of the IATED International Conference Communication, network and Information Security November 14-16, 2005 Phoenix, AZ USA – 2005

[5] *S. Liu and K. Cui, "Applications of Support Vector Machine Based on Boolean Kernel to Spam Filtering,"* Modern Applied Science, vol. 3, no. 10, October 2009.

[6] *นนท์ บุญนิธิประเสริฐ และ ชัยพร เขมะภาตะพันธ์ "Short Message Service Filtering for Thai & English Language on Mobile Phone Network"* The 7th International Conference on Computing and Information Technology (IC2IT2009)

[7] *Chaiyaporn Khemapatapan, "Thai-English Spam SMS Filtering"* Proceeding of the IEEE APPCC2010, Oct 30 – Nov 2, 2010 Auckland, New Zealand.

[8] *http://en.wikipedia.org/wiki/Naïve_bayes*

[9] *http://en.wikipedia.org/wiki/Bayesian_network*