การวิเคราะห์รูปแบบการกรองข้อความสั้นจากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่

ANALYSIS OF CONTENT-BASED AND HUMAN-INTERVENTION SMS SPAM FILTERING MODEL FOR MOBILE COMMUNICATION

อำนาจ ลมัยกลาง

AMNARD LAMAIKLANG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมโทรคมนาคม

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-EN-M-010-043

การวิเคราะห์รูปแบบการกรองข้อความสั้นจากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่

ANALYSIS OF CONTENT-BASED AND HUMAN-INTERVENTION SMS SPAM FILTERING MODEL FOR MOBILE COMMUNICATION

อำนาจ ละมัยกลาง

AMNARD LAMAIKLANG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมโทรคมนาคม

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-EN-M-010-043

ANALYSIS OF CONTENT-BASED AND HUMAN-INTERVENTION SMS SPAM FILTERING MODEL FOR MOBILE COMMUNICATION

AMNARD LAMAIKLANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR DEGREE OF

MASTER OF ENGINEERING IN TELECOMMUNICATIONS ENGINEERING

FACULTY OF ENGINEERING

KING MONGKUT'S INSTITUE OF TECHNOLOGY LADKRABANG

2013

KMITL-2013-EN-M-010-043

คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การวิเคราะห์รูปแบบการกรองข้อความสั้นจากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่

Thesis Title Analysis of Content- Based and Human -Intervention SMS Spam Filtering Model for Mobile Communication

นักศึกษา นายอำนาจ ละมัยกลาง

รหัสประจำตัว 51060925

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมโทรคมนาคม

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.สุวิพล สิทธิชีวภาค

หมายเลขวิทยานิพนธ์ KMITL-2013-EN-M-010-043

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.ยุทธพงษ์	รังสรรค์เสรี	
รศ.ดร.จีรสุดา	โกษียาภรณ์	
รศ.ดร.โกสินทร์	จำนงไทย	
รศ.ดร.พิพัฒน์	พรหมมี	
รศ.ดร.สุวิพล	สิทธิชีวภาค	

วัน / เดือน/ ปี ที่สอบ วันจันทร์ที่ 25 มีนาคม พ.ศ. 2556 เวลา 09.00-10.30 น.

สถานที่สอบ ณ อาคาร A ชั้น 5 ห้องประชุม 3

คณะวิศวกรรมศาสตร์ รับรองแล้ว

(ศาสตราจารย์ ดร.สุชัชวีร์ สุวรรณสวัสดิ์)

คณบดี คณะวิศวกรรมศาสตร์

วันที่ 25 มีนาคม พ.ศ. 2556

หัวข้อวิทยานิพนธ์	การวิเคราะห์รูปแบบการกรองข้อความสั้นจากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่
นักศึกษา	นายอำนาจ ละมัยกลาง
รหัสประจำตัว	51060925
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมโทรคมนาคม
พ.ศ.	2556
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.สุวิพล สิทธิชีวภาค

บทคัดย่อ

บริการข้อความสั้นหรือ Short Message Service (SMS) ได้รับความนิยมในการใช้งานเป็นอย่างมากในช่วงหลายปีที่ผ่านมาจึงทำให้จำนวนข้อความขยะเพิ่มขึ้นซึ่งส่งผลกระทบต่อประสิทธิภาพการทำงานของระบบศูนย์กลางบริการข้อความสั้นหรือ Short Message Service Center (SMSC) อย่างไรก็ดีเราสามารถควบคุมจำนวนข้อความขยะได้ด้วยระบบการกรองข้อความ วิทยานิพนธ์ฉบับนี้จึงนำเสนอรูปแบบการกรองข้อความสั้นที่ผสมผสานระหว่างการตรวจสอบเนื้อหาและการรับรองจากมนุษย์ (CAPTCHA) ของข้อความที่ไม่ได้ถูกจำแนกสถานะ หากไม่มีผลการตอบกลับจากมนุษย์แสดงว่าข้อความสั้นถูกส่งจากแหล่งที่สร้างข้อความขยะ ดังนั้นข้อความสั้นจึงไม่ถูกจัดส่งไปยังผู้รับปลายทาง เนื้อหาของวิทยานิพนธ์ฉบับนี้จะอธิบายรูปแบบและกระบวนการทำงานที่สามารถจำแนกข้อความขยะ ผลการวิเคราะห์ทำงานของรูปแบบดังกล่าวพบว่ามีความน่าจะเป็นที่จะคัดแยกถูกต้อง 0.9354 ในขณะที่รูปแบบที่ทำการตรวจสอบเนื้อหาเพียงอย่างเดียวมีความน่าจะเป็นที่จะคัดแยกได้ถูกต้อง 0.9022 ดังนั้นรูปแบบที่นำเสนอนี้จะส่งผลให้ SMSC สามารถทำงานได้อย่างเต็มประสิทธิภาพมากยิ่งขึ้น

Thesis	Analysis of Content-Based and Human-Intervention SMS Spam Filtering Model for Mobile Communication
Student	Mr.Amnard Lamaiklang
Student ID	51060925
Degree	Master of Engineering
Program	Telecommunications Engineering
Year	2013
Thesis Adviser	Assoc.Prof.Dr.Suvepon Sittichivapak

ABSTRACT

Short Message Service (SMS) has been the most popular means of mobile communication in recent years and hence the spam is an increasing threat to Short Message Service Center (SMSC) efficiency. The spam threat can be controlled through efficient and robust SMS filtering systems. In this paper, we present new model that is a combination of content-based (CB) filtering and human-intervention (CAPTCHA). A message, that has been classified as uncertain by CB filtering, is further checked by sending a challenge to the message sender. An automated spam generator is unlikely to send back a correct response, in which case, the message is classified as spam and don't deliver to recipients. Based on this formulation, results show that our framework achieved a higher accuracy of 0.9354 comparing to those of content-based filtering at 0.9022 consequently, promoted efficiency of SMSC operation.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ ด้วยการได้รับความกรุณาเป็นอย่างยิ่งจากท่านอาจารย์ รศ.ดร.สุวิพล สิทธิชีวภาค ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่คอยให้คำแนะนำในการค้นคว้าข้อมูล ตลอดจนเสียสละเวลาช่วยเหลือ และได้ให้ข้อคิดเห็นที่เป็นประโยชน์ต่องานวิจัย

ผู้วิจัยขอขอบพระคุณคณาจารย์ผู้ทรงคุณวุฒิ และเจ้าหน้าที่ของสถาบันฯ ที่เกี่ยวข้องทุกท่านที่ให้ความช่วยเหลือ ตลอดจนแนะนำกระบวนการในการศึกษา และการให้บริการต่างๆ ด้วยดีเสมอมา

ขอขอบพระคุณคณาจารย์ทุกๆ ท่านในสาขาวิชาวิศวกรรมโทรคมนาคม ที่ได้ถ่ายทอดความรู้ซึ่งเป็นประสบการณ์ใหม่ในหลายเนื้อหาของบทเรียนแก่ผู้วิจัยตลอดระยะเวลาการศึกษา

ผู้วิจัยขอขอบพระคุณท่านผู้บริการของบริษัท กสท โทรคมนาคม จำกัด (มหาชน) ที่ได้ให้โอกาสข้าพเจ้า ได้รับทุนการศึกษาและคัดเลือกเข้าศึกษาต่อในระดับปริญญาโทของสถาบันฯ นี้ คือ ท่านผู้ช่วยกรรมการผู้จัดการใหญ่ คุณสมยศ ธนพิรุณธร, ท่านผู้จัดการฝ่ายพัฒนาธุรกิจวงจรสื่อสารข้อมูล ดร.สมยศ อุดมโชคไพบูลย์, ท่านผู้ช่วยผู้จัดการฝ่ายพัฒนาธุรกิจวงจรสื่อสารข้อมูล คุณสมศักดิ์ พึ่งธรรมเกิดผล, ท่านผู้จัดการส่วนธุรกิจบริการเสริมอินเทอร์เน็ต คุณปิยะวุฒิ นวลประเสริฐ ที่ได้อนุเคราะห์ช่วยเหลือและส่งเสริมในการทำวิจัยครั้งนี้ และที่สำคัญคือคุณนันทวิตร ธนิตานันต์ และคุณนนท์ บุญนิธิประเสริฐที่เอื้อเฟื้อข้อมูลจนสำเร็จลุล่วงได้ด้วยดี

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณบิดา มารดา และน้องสาวที่เป็นที่รัก รวมถึงผู้มีอุปการะทุกท่าน (ญาติ – กัลยาณมิตร) ถึงแม้กลุ่มบุคคลดังกล่าวบางท่านจะไม่ได้จบการศึกษาในระดับปริญญาที่สูงส่ง แต่ด้วยแรงใจที่สูงส่งกว่าหวังให้ข้าพเจ้ามีการศึกษาที่สูงขึ้น เพื่อเป็นตัวอย่างแก่เยาวชนในชนบทให้เห็นถึงความสำคัญของการศึกษา จนสร้างแรงบันดาลใจให้ข้าพเจ้าตั้งใจศึกษาจนสำเร็จตามที่คาดหวังไว้

อำนาจ ละมัยกลาง

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

รูปที่ หน้า

สารบัญรูป(ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการติดต่อสื่อสารนับว่าเป็นส่วนหนึ่งที่สำคัญในชีวิตประจำวัน โดยเฉพาะการสื่อสารไร้สายที่ถูกพัฒนามาอย่างต่อเนื่อง เพื่ออำนวยความสะดวกสามารถตอบสนองต่อความต้องการได้ซึ่งหนึ่งในนั้นคือบริการส่งข้อความสั้นหรือ Short Message Service (SMS) รวมถึงบริการข้อความสื่อหรือ Multimedia Message Service (MMS) ได้รับความนิยมและเป็นที่แพร่หลายในการใช้บริการเป็นอย่างมากบนเครือข่ายโทรศัพท์เคลื่อนที่ (Mobile Communication) ในปัจจุบัน เช่น ใช้เป็นเครื่องมือในการสื่อสารการตลาด, เป็นอีกช่องทางในการตลาดแบบทางตรง, สร้างธุรกิจบริการเสริมต่างๆ หรือ Value Added Service (VAS) ทั้งนี้เนื่องจากข้อความสั้นมีข้อดีหลายประการเช่น เป็นสื่อที่มีประสิทธิภาพ, อัตราค่าใช้จ่ายต่อข้อความมีแนวโน้มลดลงอย่างต่อเนื่อง, สามารถกระตุ้นการรับรู้ได้เป็นอย่างดี ข้อความต่างๆที่ถูกส่งเข้ามาในบริการนี้อาจจะมีส่วนที่ถูกจัดกลุ่มเป็นข้อความขยะ (SMS Spam) ปะปนเข้ามาซึ่งจากการศึกษาในประเทศเกาหลีใต้และญี่ปุ่นนั้นพบว่ามีข้อความขยะสูงถึง 50% ของการใช้งานและผู้ใช้บริการในจีนได้รับ SMS ขยะ 8.29 sms ต่อสัปดาห์ [1] ทำให้รบกวนการใช้งานของผู้ใช้บริการโทรศัพท์เคลื่อนที่ รวมทั้งส่งผลกระทบต่อประสิทธิภาพของศูนย์กลางการรับส่งข้อความหรือ Short Message Service Center (SMSC) ที่ต้องรับภาระการทำงานเกินความจำเป็น ตัวอย่างเช่น การส่งข้อความของ ทรู คอร์ปอเรชั่นฯ (True Move) ที่มีการส่งเสริมการขายไปยังเครื่องโทรศัพท์ลูกข่ายของบริษัท โททอล แอคแซส คอมมิวนิเคชั่น (Dtac) ติดต่อกันจนทำให้เกิดการปิดรับ SMS ระหว่าง 2 ผู้ให้บริการเป็นเวลา 6 ชั่วโมง และกรณีการส่งข้อความอวยพรปีใหม่ของผู้ให้บริการในประเทศอินเดียที่อัตรามากกว่า 4 แสนข้อความต่อนาที จนทำให้ระบบหยุดทำงาน เป็นต้น หากสามารถลดปริมาณข้อความขยะออกจากระบบโทรศัพท์เคลื่อนที่ออกไปได้แล้วก็อาจจะเกิดผลดีในหลายส่วน เช่น ลดการทำงานคับคั่งของ SMSC ที่ไม่จำเป็น, SMSC ทำงานได้อย่างเต็มประสิทธิภาพ, สามารถตรวจสอบข้อความขยะก่อนถึงมือผู้ใช้บริการ Spam คือข้อความที่ไม่ได้เรียกร้องให้ส่งหรือ Unsolicited Message ที่ก่อให้เกิดความรำคาญแก่ผู้ใช้และอาจสร้างปัญหาการล่อลวงให้เสียทรัพย์สินทางโทรศัพท์มือถือ เช่น ข้อความโฆษณาขายสินค้า-บริการ การหลอกล่อให้ผู้รับทำกิจกรรมบางประเภทที่เกิดความเสียหาย ข้อความหลอกลวง (Phishing) เป็นต้น โดย Spammer จะใช้ Robot Software เข้ามาช่วยในการส่ง Spam SMS ที่มีลักษณะการส่งครั้งละหลายข้อความและหลายปลายทางในครั้งเดียว ปัจจุบันได้มีมาตรการป้องกันต่างๆ เช่น การลงทะเบียนไม่ขอรับข้อความโฆษณาจากผู้ให้บริการ การใช้ซอฟท์แวร์กรองที่เครื่องโทรศัพท์ การใช้ Software กรองที่ฝั่งเซิร์ฟเวอร์ เป็นต้น

การแก้ปัญหาด้วยซอฟท์แวร์การกรองข้อความขยะบนเครือข่ายโทรศัพท์เคลื่อนที่ที่ฝังเซิร์ฟเวอร์มีหลายวิธี เช่น Bogofilter, DMC, LR, SVM [2] ซึ่งล้วนแต่มีการพัฒนาต่อเนื่องจากพื้นฐานการกรองอีเมลขยะ (E-Mail Spam Filtering) คือ การตรวจสอบคำและการจำแนกข้อความว่าเป็นขยะหรือไม่ สำหรับงานวิจัยนี้ได้นำวิธีการ Naive Bayesian [3] ที่มีการตรวจจับคำหรือวลีสำคัญซึ่งเป็นส่วนหนึ่งของข้อความ นอกเหนือจากนี้ยังได้มีการนำแนวคิด Hybrid ที่ภายในมีกระบวนการ CAPTCHA (Completely Automatic Public Test to tell Computer and Humans Apart) [4] คือกลไกอัตโนมัติที่ใช้ทดสอบเพื่อให้ทราบว่าข้อความถูกส่งจากผู้ใช้หรือกลไก เนื่องจากส่วนใหญ่บรรดาข้อความขยะจะถูกส่งจากระบบคอมพิวเตอร์ครั้งละหลายๆ ข้อความในครั้งเดียว เพื่อตรวจสอบแหล่งที่มาของข้อมูลซึ่งมีการตรวจสอบระหว่าง SMSC และผู้ส่ง โดยวิทยานิพนธ์ฉบับนี้จะทำการวิเคราะห์เปรียบเทียบผลที่ดีกว่าของการใช้วิธีการกรองข้อความแบบผสมกับวิธีการกรองแบบตรวจสอบคำในเนื้อความเพียงอย่างเดียว

1.2 วัตถุประสงค์ของการศึกษา

1) เพื่อนำเสนอรูปแบบการกรองข้อความ (SMS Spam Filtering) ให้สามารถใช้งานกรองข้อความสั้นในระบบโทรศัพท์เคลื่อนที่ในประเทศไทย
2) เพื่อลดภาระการทำงานของระบบศูนย์กลางบริการข้อความสั้น (Short Message Service Center : SMSC) ที่ไม่จำเป็นจากการคัดแยกข้อความขยะออกไป

1.3 สมมุติฐานการศึกษา

1) สามารถคัดแยกข้อความขยะในระบบโทรศัพท์เคลื่อนที่จากการตรวจสอบเนื้อหาใน SMS นั้นได้
2) กระบวนการรับรองจากมนุษย์สามารถเพิ่มความถูกต้องในการคัดแยก SMS ได้

1.4 ขอบเขตของการศึกษา

ในวิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการกรองข้อความแบบผสม ซึ่งจะเป็นการหาค่าความถูกต้องของการกรองข้อความที่ตรวจสอบจากคำภายในข้อความที่เกิดจากการเรียนรู้ และนำมาประยุกต์เข้ากับการตรวจสอบผลการถามตอบ (Challenge-response) พร้อมกับการวิเคราะห์เปรียบเทียบผลที่เกิดขึ้นจากวิธีการที่นำเสนอโดยมีขอบเขตของการดำเนินงานวิจัยดังนี้

1) ใช้วิธีการเรียนรู้แบบ Naive Bayesian
2) ใช้วิธีการถามตอบ (Challenge-response) มาช่วยวิเคราะห์ผลการทำงาน

1.5 โครงร่างวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้ประกอบด้วย

บทที่ 1 บทนำ

บทที่ 2 การรับ-ส่งข้อความสั้นและข้อความขยะ

บทที่ 3 ทฤษฎีความน่าจะเป็นแบบเบย์และโปรโตคอลการถามตอบ

บทที่ 4 วิธีดำเนินการวิจัย

บทที่ 5 ผลการทดลองและผลการวิเคราะห์ข้อมูล

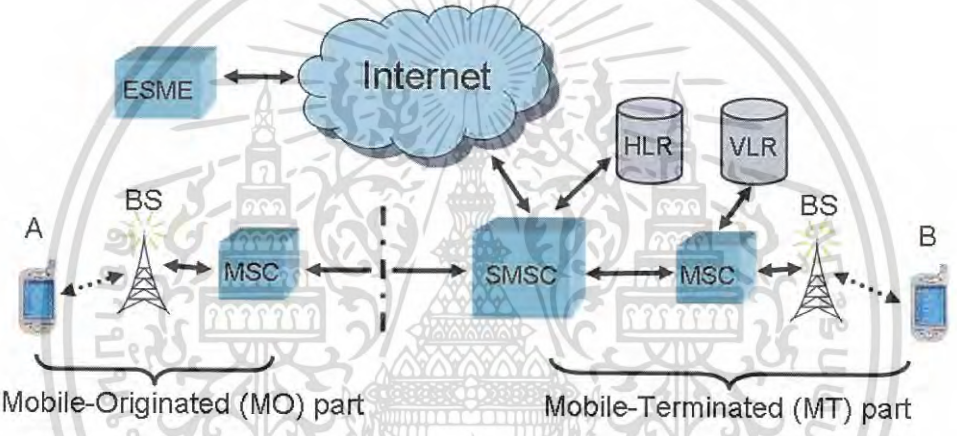
บทที่ 6 สรุปผล

บทที่ 2

ระบบการรับ-ส่งข้อความสั้นและข้อความขยะ

2.1 ระบบการรับ-ส่งข้อความสั้น

ข้อความสั้น หรือ Short Message Service (SMS) เป็นเทคโนโลยีการรับ-ส่งข้อมูลแบบเก็บและส่งต่อ (Store and Forward) ในเครือข่ายโทรศัพท์เคลื่อนที่ระบบ GSM เรียกอุปกรณ์ที่ใช้ในการเก็บและส่งต่อข้อมูลว่า Short Message Service Center (SMSC) โดยพื้นฐานการรับ-ส่งข้อความสั้นในเครือข่ายโทรศัพท์เคลื่อนที่ประกอบด้วย 2 ส่วนที่สำคัญดังแสดงในรูปที่ 2.1



รูปที่ 2.1 ระบบการรับ-ส่งข้อความสั้น

2.1.1 ผู้ส่ง (Mobile Originating:MO) ซึ่งรวมถึงตัวเครื่องโทรศัพท์มือถือ, สถานีฐาน (Base Station : BS), และชุมสายโทรศัพท์มือถือ (Mobile Switching Center : MSC) ที่ทำหน้าที่ค้นหาเส้นทางและเชื่อมต่อสัญญาณระหว่างฝั่งต้นทางและปลายทาง

2.1.2 ผู้รับ (Mobile Terminating:MT) ซึ่งรวมถึงตัวเครื่องโทรศัพท์มือถือ, สถานีฐาน (BS), ชุมสายโทรศัพท์มือถือ (MSC) ในฝั่งปลายทาง และส่วนที่สำคัญคือระบบกลางบริการข้อความสั้น (Short Message Service Center : SMSC) ที่ทำหน้าที่ควบคุมการรับ-ส่ง SMS ที่เข้ามาและ SMSC จะส่งสัญญาณร้องขอไปยัง Home Location Register (HLR) หรือ Visitor Location Register (VLR) ให้ระบุตำแหน่งเพื่อส่งต่อ SMS ไปยังผู้รับปลายทางที่ถูกต้องต่อไป

หลังจากนั้นก็จะส่ง SMS ไปในรูปแบบ Short Message Delivery Point-to-point ไปยังผู้ใช้ผ่านระบบของผู้ให้บริการ (Operator) ซึ่งจะทำการติดต่อไปยังเครื่องรับ หากเครื่องรับมีการต่อกลับมา SMS ก็จะถูกส่งไป จากนั้น SMSC จะได้รับการยืนยันตอบว่าข้อความที่ได้ส่งไปนั้นถึงปลายทางเรียบร้อยแล้วข้อความดังกล่าวก็จะมีสถานะเป็น sent ซึ่งจะไม่ถูกส่งอีก กรณีที่เครื่อง

โทรศัพท์เคลื่อนที่ของผู้รับปิด หรืออยู่นอกพื้นที่ให้บริการ SMS จะถูกเก็บไว้ใน SMSC ประมาณ 1 วัน (ระยะเวลาที่เก็บขึ้นอยู่กับผู้ให้บริการแต่ละราย) หากผู้ใช้บริการไม่เปิดเครื่องหรือกลับเข้าสู่พื้นที่ให้บริการ SMS ที่เก็บไว้ก็จะถูกลบออกจาก SMSC เพื่อป้องกันไม่ให้ SMS ที่เก็บไว้ล้น จนอาจส่งผลกระทบให้ระบบหยุดการทำงานได้

2.2 การส่ง SMS หลายปลายทาง

จากพื้นฐานที่กล่าวมานั้นเป็นลักษณะการรับ-ส่ง SMS ระหว่างโทรศัพท์มือถือด้วยกันเอง แต่ยังมีระบบ SMS ที่ยังสามารถรองรับการส่งอยู่อีกประเภทคือ External Short Message Entities (ESMEs) ที่เป็นการส่งจากหนึ่งผู้ส่งไปยังหลายผู้รับ (One-to-many message) ซึ่งได้ถูกนำมาใช้งานอย่างแพร่หลายในด้านการตลาดและบันเทิงเพราะคุ้มค่าจากการส่ง SMS ไปยังกลุ่มเป้าหมายได้ครั้งละจำนวนมาก เรียกการเชื่อมต่อแบบนี้ว่า Short Message Peer-to-Peer Protocal (SMPP Protocal) โดยสามารถใช้งานอย่างสะดวกผ่านทางระบบ Internet ได้

SMPP เป็น Protocol มาตรฐานในการส่งข้อมูล SMS MMS หรือ PUSH Message ภายในโครงข่ายโทรศัพท์เคลื่อนที่ โดยประกอบด้วยโครงสร้าง 2 ส่วน ดังแสดงในตารางที่ 2.1 คือ ส่วนที่ 1 PDU Header ที่ใช้ในการระบุ ความยาว ชนิด และลำดับของข้อความ ส่วนที่ 2 PDU Body ใช้บรรจุข้อมูลที่ต้องการส่งผ่าน เช่น ข้อความภายใน SMS หรือ Link สำหรับ PUSH Message ดังตัวอย่างที่แสดงในรูปที่ 2.2 และ 2.3 เป็นต้น

ตารางที่ 2.1 โครงสร้างของ SMPP PDU Format

PDU Header (mandatory)				PDU Body (Optional)
command length	*command id*	*command status*	*sequence number*	*PDU Body*
4 octets	Length = (Command Length value - 4) octets			

```
00 00 00 2F 00 00 00 02 00 00 00 00 00 00 00 01 53 4D 50 50 33 54 45 53 54 00
73 65 63 72 65 74 30 38 00 53 55 42 4D 49 54 31 00 00 01 01 00
```

รูปที่ 2.2 ตัวอย่าง Data-stream ของ SMPP PDU แบบ Hex format

00 00 00 2F Command Length 0x0000002F (ความยาวของ PDU)
00 00 00 02 Command ID 0x00000002 (คำสั่งสำหรับ bind transmitter)
00 00 00 00 Command Status 0x00000000 (สถานะของข้อความ)
00 00 00 01 Sequence Number 0x00000001 (ลำดับของ PDU Message)

รูปที่ 2.3 ตัวอย่าง Header ใน Data-stream ของ SMPP PDU

การส่งผ่านของ SMS ระหว่าง SMSC และโทรศัพท์มือถือสามารถทำได้เมื่อไหร่ก็ตามที่ใช้ Mobile Application Part (MAP) ของ SS7 Protocol ข้อความจะถูกส่งด้วยการทำงานของ MAP - MO และ MT-ForwardSM ซึ่งอัตราความยาวที่รับได้ จำกัดโดย signaling protocol ถึง 140 octets (140 octets = 140 x 8 bits = 1120 bits) โดยข้อความจะถูกเปลี่ยนเป็นรหัสโดยใช้ตัวอักษร จะขึ้นอยู่กับว่าตัวอักษรไหนที่สมาชิกได้กำหนดค่าในเครื่อง ซึ่งทำให้เกิดขนาดสูงสุดของแต่ละข้อความที่ 160 7-bit characters, 140 8-bit characters, และ 70 16-bit characters (รวมวรรค) การสนับสนุนของตัวอักษร GSM 7-bit นั้นบังคับกับเครื่องโทรศัพท์มือถือของ GSM และองค์ประกอบของเครือข่ายแต่ละอักขระในภาษาต่างๆ เช่น อารบิค จีน เกาหลี ญี่ปุ่น หรือ Cyrillic เป็นต้น โดยจะกำหนดอักขระเป็นรหัสโดยใช้ 16-bit UTF-16 ในการเปลี่ยนรหัส เส้นทางของข้อมูลและ meta data ข้อมูลอื่นๆนั้นถูกเพิ่มเติมลงไปในขนาดการโหลด

เนื้อหาที่ใหญ่กว่า (Concatenated SMS, multipart หรือ segmented SMS or "long SMS") สามารถถูกส่งโดยใช้หลายข้อความ โดยที่แต่ละข้อความจะมีข้อมูลด้านบนของผู้ใช้ user data header (UDH) ประกอบอยู่ด้วย เนื่องด้วย UDH ถูกบรรจุอยู่ใน payload จำนวนของอักขระจึงลดลงเหลือ : 153 for 7-bit encoding, 133 for 8-bit encoding และ 67 for 16-bit encoding โทรศัพท์มือถือที่ได้รับข้อความจะรวบรวมข้อความทั้งหมดและนำเสนอแก่ผู้ใช้โดยรวมเป็น 1 SMS ยาวๆ ขณะที่ทฤษฎีมาตรฐานอนุญาตให้สูงสุด 255 ส่วน และ ข้อความยาวๆส่วนมากจะถูกเรียกเก็บเงินเป็นหลายๆข้อความด้วยเช่นกัน

2.3 ข้อความขยะ (SMS Spam)

นิยามของข้อความขยะ [5]

ขยะหรือ Spam คือ "ข้อความที่ไม่ได้เรียกร้องให้ส่ง" (unsolicited message) ที่ก่อให้เกิดความรำคาญแก่ผู้ใช้และอาจสร้างปัญหาการล่อลวงให้เสียทรัพย์สินทางโทรศัพท์มือถือ ถูกส่งผ่านสื่ออิเล็กทรอนิกส์ในรูปแบบต่างๆ เช่น อีเมล (email), ข้อความสั้น (SMS), ข้อความสื่อ (MMS) หรือข้อความด่วน (Instant Message) เป็นต้น ในที่นี้จะใช้คำว่า SMS Spam เพื่อบ่งบอกว่าเป็น Spam ที่ส่งผ่านระบบ SMS ของระบบโทรศัพท์เคลื่อนที่ ซึ่งเป็นประเด็นที่สนใจในการจัดทำวิทยานิพนธ์ฉบับนี้ โดยสามารถยกตัวอย่าง SMS ที่เข้าข่ายเป็น SMS Spam ดังแสดงตัวอย่างในตารางที่ 2.2 ในระบบโทรศัพท์มือถือได้ดังนี้

- SMS โฆษณาขายสินค้าหรือบริการ
- SMS หรือ MMS ที่ล่อหลอก (Trick) ให้ผู้รับโทรศัพท์เข้าไปยังหมายเลขที่คิดราคาค่าโทรสูงกว่าปกติ เช่น SMS แจ้งให้โทรศัพท์กลับไปเพื่อยืนยันการสิทธิ์การรับรางวัล เป็นต้น
- ข้อความลวง (phishing) ที่หลอกล่อให้ผู้รับส่งข้อมูลส่วนบุคคลหรือข้อมูลทางธุรกรรมนำไปใช้ในทางที่ผิดกฎหมายต่อได้

ตารางที่ 2.2 ตัวอย่าง SMS Spam และแบบ SMS แบบปรกติ

SPAM	NORMAL
Get lots of xxx pictures in your e-mail!	Contact Me Privately (jackparkinson12@live.com)
Get free xxx account passwords! Press *4555	Hello, my friend!
ดาวน์โหลดริงโทน!! พิมพ์ ok ส่งที่ *123456	โหลดเอกสาร word ได้ที่ www.123.com/zip.zip
มันส์กับเกมEuro2008แล้วลุ้นPSP!	จองหนังให้แล้วนะ รีบมาด้วย

นอกจากนี้อาจจะยังมีกรณีที่ SMS Spam ที่ถูกส่งมาโจมตีระบบจาก Spammer โดยอาจจะใช้ Robot Software เข้ามาช่วยในการส่ง Spam SMS ที่มีลักษณะการส่งครั้งละหลายข้อความและหลายปลายทางในครั้งเดียวจากการลักลอบใช้งานซึ่งอาจจะส่งผลให้ SMS ค้างที่ SMSC จนไม่สามารถทำงานต่อไปได้ และในอุตสาหกรรม GSM ได้ระบุตัวเลขของการโกงที่เป็นไปได้บนมือถือ ซึ่งเกิดจากการละเมิดของบริการส่งข้อความภัยคุกคามที่รุนแรงที่สุดคือ SMS Spoofing เกิดขึ้นเมื่อผู้โกง

แจกจ่ายข้อมูลที่อยู่เพื่อปลอมเป็นผู้ใช้ที่ได้เปิดโรมมิ่งไปเครือข่ายต่างประเทศและส่งข้อความไปยังเครือข่าย บ่อยครั้งข้อความเหล่านี้ถูกส่งไปยังปลายทางนอกเครือข่าย ด้วย home SMSC นั้นเพียงพอที่จะถูกโจมตีเพื่อส่งข้อความไปยังเครือข่าย จึงก่อให้เกิดเป็นภาระค่าใช้บริการที่สูงมากเกินความจำเป็นด้วยเช่นกัน ทั้งนี้การป้องกัน SMS Spam ก็มีด้วยกันหลายวิธีการ เช่น การใช้ซอฟท์แวร์กรอง SMS Spam ที่ฝั่ง Server ของผู้ให้บริการ, การใช้ซอฟท์แวร์กรอง SMS Spam ที่เครื่องโทรศัพท์ของผู้ใช้บริการ, การสร้างข้อตกลงเบื้องต้นระหว่างผู้ใช้บริการและผู้ให้บริการ เป็นต้น

2.4 การจำแนกคำภาษาไทย (Thai word classifacation)

2.4.1 การตัดคำ (Word Segmentation) [6] การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้น มีปัญหาเบื้องต้นคือ การตัดคำในภาษาไทย ซึ่งลักษณะการเขียนภาษาไทยจะมีการเขียนติดต่อกันเป็นสายอักขระโดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ดังเช่นภาษาอังกฤษซึ่งใช้ช่องว่าง (Space) คั่นระหว่างคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งของการแบ่งสายอักขระไทยออกเป็นคำๆ จึงได้มีการพัฒนาวิธีการตัดคำแบ่งได้เป็น หลักการตัดคำโดยใช้กฎ (Rule Base Approach) หลักการตัดคำโดยใช้อัลกอริทึม (Algorithm Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) แต่ละวิธีการก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงานและปริมาณการใช้ทรัพยากรต่างๆ ที่แตกต่างกัน จากการศึกษาเรื่องตัดคำสำหรับการจัดหมวดหมู่เอกสารภาษาไทยพบปัญหาด้านการหาขอบเขตของคำ เนื่องจากไม่มีการเขียนแบ่งพยางค์ คำ หรือประโยค ไม่มีหลักเกณฑ์ตายตัวในการใช้ช่องว่างในภาษาเขียน การสะกดคำมีรูปแบบซับซ้อน มีคำยืม คำทับศัพท์ คำเฉพาะจำนวนมาก และคำมีความกำกวมสูง จากการศึกษาเปรียบเทียบประสิทธิภาพวิธีดังกล่าว พบว่าวิธีตัดคำที่เหมาะสมกับการจัดหมวดหมู่เอกสารคือวิธีการตัดคำแบบยาวที่สุด (Longest Matching) ซึ่งมีวิธีการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดแล้วทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้มีการย้อนรอย (Back Tracking) กลับไปเลือกพยางค์ที่ยาวรองมาแทนทำต่อไปเรื่อยๆจนสิ้นสุดสายอักขระ

2.4.2 การกำจัดคำหยุด (Stop-Word List Removal) เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลงคำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสารและปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุด

เป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล ตัวอย่างคำหยุดเช่น ที่ ใน ว่า และ จะ มี ได้ของ ให้ เป็นต้น

2.4.3 การหารากศัพท์ (Stemming) จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลงและเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ การหารากศัพท์ของคำภาษาไทยนั้นจะใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกัน หรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์ เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์

2.4.4 การสกัดคุณลักษณะ (Feature Extraction) วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะเอกสารคือการดึงคุณลักษณะ (Feature) ของเอกสารออกมา กับการลดขนาดเอกสารลง ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่าจะใช้อะไร เป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของเอกสาร และใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดียวแล้ว ยังสามารถใช้ วลี หรือกลุ่มของคำ ประโยค แทนคุณลักษณะของเอกสารได้เช่นกัน ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารประเภทข้อความคือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความจริง (Boolean) แทนด้วยค่าความถี่ของคำ (Word Frequency) หรือแทนด้วยค่าน้ำหนักของคำแบบอื่นๆ [1,2,5] ซึ่งวิทยานิพนธ์ฉบับนี้ใช้การเลือกคุณลักษณะแบบคำเดียว (Single word) ซึ่งได้จากการตัดคำโดยใช้พจนานุกรมเรียบร้อยแล้วผลลัพธ์ที่ได้จากการตัดคำจะได้เป็นคำเดียวจำนวนมากเพื่อมาใช้เป็นตัวแทนเอกสารในการเรียนรู้

บทที่ 3

ทฤษฎีความน่าจะเป็นแบบเบย์และโปรโตคอลการถามตอบ

รูปแบบการเรียนรู้สำหรับการจำแนกข้อมูล (Machine leaning model for classification) มีด้วยกันหลายวิธีการ เช่น Naive Bayesian, Decision Tree, Neural Network, Support vector Machine, Hybrid เป็นต้น และหนึ่งในวิธีการที่เป็นที่นิยมคือ Naive Bayesian ดังตัวอย่างงานวิจัยที่ถูกใช้ทั้งในและต่างประเทศ เช่น ในงานวิจัย [7] เป็นการนำไปใช้จำแนกประเภทไฟล์เอกสารของไทย ซึ่งพบว่าใช้ได้ได้ดีมีประสิทธิภาพในระดับหนึ่ง ในงานวิจัย [8] ถูกนำมาใช้ในการจำแนกข้อความแบบอัตโนมัติซึ่งพบว่า Naive Bayesian สามารถเรียนรู้เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูล ด้วยทฤษฎีความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) และกำหนดให้เงื่อนไขนั้นมีความเป็นอิสระต่อกัน (Conditional Independence) ของ Attribute หรือคุณลักษณะของกลุ่มข้อมูล โดยผลที่ได้พบว่าสามารถจำแนกข้อมูลได้และใช้เวลาในการเรียนรู้น้อยกว่าวิธีการอื่นๆ นั่นหมายถึงมีการใช้ระยะเวลาการทำงานน้อยที่สุด ในงานวิจัย [9] ได้นำมาจำแนกข้อความโดยพบว่าสามารถทำงานได้ดีสำหรับกลุ่มข้อมูลขนาดใหญ่ เนื่องจากเป็นการเรียนรู้แบบพื้นฐานซึ่งเป็นจุดเริ่มต้น และจะให้ผลการจำแนกได้แม่นยำดียิ่งขึ้นเมื่อนำไปใช้ร่วมกับเทคนิควิธีการอื่นๆ ในงานวิจัย [10] นำมาจำแนกประเภทของหนังสือในภาษาจีน ซึ่งก็เลือกใช้วิธีนี้เนื่องจากสามารถทำงานได้ง่ายและมีประสิทธิภาพ ในงานวิจัย [11] ได้กล่าวไว้ว่า Naive Bayesian เป็นเทคนิคการจำแนกที่ง่ายที่สุด เป็นที่แพร่หลายและถูกนำมาใช้เป็นระยะเวลาหนึ่งแล้ว ดังนั้นในวิทยานิพนธ์ฉบับนี้จึงเลือกใช้วิธีการของ Naive Bayesian สำหรับจำแนกข้อความขยะในระบบโทรศัพท์ของไทย และเนื่องจากไม่ค่อยมีงานวิจัยที่เกี่ยวกับการเรียนรู้และจำแนกข้อความในภาษาไทยซึ่งโครงสร้างของภาษาค่อนข้างมีความซับซ้อนกว่าภาษาอังกฤษมาก

การเรียนรู้แบบมีผู้สอน (supervised learning) เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลสอน (training data) ข้อมูลสอนประกอบด้วยวัตถุเข้า (มักจะเป็น เวกเตอร์) และผลที่ต้องการผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่อง (การถดถอย : regression) หรือ ใช้ทำนายประเภทของวัตถุ (การแบ่งประเภท : classification) ภารกิจของเครื่องเรียนรู้แบบมีผู้สอนคือการทำนายค่าของฟังก์ชันจากวัตถุเข้าที่ถูกต้องโดยใช้ตัวอย่างสอนจำนวนน้อย (คู่ของข้อมูลเข้าและผลที่เป็นเป้าหมาย : training examples) โดยเครื่องเรียนรู้จะต้องวางนัยทั่วไป (generalize) จากข้อมูลที่มีอยู่ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุผล

การแก้ปัญหาการเรียนรู้แบบมีผู้สอน (เช่น การเรียนรู้เพื่อรู้จำลายมือ) มีขั้นตอนต่าง ๆ ที่ต้องพิจารณา ได้แก่

1. กำหนดชนิดของตัวอย่างสอน ก่อนจะเริ่มทำอย่างอื่น จะต้องตัดสินว่าข้อมูลชนิดใดที่จะใช้เป็นตัวอย่าง เช่นในกรณีการรู้จักลายมือ ตัวอย่างอาจจะเป็นตัวอักษรตัวเดียว คำ หรือบรรทัด
2. เก็บตัวอย่าง ชุดตัวอย่างสอนจะต้องมีลักษณะเป็นตามที่ใช้จริง ดังนั้นชุดข้อมูลตัวอย่างและผลที่สอดคล้องจะต้องถูกจัดเก็บจากผู้เชี่ยวชาญหรือจากการวัด
3. กำหนดวิธีการแทนลักษณะ (feature) ของข้อมูลเข้า ความถูกต้องของฟังก์ชันจะขึ้นอยู่กับการแทนข้อมูลอย่างมาก โดยทั่วไปวัตถุเข้าจะถูกแปลงเป็นเวกเตอร์ของลักษณะ ใช้อธิบายวัตถุที่ต้องการแบ่งประเภท จำนวนลักษณะจะต้องไม่มากจนเกินไป เพราะจะทำให้เกิดปัญหา Curse of dimensionality เนื่องจากมิติที่กว้างเกินไปจนทำให้มีพื้นที่ว่างมากจนเครื่องเรียนรู้ไม่สามารถวางนัยทั่วไปได้ แต่จำนวนลักษณะก็จะต้องมากพอที่จะทำให้สามารถทำนายผลได้แม่นยำ
4. กำหนดโครงสร้างของฟังก์ชันที่ต้องการ และขั้นตอนวิธีการเรียนรู้ที่สอดคล้อง เช่น อาจจะต้องเลือกว่าจะใช้ ข่ายงานประสาทเทียม หรือ ต้นไม้ตัดสินใจ
5. ทำการออกแบบให้สมบูรณ์ แล้วใช้ขั้นตอนวิธีการเรียนรู้กับตัวอย่างที่เก็บมา อาจจะปรับพารามิเตอร์ต่างๆ ของขั้นตอนวิธีให้เหมาะที่สุดโดยใช้ชุดย่อยของชุดตัวอย่าง (เรียกว่า ชุดตรวจสอบ -- validation set) หรือ ใช้การตรวจสอบไขว้ (cross-validation) หลังจากปรับค่าต่างๆ แล้ว อาจจะวัดประสิทธิภาพของขั้นตอนวิธีโดยใช้ชุดทดสอบ (test set) ซึ่งแยกต่างหากจากชุดสอน

3.1 ทฤษฎีความน่าจะเป็นแบบเบย์ (Bayesian) [12]

ในทฤษฎีความน่าจะเป็น สถิติ การอนุมาน และปัญญาประดิษฐ์บางครั้งจะพบคำว่า แบบเบย์ (Bayesian) มาขยายชื่อทฤษฎีหรือโมเดลต่าง ๆ โดยทุกครั้งที่พบคำขยายนี้หมายความว่าได้มีการนำปรัชญาหรือหลักการของ ทฤษฎีความน่าจะเป็นแบบเบย์ (บางท่านเรียก การอนุมานแบบเบย์ หรือ สถิติแบบเบย์) มาใช้กับสาขาความรู้นั้น ๆ ถ้าจะกล่าวอย่างไม่เป็นทางการทฤษฎีความน่าจะเป็นแบบเบย์ หมายความถึง ความน่าจะเป็น ที่เป็นความเชื่อมั่นส่วนบุคคลในเหตุการณ์หนึ่ง ๆ ซึ่งต่างจากทฤษฎีความน่าจะเป็นของคอลโมโกรอฟ (ที่มักถูกเรียกว่าทฤษฎีความน่าจะเป็นเชิงความถี่) ที่มักแปลความหมายของความน่าจะเป็น (โดยต้องแปลควบคู่ไปกับการทดลองเสมอ) ตัวอย่างเช่น ความน่าจะเป็นของเหตุการณ์ A คือ อัตราส่วนของจำนวนครั้งของเหตุการณ์ A ที่ทดลองสำเร็จเทียบกับจำนวนครั้งที่ทดลองทั้งหมด จุดแตกต่างสำคัญระหว่างทฤษฎีทั้งสองประเภทมีดังนี้

3.1.1 ความหมายของความน่าจะเป็น โดยแนวคิดแบบเบย์มอง ความน่าจะเป็น เป็นความเชื่อส่วนบุคคลเชิงความถี่ มองความน่าจะเป็น เป็นคุณสมบัติหนึ่งที่ถูกฝังอยู่ในวัตถุ (ไม่ขึ้นกับตัวบุคคล)

3.1.2 การนำทฤษฎีไปใช้งาน ในการนำทฤษฎีความน่าจะเป็นเชิงความถี่ไปใช้จะต้องมีการทดลองเชิงแนวคิด (conceptual experiment) ควบคู่ไปด้วยเสมอ เหตุการณ์ใด ๆ ก็ตามที่ไม่มีการทดลองเชิงแนวคิดที่สมเหตุสมผลพอจะไม่สามารถนำทฤษฎีความน่าจะเป็นเชิงความถี่ไปใช้งานได้ เช่น ไม่สามารถจินตนาการการทดลองเพื่อทดสอบว่ามีมนุษย์ต่างดาวอยู่หรือไม่ได้ ฉะนั้นประโยคความน่าจะเป็นที่จะมีมนุษย์ต่างดาวไม่มีความหมายในทฤษฎีความน่าจะเป็นเชิงความถี่ แต่สามารถนำทฤษฎีความน่าจะเป็นแบบเบย์มาอ้างความน่าจะเป็นประเภทนี้ได้ในมุมมองนี้อาจกล่าวได้ว่าทฤษฎีความน่าจะเป็นแบบเบย์สามารถนาไปประยุกต์ใช้งานได้กว้างขวางมากกว่ากล่าวโดยสรุปทฤษฎีความน่าจะเป็นแบบเบย์ มีปรัชญาที่ต่างจากทฤษฎีความน่าจะเป็นเชิงความถี่เกือบสิ้นเชิงถึงแม้จะมีสัจพจน์พื้นฐานแบบเดียวกัน โดยในทฤษฎีความน่าจะเป็นแบบเบย์นั้นมองความน่าจะเป็นสถิติหรือการอนุมานเป็นเรื่องเดียวกัน

3.2 วิธีการเรียนรู้แบบเบย์ (Bayesian Learning)

การเรียนรู้แบบเบย์ เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็นซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำโมเดลมาหาว่าสมมุติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ความรู้ก่อนหน้า หมายถึง ความรู้ที่เรามีเกี่ยวกับสมมุติฐานแต่ละตัวก่อนที่เราจะเก็บข้อมูล เมื่อใช้งานเราจะนำความน่าจะเป็นของข้อมูลที่เก็บได้มาปรับสมมุติฐานซ้ำอีกครั้ง ข้อดีของวิธีการเรียนรู้แบบนี้คือเราสามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior Knowledge) เข้ามาช่วยนำการเรียนรู้ได้ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดีไม่ด้อยกว่าวิธีการเรียนรู้ประเภทอื่น

3.2.1 การแบ่งกลุ่มเบเซียนประเภทนาอีฟเบเซียน (Naive Bayesian Classifier)

เทคนิคการแบ่งกลุ่มแบบเบเซ่ยนอาศัยพื้นฐานจากกฎของบย์ในการเรียนรู้ประกอบด้วยอัลกอริทึมประเภทต่าง ๆ คือ นาอีฟเบเซียน กิ๊ป คลาสสิไฟเออร์ (Gibbs Classifier) เบย์ออฟติมอลคลาสสิไฟเออร์ (Bayes Optimal Classifier) และ เบเซียนบีลีฟเน็ตเวิร์ค (Bayesian Belief Network) ในการวิจัยนี้ได้เลือกใช้อัลกอริทึมประเภทเบย์อย่างง่าย หรือ นาอีฟเบเซียน ในการวิเคราะห์ข้อมูลผู้สำเร็จการศึกษา เนื่องจากเป็นอัลกอริทึมที่สามารถเรียนรู้และเข้าใจง่ายและยังสามารถเรียนรู้ได้รวดเร็ว

3.2.2 วิธีการเรียนรู้เบย์อย่างง่าย (Naive Bayesian Learning) [13]

การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพอีกวิธีหนึ่ง โดยที่ใช้งานได้ดีและเหมาะสมกับกรณีของเซตตัวอย่างมีจานวนมากและมี Attribute ของตัวอย่างไม่ขึ้นต่อกัน มีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnois) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการ

อื่น เนื่องจากเป็นวิธีการจำแนกข้อมูลที่มีประสิทธิภาพและมีอัลกอริทึมในการทำงานไม่ซับซ้อนเหมือนวิธีการอื่น

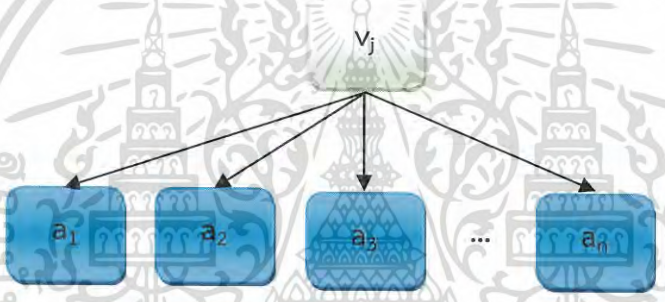
กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม v_j สาหรับข้อมูลที่มี Attribute ทั้งหมด n ตัว X = {a_1, a_2,..., a_n} ดังแสดงในรูปที่ 3.1 หรือ ใช้สัญลักษณ์ว่า P(a_1, a_2,..., a_n | v_j) คือ

$$P(a_1, a_2, a_3, ..., a_n \mid v_j) = \prod_{i=1}^{n} P(a_i, v_j) \quad (3.1)$$

โดยที่ $\prod$ หมายถึง ผลคูณของค่าทั้งหมด

i = 1, 2, 3,..., n

j = 1, 2, 3,..., n



รูปที่ 3.1 แสดงลักษณะของ Attribute แต่ละตัวที่เป็นอิสระต่อกัน

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้มีวิธีการดังนี้คือ

1.) หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า P(a_1, a_2,..., a_n | v_j) จากสมการ (3.1) มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้น ๆ คือ P(v_j) ได้เท่ากับ V_{NB}

2.) นำค่าที่ได้มาเปรียบเทียบกันกลุ่มที่มีค่าความน่าจะเป็นสูงสุดคือคำตอบดังนั้น เราจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่ายดังสมการ

$$V_{NB} = \arg\max_{v_j \in V} P(v_j) \times \prod_{i=1}^{n} P(a_i, v_j) \quad (3.2)$$

ตัวอย่างการใช้อัลกอริทึมการเรียนรู้เบย์อย่างง่าย โดยใช้ชุดตัวอย่างสอน (Training Set) ดังตารางต่อไปนี้

ตารางที่ 3.1 Training Set สำหรับการเรียนรู้เบย์อย่างง่าย

Class ↓

Attribute →

Name	Hair	Height	Weight	Lotion	Result
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	No	None
Annie	Blonde	Short	Average	No	Sunburned
Emily	red	Average	Heavy	No	Sunburned
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	Yes	None
Katie	Blonde	Short	Light	Yes	None

Value

สมมติว่าตัวอย่างที่ต้องการจำแนกประเภท คือ ผลลัพธ์ในตารางที่ 3.2

ตารางที่ 3.2 Test Data Set ที่ต้องการจำแนกประเภท

Name	Hair	Height	Weight	Lotion	Result
Judy	Blonde	Average	Heavy	No	?

ค่าความน่าจะเป็นของประเภท Sunburned และ none คำนวณได้จากสมการที่ (3.2) กรณีความน่าจะเป็นของประเภท Sunburned แทนด้วย V_{NB} = + จะได้ว่า

$$P(+)P(\text{Blonde}|+)P(\text{Average}|+)P(\text{Heavy}|+)P(\text{No}|+) = \frac{3}{8}\times\frac{2}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{3}{3} = \frac{1}{18} \qquad (3.3)$$

โดยที่
- P(+) หมายถึง ความน่าจะเป็นของประเภท Sunburned ซึ่งผลลัพธ์ที่ได้เป็นประเภท Sunburned จำนวน 3 คนจากทั้งหมด 8 คน
- P(blonde|+) หมายถึง ความน่าจะเป็นของคนที่มีผมสี blonde ในประเภท Sunburned ซึ่งมีจำนวน 2 คนจากทั้งหมด 3 คน
- P(average|+) หมายถึง ความน่าจะเป็นของคนที่มีความสูง average ในประเภท Sunburned ซึ่งมีจำนวน 2 คนจากทั้งหมด 3 คน

- P(heavy|+) หมายถึง ความน่าจะเป็นของคนที่มีน้ำหนัก heavy ในประเภท Sunburned ซึ่งมีจำนวน 1 คนจากทั้งหมด 3 คน
- P(no|+) หมายถึง ความน่าจะเป็นของคนที่ไม่มี Lotion ในประเภท Sunburned ซึ่งมีจำนวน 3 คนจากทั้งหมด 3 คน

กรณีความน่าจะเป็นของประเภท none แทนด้วย V_{NB} = - จะได้ว่า

$$P(-)P(\text{Blonde}|-)P(\text{Average}|-)P(\text{Heavy}|-)P(\text{No}|-) = \frac{5}{8}\times\frac{2}{5}\times\frac{1}{5}\times\frac{2}{5}\times\frac{2}{5} = \frac{1}{125}$$

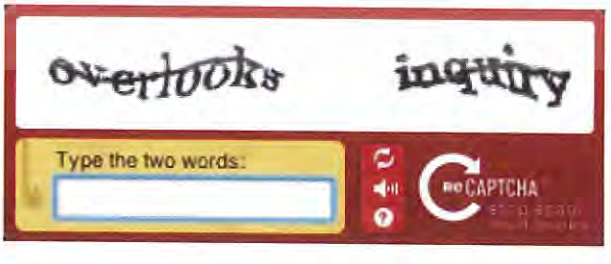
โดยที่
- P(-) หมายถึง ความน่าจะเป็นของประเภท none ซึ่งผลลัพธ์ที่ได้เป็นประเภท Sunburned จำนวน 5 คนจากทั้งหมด 8 คน
- P(blonde|-) หมายถึง ความน่าจะเป็นของคนที่มีผมสี blonde ในประเภท none ซึ่งมีจำนวน 2 คนจากทั้งหมด 5 คน
- P(average|-) หมายถึง ความน่าจะเป็นของคนที่มีความสูง average ในประเภท none ซึ่งมีจำนวน 1 คนจากทั้งหมด 5 คน
- P(heavy|-) หมายถึง ความน่าจะเป็นของคนที่มีน้ำหนัก heavy ในประเภท none ซึ่งมีจำนวน 2 คนจากทั้งหมด 5 คน
- P(no|-) หมายถึง ความน่าจะเป็นของคนที่ไม่มี Lotion ในประเภท none ซึ่งมีจำนวน2 คนจากทั้งหมด 5 คน

ดังนั้นคำตอบที่ได้คือ V_{NB} = + หมายถึง ประเภท Sunburned
เนื่องจากประเภท Sunburned มีค่าความน่าจะเป็นมากที่สุด

3.3 โปรโตคอลการถามตอบ (Challenge-response Protocal)

การถามตอบ (challenge-response) [4][14][17] เป็นส่วนหนึ่งของกระบวนการพิสูจน์ตัวตน (Authentication) โดย challenge คือคำถามที่ระบบจะถามมา (ซึ่งอาจจะไม่ซ้ำกันเลยในแต่ละครั้ง) ถ้าเรารู้ข้อมูลลับส่วนตัวของเรา เราก็จะสามารถตอบ challenge นั้นด้วย response ที่เกิดจากผลของchallenge กับรหัสลับของเรารวมกัน โดยที่ไม่ต้องส่งตัวรหัสลับออกไปทางเครือข่ายเลย ผู้ส่งที่ระบบไม่รู้จักจะถูกร้องถามให้พิสูจน์ยืนยันตัวเองก่อน โดยพิมพ์อักษรลอกตามภาพลวดลายที่ปรากฏบนหน้าจอ ซึ่งเรียกว่าเทคนิค CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) CAPTCHA ยืนอยู่บนแนวคิดที่มนุษย์ย่อมต้องมองเห็นและแปลความหมายของภาพลวดลายออก แต่คอมพิวเตอร์ที่ถูกตั้งโปรแกรมให้ส่ง Spam จะไม่สามารถมองออกและแปลความหมายได้

CAPTCHA ออกเสียงว่า แคปช่า ซึ่งย่อมาจาก Completely Automated Public Turing Computer and Humans Apart (การทดสอบของทัวริงสาธารณะแบบอัตโนมัติเพื่อแยกแยะว่าเป็นคอมพิวเตอร์กับมนุษย์อย่างสมบูรณ์) คือ กลไกอัตโนมัติที่ใช้ทดสอบเพื่อให้ทราบว่ามนุษย์ หรือ คอมพิวเตอร์ เป็นการทดสอบการตอบสนองโดยใช้ทดสอบกับระบบคอมพิวเตอร์ เพื่อตัดสินใจว่าผู้ใช้หรือผู้ที่กำลังติดต่อกับเว็บเซิร์ฟเวอร์เป็นมนุษย์หรือไม่ วัตถุประสงค์สำคัญก็เพื่อความปลอดภัยโดยเฉพาะเว็บไซต์ที่ต้องมีการป้อนข้อมูลส่วนตัว คิดค้นขึ้นในปี ค.ศ. 2000 โดย ลูอิส วอน อาห์น (Luis von Ahn) แมนูล บลัม (Manuel Blum) นิโคลัส เจ. ฮอปเปอร์ (Nicholas J. Hopper) และ จอห์น แลงฟอร์ด (John Langford) เนื่องจากแฮกเกอร์ส่วนใหญ่จะใช้สิ่งที่เรียกว่า **"บอตส์" (bots)** ในการโจมตีผู้ใช้ ซึ่งบอตที่ว่านี้สามารถสร้างขึ้นโดยคอมพิวเตอร์ แต่เนื่องจากคอมพิวเตอร์ไม่สามารถแก้ปัญหาการทดสอบด้วย CAPTCHA ได้ จะต้องอาศัยมนุษย์ที่เพ่งดูกราฟฟิกยุ่งเหยิงเหล่านี้และแกะตัวอักษรออกมาเพื่อพิมพ์ยืนยันอีกที วิธีการง่ายๆที่พบคือนำตัวอักษรมาแปลงให้เป็นรูปภาพแล้วถามผู้ใช้ว่าตัวอักษรในรูปภาพนั้นคืออะไร เพราะปกติมนุษย์จะอ่านตัวอักษรจากรูปภาพได้โดยไม่รู้สึกว่าต่างอะไรกับข้อมูลตัวอักษร (text) ทั่วๆไป แต่สำหรับคอมพิวเตอร์มันจะรู้แค่ว่านี่เป็นไฟล์ภาพเท่านั้น แต่ไม่รู่ว่าเป็นภาพอะไร เหตุที่ต้องมี CAPTCHA ก็เพื่อป้องกันผู้ใช้ที่เป็น bot นั่นเอง เช่น เว็บเมล์ของ google มีผู้ใช้งานมาก และบางคนก็อาศัยฟรีเมล์นี้เป็นแหล่งกระจาย Spam โดยทั่วไป google จะทำการระงับ account เหล่านี้ แต่คนกลุ่มนี้จะสมัครใหม่และวิธีที่จะไม่ให้เสียเวลาคือ การใช้งานผ่าน bot หรือโปรแกรมอัตโนมัติที่ช่วยสมัคร e-mail account ให้ ปัจจุบันมีคนพัฒนาโปรแกรมประเภท OCR เพื่อช่วยแปลงอักษรในภาพมาเป็นข้อมูลที่เป็นตัวอักษร (text) ซึ่งจริงๆ ถูกคิดค้นมาเพื่อใช้ประโยชน์อย่างอื่น เช่น มีหนังสือที่เป็นกระดาษก็เอามาผ่านโปรแกรม OCR เพื่อจะได้ข้อมูลที่เป็นตัวอักษร (text) ซึ่งสามารถนำไปใช้ในโปรแกรมประมวลผลคำ (word processor) ได้ อย่างถ้าเป็นนักศึกษาก็ scan หนังสือเป็นไฟล์ภาพแล้วนำมาผ่านโปรแกรม OCR ทำเป็นรายงานใน เวิร์ดฯได้เลย ดังนั้นจึงมีการพยายามป้องกันโปรแกรม OCR ให้ทำงานยากขึ้น เช่น ทำให้ตัวอักษรบิดเบี้ยว หรือใส่สิ่งรบกวนลงไป เช่น เส้น จุด หรือรูปต่างๆ ดังแสดงตัวอย่างในรูปที่ 3.2 เป็นต้น



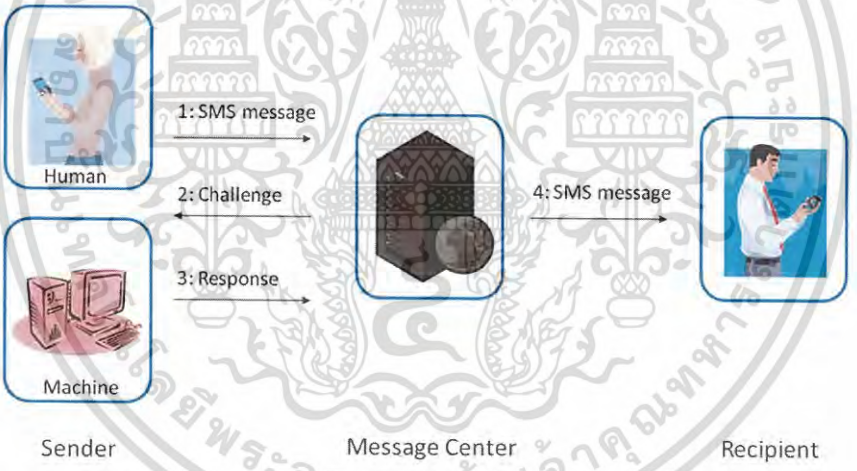
รูปที่ 3.2 ตัวอย่างของ CAPTCHA เพื่อใช้ในการพิสูจน์ตัวตน

บทที่ 4

วิธีดำเนินการวิจัย

4.1 แนวคิดรูปแบบการกรองข้อความจากเนื้อหาร่วมกับการรับรองจากมนุษย์

เพื่อลดภาระการทำงานของศูนย์กลางบริการข้อความสั้น (SMSC) ที่ไม่จำเป็นจากข้อความขยะ (SMS Spam) ให้สามารถทำการรับ-ส่ง SMS ของผู้ใช้บริการทั่วไปเป็นไปอย่างปกติไม่ถูกรบกวน การกรองข้อความจากเนื้อหา (Content-Based filtering : CB filtering) ที่อาจจะยังไม่สามารถคัดแยกหรือจำแนก SMS ได้อย่างถูกต้องทั้งหมดนั้น มาทำงานร่วมกับการรับรองจากมนุษย์ (Human Intervention) โดยวิธีการถามตอบ (Challenge-response) ซึ่งใช้กระบวนการ (Completely Automated Public Turing Computer and Humans : CAPTCHA) จึงเป็นอีกวิธีการที่นำเสนอ ซึ่งในวิทยานิพนธ์จะเรียกอีกอย่างว่า Hybrid เพื่อให้เกิดความน่าจะเป็นในการคัดแยก SMS ที่ถูกต้องมากยิ่งขึ้นดังแสดงในรูปที่ 4.1



รูปที่ 4.1 รูปแบบการกรองข้อความแบบ Hybrid

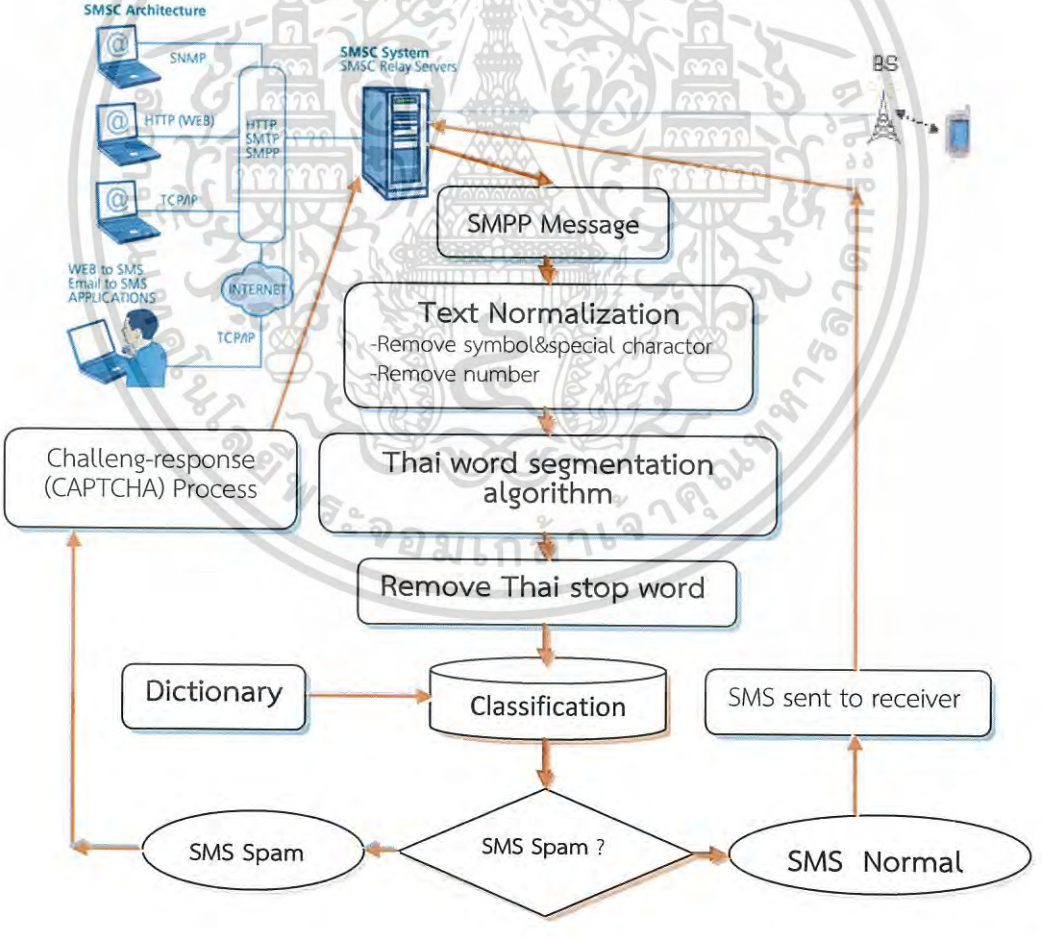
4.1.1 วิธีการกรองที่เหมาะสม

เนื่องจากยังไม่มีงานวิจัยที่ศึกษาการกรองข้อความจากบริการส่งข้อความ SMS ในประเทศไทยอย่างจริงจัง จึงต้องใช้การศึกษาวิธีกรองข้อความที่มีใช้งานในต่างประเทศเป็นพื้นฐานอ้างอิง โดยวิธีการที่ถูกใช้งานอย่างแพร่หลาย การจำแนกประเภทข้อความด้วยวิธีการตรวจสอบเนื้อหาหรือ Content-Based (CB) นิยมใช้ Naive Bayesian หลักการของวิธีการนี้ใช้การคำนวณความน่าจะเป็นซึ่งถูกใช้ใน การทำนายผล Naive-Bayesian เป็นเทคนิคในการแก้ปัญหาแบบ classification ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้โดย

จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้ในงานจัดหมวดหมู่เอกสารข้อความ (Text Classification), การวินิจฉัย (Diagnosis) ได้ดี อัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นพบว่าใช้งานได้ดี ซึ่งจะทำการศึกษาและปรับปรุงการทำงานบางส่วนให้สามารถใช้งานสอดคล้องกับภาษาไทยได้ เพื่อวิจัยเปรียบเทียบในด้านประสิทธิภาพความถูกต้องต่อไป

4.1.2 เนื้อหาของ SMS ในภาษาไทย

จากปัญหาของ SMS สำหรับภาษาไทยที่ค่อนข้างมีความซับซ้อนกว่าภาษาอังกฤษรายละเอียดดังแสดงในบทที่ 1 ของวิทยานิพนธ์ ซึ่งจะทราบหลักการทำงานและข้อบกพร่องในการกรองข้อความ โดยจะนำผลการทดสอบมาทำการวิเคราะห์เพื่อแก้ไขและปรับปรุงวิธีการกรองให้มีความสอดคล้องกับข้อความ SMS ในประเทศไทยดังแสดงในรูปที่ 4.2



รูปที่ 4.2 ขั้นตอนการกรองเนื้อความของ SMS สำหรับภาษาไทย

4.1.3 เครื่องมือที่ใช้ในงานวิจัย

การทดลองการทำงานของตัวกรองเนื้อหา (CB Filtering) ในวิทยานิพนธ์ฉบับนี้ได้จากการจำลองการทำงานจากโปรแกรมที่สร้างขึ้นมาและติดตั้งบนคอมพิวเตอร์แทนการติดตั้งที่ SMSC หรือ SMS Gateway โดยมีข้อมูลจำนวน 2 ชุดคือ ชุดข้อมูลฝึกสอน (Training Data:TD) ที่นำตัวอย่างข้อความขยะไปฝึกให้มีการเรียนรู้โดยชุดตัวอย่างคำที่เป็น SMS ขยะที่ได้มาจากผลการสำรวจของงานวิจัย [4] และชุดข้อมูลทดสอบชุดใหม่ (New Data:ND) ที่ผสมระหว่างภาษาไทยและอังกฤษซึ่งนำมาจากระบบบริการ CAT CDMA ของบริษัท กสท โทรคมนาคม จำกัด (มหาชน) ที่ให้บริการอยู่ในปัจจุบัน

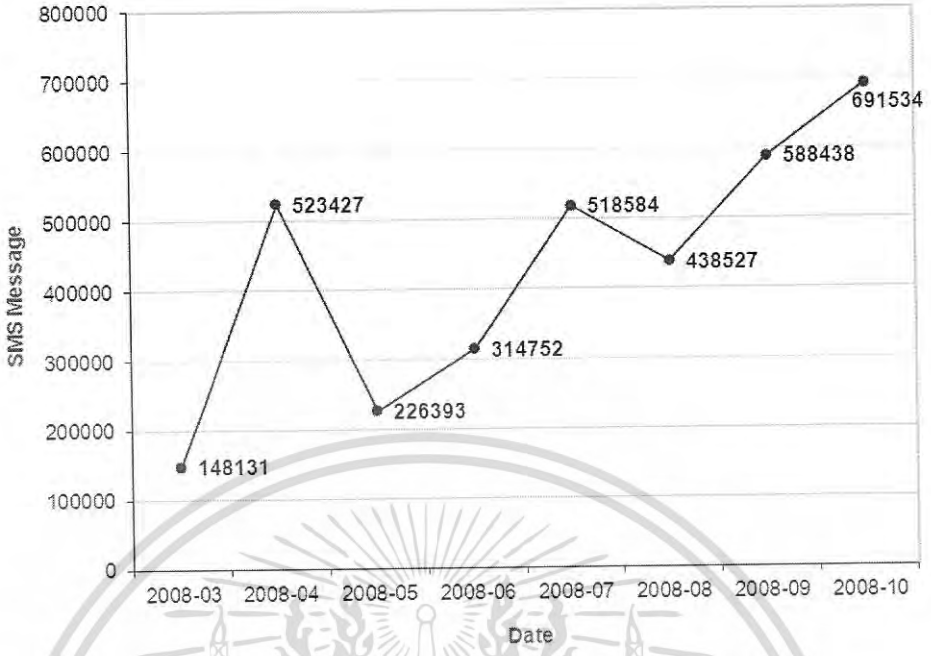
จากการคัดแยกข้อความขยะในจากตัวกรองเนื้อหา (CB Filtering) ซึ่งเป็นขั้นตอนแรกคัดแยกว่าเป็น spam หรือ ham โดยอาศัยค่า threshold แต่แทนที่จะมีค่า threshold เดียว ในวิทยานิพนธ์นี้จะใช้ threshold สองค่า คือ ค่า threshold บน และค่า threshold ล่าง ซึ่งจะแบ่งข้อความออกเป็น 3 กลุ่ม จากนั้นจะนำกลุ่มคำในช่วงพื้นที่สีเทามาวิเคราะห์จากการรับรองของมนุษย์ว่าเป็นข้อความที่ถูกส่งจากมนุษย์จริงหรือไม่

4.2 ขั้นตอนการเตรียมการดำเนินงานวิจัย

4.2.1 การรวบรวมและการวิเคราะห์ SMS ในประเทศไทย [15]

ก่อนทำการออกแบบตัวกรองข้อความสั้นนั้น จำเป็นต้องมีการวิเคราะห์พฤติกรรมของระบบ SMS เพื่อเป็นแนวทางในการพัฒนา โดยจะทำการรวบรวมข้อมูลการใช้งานบริการส่งข้อความ SMS จากระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยของบริษัท กสท โทรคมนาคม จำกัด (มหาชน) ชื่อบริการ CAT CDMA เป็นระยะเวลาประมาณ 8 เดือน เพื่อใช้ในการศึกษาเกี่ยวกับลักษณะและแนวโน้มทางสถิติของข้อความในระบบส่งข้อความ โดยมีวิธีดำเนินการดังนี้

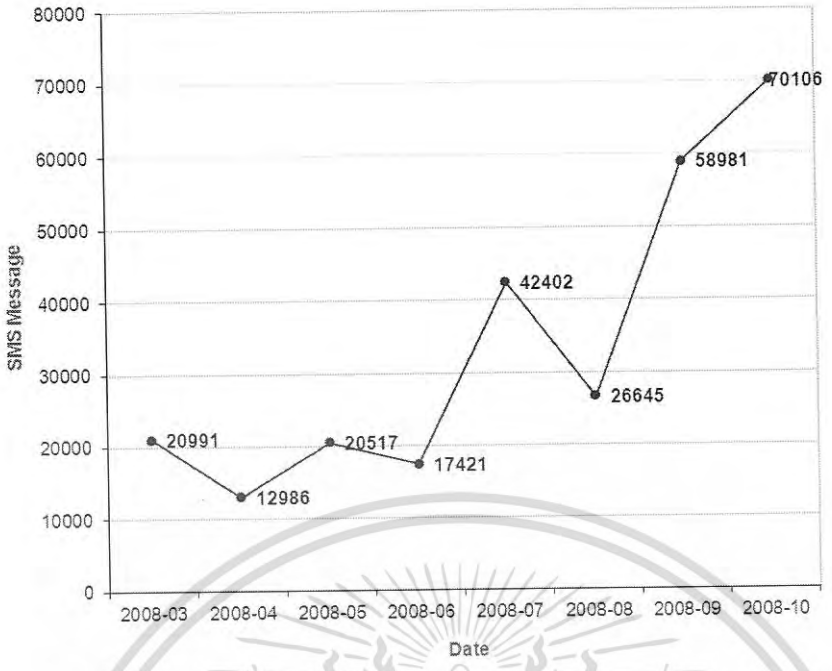
- ติดต่อผู้ให้บริการโทรศัพท์เคลื่อนที่ เพื่อขอความอนุเคราะห์ข้อมูลการ รับ – ส่งข้อความสั้น SMS เป็นระยะเวลา 8 เดือน
- ทำการเก็บรวบรวม Call Data Record (CDR) ของบริการ SMS ที่เครื่อง SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่
- นำเนื้อความของ SMS ที่ได้รับความอนุเคราะห์จากผู้ให้บริการ นำมาจัดเตรียมลงบันทึกสู่ฐานข้อมูล เพื่อใช้ในการเรียกดึงข้อมูลได้อย่างเป็นระบบและมีความรวดเร็ว
- นำสถิติข้อมูลการใช้งานมาแสดงในลักษณะเปรียบเทียบระยะเวลาและยอดการใช้งานดังรูปที่ 4.3



รูปที่ 4.3 ตัวอย่างข้อมูลยอดการส่ง SMS ของบริการ CAT CDMA

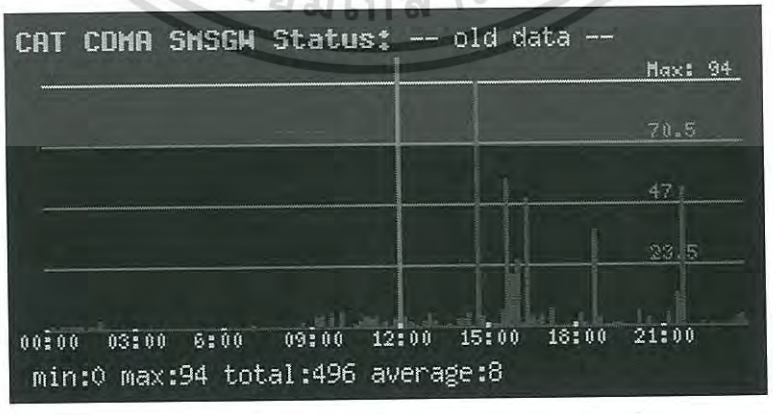
จากรูปที่ 4.3 แสดงให้เห็นแนวโน้มการส่ง SMS ที่มีการเพิ่มสูงขึ้นในแต่ละเดือน โดยเฉพาะสำหรับในเดือนที่มีเทศกาลพิเศษ เช่น เทศกาลสงกรานต์ จะพบว่ามีสถิติการใช้งานสูงกว่าการใช้งานในเดือนอื่นๆ เนื่องจากเป็นช่วงวันหยุดพักผ่อน ซึ่งผู้ใช้งานเดินทางไปท่องเที่ยวในต่างจังหวัด ทำให้มีการสื่อสารถึงบุคคลอื่นสูงกว่าปกติ

จากกลุ่มตัวอย่าง จำนวน SMS ที่มีการส่งในระบบบริการ CAT CDMA จะเป็นจำนวนที่มีการส่งจากทั้งโทรศัพท์เคลื่อนที่และส่งจากเว็บไซต์ซึ่งเป็นอีกช่องทางที่ให้บริการเสริม ซึ่งหากพิจารณาจำนวน SMS ที่แยกจากการส่งผ่านเว็บไซต์ไปที่ SMSC ที่เป็นการส่งจาก 1 ข้อความไปยังหลายปลายทาง ก็จะพบว่ามีจำนวนการใช้งานที่มีแนวโน้มเพิ่มขึ้นเรื่อยๆ เนื่องจากสามารถใช้งานได้ง่าย เช่น ส่งได้ครั้งละหลายๆ หมายเลขปลายทางพร้อมกัน สามารถบริหารจัดการเบอร์ปลายทางได้ สามารถเชื่อมต่อการส่ง SMS จากระบบฐานข้อมูลที่มีอยู่แล้ว สามารถส่งโดยผ่านช่องทางอินเทอร์เน็ตได้ เป็นต้น

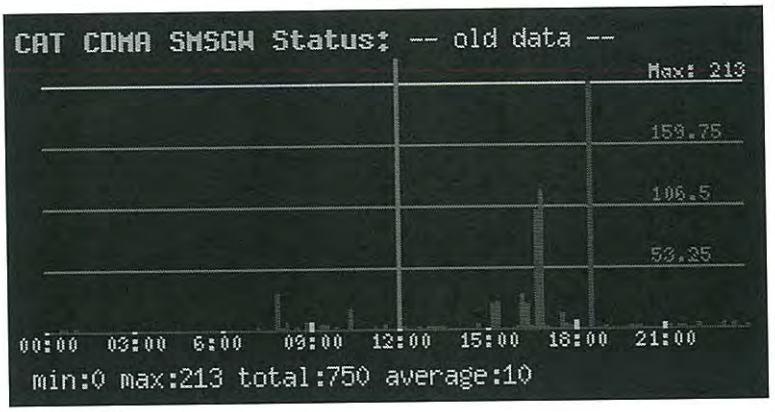


รูปที่ 4.4 ตัวอย่างข้อมูลการส่ง SMS ของบริการ CAT CDMA ผ่าน TCP/IP

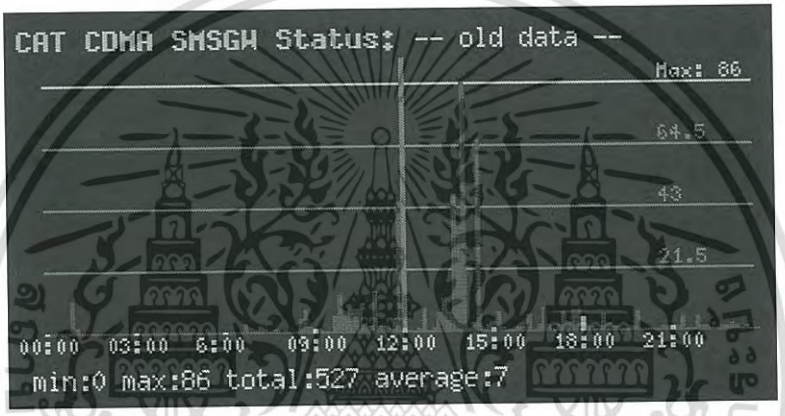
จากรูปที่ 4.4 แสดงให้เห็นแนวโน้มการใช้บริการส่งข้อความ SMS ผ่าน TCP/IP ที่มีการเพิ่มสูงขึ้นในแต่ละเดือนเช่นเดียวกับการส่งข้อความจากโทรศัพท์เคลื่อนที่ แต่สำหรับในช่วงเดือนที่มีเทศกาลพิเศษ เช่น เทศกาลสงกรานต์ จะมีสถิติการใช้งานต่ำกว่าการใช้งานในเดือนอื่นๆ ซึ่งสวนทางกับการส่งจากโทรศัพท์เคลื่อนที่ เนื่องจากเป็นช่วงวันหยุดพักผ่อน ซึ่งผู้ใช้งานเดินทางไปท่องเที่ยวในต่างจังหวัด ไม่ได้ใช้การส่ง SMS จากที่ทำงานทำให้การใช้งานบริการส่งข้อความผ่าน TCP/IP จากบุคคลทั่วไปลดลง หากพิจารณาข้อมูลจากระบบในเชิงความถี่ของข้อมูลที่ SMS Gateway สามารถสังเกตยอดการใช้งานได้ตามรูปที่ 4.5 – 4.8



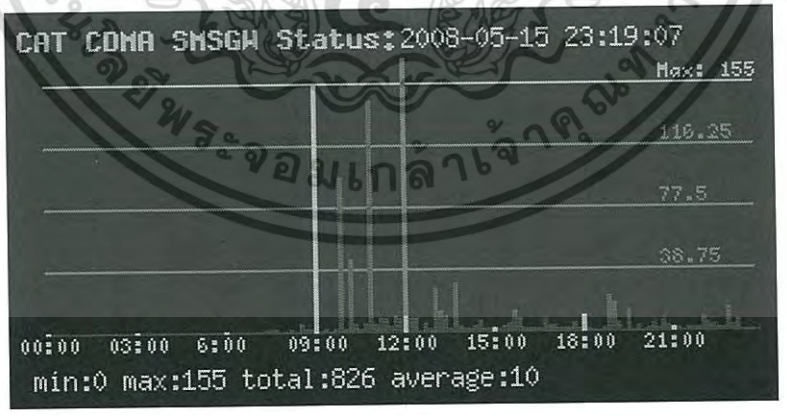
รูปที่ 4.5 ตัวอย่างข้อมูลการส่ง SMS ของบริการ CAT CDMA ผ่านเว็บของวันที่ 12/05/2008



รูปที่ 4.6 ตัวอย่างข้อมูลการส่ง SMS ของบริการ CAT CDMA ผ่านเวปของวันที่ 13/05/2008



รูปที่ 4.7 ตัวอย่างข้อมูลการส่ง SMS ของบริการ CAT CDMA ผ่านเว็บของวันที่ 14/05/2008



รูปที่ 4.8 ตัวอย่างข้อมูลการส่ง SMS ของบริการ CAT CDMA ผ่านเว็บของวันที่ 15/05/2008

รูปที่ 4.8 แสดงความถี่ในการส่งข้อความตามช่วงเวลาต่างๆ ของบริการ CAT4SMS ซึ่งมีความถี่ในการส่งหนาแน่นในช่วงเวลา 09:00 น. ถึง 16:00 น. ในแต่ละวัน ลักษณะของ

การส่งข้อความนั้นจะเป็นการส่งข้อในลักษณะส่งจาก 1 ข้อความ/ต้นทาง ไปยังผู้รับหลายปลายทาง

4.2.2 การนิยามข้อความขยะ

การกรองข้อความเพื่อแยก SMS ที่เป็น Spam ออกจากระบบโทรศัพท์เคลื่อนที่ได้นั้น จำเป็นต้องกำหนดความหมายของข้อความที่บ่งชี้ว่ามีแนวโน้ม Spam ให้มีความชัดเจน ขณะที่ผู้ใช้งานโทรศัพท์เคลื่อนที่แต่ละคนอาจมีทัศนคติในการตัดสินว่าข้อความใดเป็นข้อความ Spam หรือข้อความปรกติที่แตกต่างกัน ตัวอย่างความแตกต่างทางความคิดในการตัดสินข้อความ Spam เช่น ผู้ใช้งานโทรศัพท์เคลื่อนที่ที่มีรูปร่างอ้วนจะมีความสนใจในข้อความ SMS ที่เกี่ยวข้องกับการลดความอ้วน ได้แก่ ข้อความส่งเสริมการขายยา, อาหารเสริม, สถานออกกำลังกาย เป็นต้น เพื่อลดความอ้วน ในขณะที่ผู้ใช้งานโทรศัพท์เคลื่อนที่ที่มีลักษณะรูปร่างปกตินั้นกลับไม่ต้องการรับข้อมูลข่าวสารทาง SMS ดังกล่าว

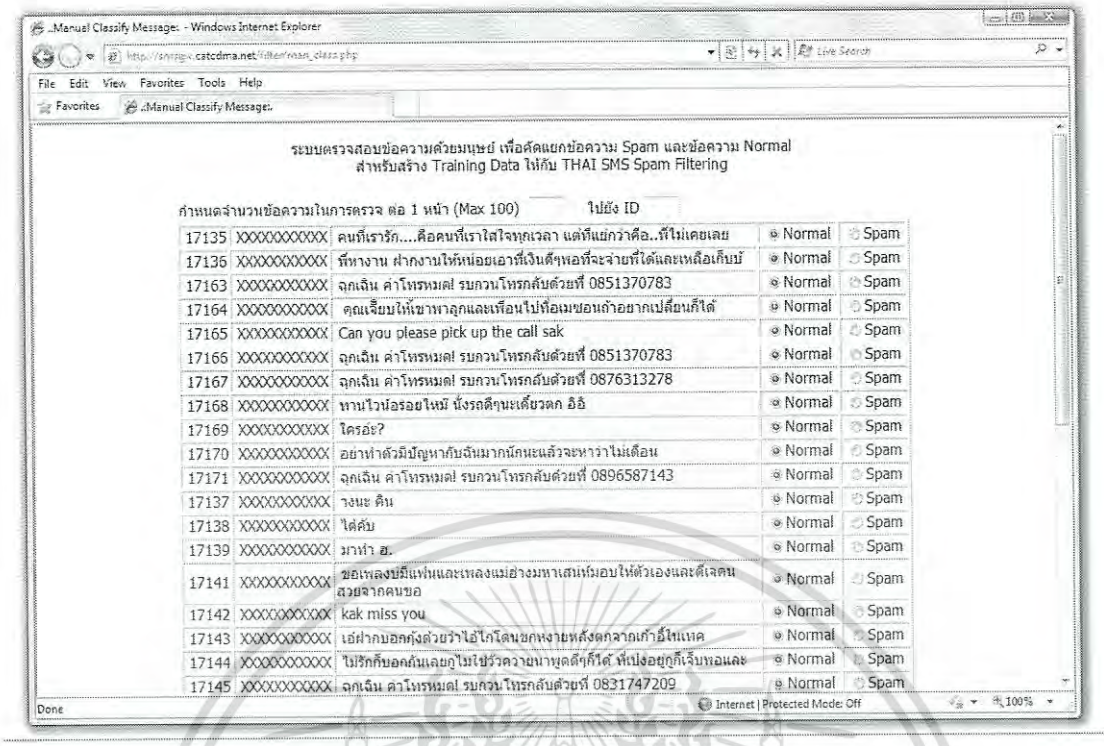
ความแตกต่างทางความคิดดังตัวอย่างข้างต้น ทำให้การตัดสินความหมายของข้อความ SMS เพื่อใช้เป็นข้อมูลฝึกสอนสำหรับออกแบบวิธีกรองข้อความ ไม่สามารถทำได้ถูกต้องทุกข้อความ การทำแบบสำรวจความคิดเห็น เพื่อใช้กำหนดความหมายของข้อความ Spam กับผู้ใช้งานโทรศัพท์เคลื่อนที่ ทำให้สามารถเข้าใจทัศนคติของกลุ่มผู้ใช้งานที่มีต่อข้อความ Spam ได้ชัดเจน โดยใช้แบบสำรวจปลายปิด เพื่อให้กลุ่มตัวอย่างสามารถแสดงความคิดเห็นได้อย่างสะดวกรวดเร็ว และจัดทำแบบสำรวจเป็น 2 ช่องทาง ได้แก่ แบบสำรวจที่จัดทำขึ้นเป็นเอกสาร จำนวน 2 หน้ากระดาษ A4 และแบบ Online ในรูปของ HTML เชื่อมต่อกับ ฐานข้อมูล Microsoft SQL และการนำข้อมูลดังกล่าวมาใช้กำหนดทิศทางของการออกแบบวิธีกรองข้อความ จะช่วยเพิ่มความถูกต้องในการกรองได้อีกด้วย โดยมีวิธีดำเนินการดังนี้

1. สร้างแบบสำรวจความคิดเห็นโดยให้ความสำคัญในการนิยามความหมายของคำว่า “ข้อความ Spam”

2. จัดทำแบบสำรวจในรูปสิ่งพิมพ์ และ HTML เพื่อเป็นช่องทางการรวบรวมข้อมูล

3. เก็บรวบรวมความคิดเห็น และจัดเตรียมข้อมูลลงสู่ฐานข้อมูลเพื่อใช้ในการสรุปผล

พัฒนาเครื่องในการคัดแยกข้อความ Spam แบบ Online Web Application เพื่อให้การคัดแยกข้อความด้วยมนุษย์จากข้อสรุปนิยามของ “ข้อความ Spam” ดังรูปที่ 4.9



รูปที่ 4.9 ระบบจำแนกข้อความด้วยมนุษย์ ผ่าน Web Application

จากรูปที่ 4.9 แสดงหน้าจอการทำงานของเครื่องมือการคัดแยกข้อความ Spam แบบ Online ผ่าน Web Application ซึ่งพัฒนาด้วยภาษา HTML และ PHP โดยเชื่อมต่อกับฐานข้อมูล Microsoft SQL ที่เก็บรวบรวมข้อมูล SMS

เพื่อหากลุ่มตัวอย่างของข้อความ Spam จึงได้มีการจัดทำแบบสำรวจความคิดเห็นของผู้ใช้งานโทรศัพท์เคลื่อนที่ในประเทศไทยจากกลุ่มตัวอย่าง เพื่อใช้ในการเรียนรู้ข้อความ Spam ซึ่งมีสาระสำคัญของแบบสอบถามดังนี้

1. ความหมายของข้อความ Spam ซึ่งจะนำมาใช้กำหนดชุดข้อมูลฝึกสอน เพื่อให้การกรองข้อความ SMS สามารถทำได้ตรงกับกลุ่มผู้ใช้งานในประเทศไทยมากที่สุด
2. ตัวอย่างคำและรูปแบบของข้อความที่พบได้ในข้อความ Spam เพื่อใช้กำหนดเงื่อนไขเพิ่มเติม จากระบบการกรองที่มีอยู่ในปัจจุบัน ให้มีความแม่นยำในการกรองสูงขึ้น
3. ผลกระทบและการแก้ไขปัญหาข้อความ Spam ใช้ในการอ้างอิงถึงความรุนแรงของปัญหาข้อความ Spam ที่เกิดขึ้นในประเทศไทย

การสำรวจความคิดเห็นจากกลุ่มตัวอย่างผู้ที่ใช้งานโทรศัพท์เคลื่อนที่จำนวน 468 ตัวอย่างแสดงดังตารางที่ 4.1 และตารางที่ 4.2

ตารางที่ 4.1 ผลสำรวจข้อมูลในส่วนของข้อมูลทั่วไปของกลุ่มตัวอย่าง

คำถาม/คำตอบ	จำนวนผู้ตอบ (คน)
<u>เพศ</u>	
ชาย	265
หญิง	203
<u>อายุ</u>	
ต่ำกว่า 20 ปี	56
21 – 30 ปี	307
31 – 40 ปี	75
40 ปีขึ้นไป	30
<u>สถานะ</u>	
โสด	393
สมรส	71
หม้ายหรือหย่า	4
<u>อาชีพ</u>	
นักเรียน / นักศึกษา	239
รับราชการ / รัฐวิสาหกิจ	139
พนักงานบริษัท / ธุรกิจส่วนตัว	90

<u>จำนวนโทรศัพท์เคลื่อนที่ที่ใช้งานพร้อมกัน</u>	
1 เครื่อง	294
2 เครื่อง	161
มากกว่า 2 เครื่อง	13
<u>จำนวนข้อความ SMS ที่ได้รับโดยเฉลี่ยใน 1 วัน</u>	
ไม่มี	39
น้อยกว่า 3 ข้อความ	276
น้อยกว่า 5 ข้อความ	112
มากกว่า 5 ข้อความ	41
<u>จำนวนข้อความ SMS ที่ส่งไปยังหมายเลขอื่นโดยเฉลี่ยใน 1 วัน</u>	
ไม่มี	183
น้อยกว่า 3 ข้อความ	236
น้อยกว่า 5 ข้อความ	37
มากกว่า 5 ข้อความ	12

ตารางที่ 4.2 ผลสำรวจข้อมูลในส่วนผลกระทบของผู้ที่ได้รับข้อความ Spam

คำถาม/คำตอบ	จำนวนผู้ตอบ (คน)
<u>ความหมายของข้อความ Spam ทาง SMS (ตอบได้มากกว่า 1 ข้อ)</u>	
ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ	361
ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่	275
ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้	153
ข้อความหยาบคาย หรือข้อความที่ไม่มีสาระสำคัญ	65
ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง	133
<u>ความถี่ของข้อความ Spam ที่ท่านได้รับในแต่ละวัน</u>	
ไม่มี	86
น้อยกว่า 3 ข้อความ	289
น้อยกว่า 5 ข้อความ	70
มากกว่า 5 ข้อความ	21
<u>ภาษาของข้อความ Spam ที่ท่านได้รับ</u>	
ภาษาอังกฤษเพียงอย่างเดียว	21
ภาษาไทยหรือภาษาไทยปนภาษาอังกฤษ	429
<u>คำใดบ้างที่ท่านคิดว่าจะพบในข้อความ Spam</u>	ดูตารางที่ 4.3
<u>ช่วงเวลาที่ท่านได้รับข้อความ Spam</u>	
06:01 น. – 20:00 น.	266
20:01 น. – 06:00 น.	37
ตลอดทั้งวัน ไม่แน่นอน	165

ผลกระทบของข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)	
ทำให้แบตเตอร์รี่หมดเร็วขึ้น	101
ก่อความรำคาญและทำให้ใช้งานไม่สะดวก	357
ถูกละเมิดสิทธิส่วนบุคคล	97
เสียค่าบริการจากโฆษณาในข้อความ Spam เพิ่มขึ้น	53
เสียพื้นที่ในการเก็บข้อความที่จำเป็น (Inbox เต็ม)	66
วิธีการแก้ปัญหาข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)	
ไม่ดำเนินการใดๆ	86
รับข้อความและเปิดอ่านตามปกติ	181
ลบข้อความทิ้งทันที	297
ปิดเครื่องโทรศัพท์ทันที	470
แจ้งผู้ให้บริการโทรศัพท์เคลื่อนที่เพื่อให้ดำเนินการแก้ไข	54
เลือกใช้โทรศัพท์เคลื่อนที่ที่สามารถกรองข้อความ Spam ได้	27

ตารางที่ 4.3 ข้อความ Spam ที่ได้จากผลสำรวจ

ลำดับ	คำ	ลำดับ	คำ	ลำดับ	คำ	ลำดับ	คำ
1	bonus	21	ซื้อ	41	ลักษณ์	61	เชิญชวน
2	download	22	ดวง	42	ลุ้น	62	เซียมซี
3	duty	23	ดารา	43	สนุก	63	เด็ด
4	free	24	ดาวน์	44	สมัคร	64	เติม
5	mail	25	ดาวน์โหลด	45	สลาก	65	เบอร์
6	mms	26	ดูดวง	4	สอบถาม	66	เพลง
7	promotion	27	ดูหมอ	47	สิทธิพิเศษ	67	เพิ่ม
8	push	28	ด่วน	48	สินค้า	68	เพิ่มเติม
9	ringtone	29	บริการ	49	สุขภาพ	69	เพียง
10	sms	30	พยากรณ์	50	ส่ง	70	เวลา
11	vote	31	พิเศษ	51	ส่งเสริม	71	เสียง
12	www	32	ฟรี	52	ส่วนลด	72	แม่น
13	xxx	33	ฟุตบอล	53	หาคู่	73	โฆษณา
14	กด	34	ราคา	54	ห้างสรรพสินค้า	74	โชคดี
15	ขาย	35	รางวัล	55	อ้วน	75	โทน
16	คลิป	36	รายการ	56	ฮิต	76	โทรศัพท์
17	คอร์ส	37	รายละเอียด	57	เกมส์	77	โบนัส
18	คูปอง	38	รูปภาพ	58	เครดิต	78	โหลด
19	ค่าบริการ	39	ร้านค้า	59	เงิน	79	ใหม่
20	ชิง	40	ลด	60	เงินพิเศษ		

จากผลสำรวจเพื่อหานิยามข้อความ Spam สำหรับเป็นข้อมูลในการเรียนรู้ สามารถสรุปโดยเรียงลำดับตามคะแนนจากผลสำรวจได้ดังนี้

1. ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ จำนวน 361 คะแนน
2. ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่ จำนวน 275 คะแนน
3. ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้ จำนวน 153 คะแนน
4. ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง จำนวน 133 คะแนน

และจากผลสำรวจสามารถสรุปผลกระทบจากการที่ได้รับข้อความ Spam โดยเรียงลำดับตามคะแนนได้ดังนี้

1. ก่อความรำคาญและทำให้ใช้งานไม่สะดวก จำนวน 357 คะแนน
2. ทำให้แบตเตอร์รี่หมดเร็วขึ้น จำนวน 101 คะแนน
3. ถูกละเมิดสิทธิส่วนบุคคล จำนวน 97 คะแนน
4. เสียพื้นที่ในการเก็บข้อความที่จำเป็น (Inbox เต็ม) จำนวน 66 คะแนน

จากการทำแบบสำรวจความคิดเห็นพบว่าผู้ใช้งานโทรศัพท์เคลื่อนที่ในประเทศไทยส่วนใหญ่ตัดสินข้อความในลักษณะโฆษณาขายสินค้า ข่าวสารทั่วไปที่ผู้ให้บริการโทรศัพท์เคลื่อนที่แจ้งเตือน ข้อความที่ไม่ทราบที่มา และข้อความที่มีความหมายใกล้เคียงกันที่ส่งหลายครั้งแสดงตัวอย่างดังตารางที่ 4.4 ซึ่งแตกต่างจากนิยามที่กำหนดไว้ในงานวิจัยที่มุ่งเน้นเฉพาะข้อความที่โฆษณาขายสินค้าหรือการชักชวนให้ใช้บริการพิเศษซึ่งมีอัตราค่าบริการสูงกว่าปรกติ และใช้ข้อมูลดังกล่าวอ้างอิงการคัดแยกข้อความที่รวบรวมจาก SMSC

ตารางที่ 4.4 ตัวอย่างของข้อความปกติกับข้อความ Spam ที่คัดแยกด้วยมนุษย์

ข้อความปรกติ	ข้อความ Spam
ถ้าไปทำได้ ตอนนี้อยากหายตัวไปนอนข้างๆ เธอในคืนนี้ เหงาจัง	2,350บเครื่อง+SimCATCDMAคุยทั้งวันทั้งปีทุกเครือข่าย149บ/ดที่mshopศก
มองโลกในแง่ดี ซิค๊ะ,	www.pec9.comเปิดคอร์สสอนสดติวเข้มชั้น ป.4-ม.6 เริ่มเรียน 1 มิ.ย. 51
สุขสันต์วันเกิด ขอให้มีความสุขมากๆนะ นึกอะไรก็ขอให้ได้ดั่งใจนึก	ลุ้นดูดวงฟรีกับปู โลกเบี้ยวแค่รับดวงรายวันกด*298*193#(4บ/วัน)022076888

ตารางที่ 4.4 ตัวอย่างของข้อความปกติกับข้อความ Spam ที่คัดแยกด้วยมนุษย์ (ต่อ)

ฉุกเฉิน ค่าโทรหมด! รบกวนโทรกลับด้วยที่ 084XXXXXXX	สุดคุ้ม!ใช้ฟรี15วัน โมโนลูกทุ่งฮิต แถมผลหวย โทร*45223111110/027302424
ไม่เคยกลัวคำขู่	ด่วนเซนต์เทเรซารับสมัครน.ศ.พยาบาลอีก10คน สุดท้ายโทร0867227493/037395313
เอากระโปรงผู้หญิงไปใส่ซะไป	ส่งความรักด้วยเสียงผ่านVoiceSMSฟรี ถึง31มี.ค. 51 โทร50100(เฉพาะภูมิภาค)
เธอเป็นคนเดียวที่ทำร้ายจิตใจฉันมากที่สุด ถ้าฉันตายเธอคงดีใจนะ	เชิญแข่งลดน้ำหนักชิงทองและครอสไม่จำกัดจำนวนยูนิเซ็นตปิ่นเกล้า028848692
ไม่คิดถึงเค้าแล้วหรอ?	สมัครสมาชิกทรูมูฟซูปเปอร์สตาร์ภายในวันที่ 20 พ.ค. นี้ รับฟรี!
การที่เรารักใครสักคนบางครั้งเราก็ไม่ต้องการให้เขารักเราตอบขอแค่ห่วงใย	ฟรีกระเป๋ามูลค่า350บ.ที่บูธTMBMoneyExpo 8-11พค.51_โชว์บัตร Ready Cash
มองในแง่ดีเสมอแต่พระเจ้าซิ!กำลังไม่ให้โอกาส!	ร่วมเล่นเกมส์ค้นหาสมบัติ พร้อลุ้นรับตุ๊กตา Mickey & Minnie Mouse Big S
ขี้เกียจทะเลาะกับผู้หญิงปัญญาอ่อน	ดูภาพมันส์ๆจาก100 Rock Uncensored BKK และลุ้นรับ CD พร้อมลายเซ็นต์จ
กำลังจะนอนค่ะ	ดวงคุณเดือนพฤษภาจะมีโชคลาภหรือไม่?ปูโลกเบี้ยวมีคำตอบกด*4988แม่นมากๆ
ไหงงั้น?	คุณมากกว่าใคร ดอกเบี้ย 0% 1 ปี ฟรีโอน+จดจำนอง โทร 1375
เค้าคนนั้นก็พอใจแล้วฝันดีนะคับหมูขาวของชัย	ด่วน!มีจำกัดโมบายดูทีวี-ใส่Sim2ระบบลูกเล่นครบเพียง8,900ที่M-Shopศก.
Sorry. Good night.	เจาะลึก!แม่นยำ!ดวงคุณสัปดาห์นี้จะเป็นอย่างไร อ.ลักษณ์ฟันธงโทร1900190065

4.3 การกรองข้อความจากเนื้อหา (CB Filtering)

จากบทที่ 3 การจำแนกข้อความ (Text classification) ใช้วิธีการตรวจสอบเนื้อหาหรือ Content-Based (CB) นิยมใช้เทคนิค Naive Bayesian ที่เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพ เหมาะสมกับกรณีของเซตตัวอย่างที่มีจำนวนมากและมี Attribute ของกลุ่มตัวอย่างที่ไม่ขึ้นต่อกัน (Conditional Independence) มีการนำไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) ซึ่งพบว่าใช้งานได้ดี โดยรายละเอียดการทำงาน Content-based filtering ในวิทยานิพนธ์ฉบับนี้สามารถอธิบายได้ดังสมการต่อไปนี้

กำหนดให้กลุ่มตัวอย่างสำหรับการทดสอบสอบ (Sample set) จำแนกออกเป็น 2 ประเภท คือ ข้อความปกติหรือในวิทยานิพนธ์ฉบับนี้เรียกว่า ham และข้อความความขยะหรือเรียกกว่า spam สามารถเขียนแทนด้วย $C=\{C_{ham},C_{spam}\}$

สำหรับคุณลักษณะ (Attribute) ของข้อความหรือ sms ที่มีลักษณะต่างๆ กันแทนด้วย $y_i=(W_1, \dots, W_n)$ ซึ่งคำที่อยู่ในข้อความนั้นๆ แทนด้วย W_n โดยข้อความนั้นเป็นส่งที่ต้องการจำแนก y_i ว่าเป็น ham หรือ spam ซึ่งสามารถเขียนแทนสมการรูปแบบความน่าจะเป็นของการจำแนกได้ดังสมการ 4.1

$$p(C \mid W_1,\dots,W_n) \tag{4.1}$$

เมื่อ C คือตัวแปรที่แทน Class ของการจำแนกข้อความ ซึ่งเป็นผลลัพธ์ของการทำงาน

$W_1, \dots, W_n$ แทนตัวแปรของคำที่อยู่ในข้อความ ซึ่งในหนึ่งข้อความจะประกอบด้วยหลายๆ คำมาเรียงต่อกัน

จากทฤษฎีของเบย์ สามารถเขียนสมการได้ดังนี้

$$p(C \mid W_1,\dots,W_n) = \frac{p(C)\cdot p(W_1,\dots,W_n \mid C)}{p(W_1,\dots,W_n)} \tag{4.2}$$

จากสมการที่ 4.2 เราพิจารณาเฉพาะในส่วนบนของสมการหรือเศษ ทั้งนี้เนื่องจากส่วนล่างหรือตัวส่วนนั้นเป็นตัวแปรที่ไม่ได้ขึ้นอยู่กับตัวแปร C และตัวแปร W_n มี Feature ที่กำหนดไว้แล้ว จึง

เป็นค่าคงที่ ดังนั้นเราสามารถประมาณค่าตัวเลขของสมการโดยใช้รูปแบบความน่าจะเป็นร่วม (joint probability model) ได้ดังสมการที่ 4.3

$$p(C,W_1,...,W_n) \tag{4.3}$$

จากสมการที่ 4.3 เมื่อใช้กฎลูกโซ่ (chain rule) มาใช้ในการเขียนสมการตามนิยามของความน่าจะเป็นแบบมีเงื่อนไข (Conditional probability) [10] ได้ดังสมการที่ 4.4 ดังนี้

$$\begin{aligned}
&p(C,W_1,...,W_n)\\
&\alpha \;\; p(C)\cdot p(W_1,...,W_n \mid C)\\
&\alpha \;\; p(C)\cdot p(W_1 \mid C)\cdot (W_2,...,W_n \mid C,W_1)\\
&\alpha \;\; p(C)\cdot p(W_1 \mid C)\cdot p(W_2 \mid C,W_1)\cdot (W_3,...,W_n \mid C,W_1,W_2)\\
&\alpha \;\; p(C)\cdot p(W_1 \mid C)\cdot p(W_2 \mid C,W_1)\cdot p(W_3 \mid C,W_1,W_2)\cdot (W_4,...,W_n \mid C,W_1,W_2,W_3)\\
&\alpha \;\; p(C)\cdot p(W_1 \mid C)\cdot p(W_2 \mid C,W_1)\cdot p(W_3 \mid C,W_1,W_2)\cdots (W_n \mid C,W_1,W_2,W_3,...,W_{n-1})
\end{aligned} \tag{4.4}$$

กำหนดให้ naive ซึ่งเป็นเงื่อนไขที่กำหนดให้เป็นอิสระต่อกัน (Conditional independence) โดยทุกๆคำหรือ W_n มี Feature ที่เป็นอิสระต่อกันอยู่แล้ว ดังนั้นเมื่อเราต้องการจำแนกให้เป็น Class จะสามารถเขียนได้ดังนี้

$$p(W_i \mid C,W_j) = p(W_i \mid C)\cdot p(W_i \mid C,W_j,W_k) = p(W_i \mid C)\cdot p(W_i \mid C,W_j,W_k,W_l) = p(W_i \mid C)$$

โดยที่ $i \neq j,k,l$ และจากรูปแบบร่วม (joint model) สามารถลดรูปสมการให้เหลือดังนี้

$$p(C,W_1,...,W_n) \;\; \alpha \;\; p(C)\cdot p(W_1 \mid C)\cdot p(W_2 \mid C),...,W_n \mid C) \tag{4.5}$$

$$p(C,W_1,...,W_n) \ \alpha \ p(C)\prod_{i=1}^{n} W_i \mid C \tag{4.6}$$

จากค่าความน่าจะเป็นของการจำแนกข้อความตามสมการที่ 4.6 เรานำมาประยุกต์ใช้สำหรับวิทยานิพนธ์ฉบับนี้เพื่อการจำแนกข้อความได้ดังสมการที่ 4.7 ดังนี้

$$p(C \mid y_i) = p(C)\prod_{i=1}^{n} (W_i \mid C) \tag{4.7}$$

โดยที่ $p(C \mid y_i)$ คือความน่าจะเป็นของการจำแนกข้อความที่ y_i ใดๆ ถ้ามีค่ามากกว่า 0.5 มีแนวโน้มเป็น ham และถ้าค่าน้อยกว่า 0.5 มีแนวโน้มเป็น spam

W_i คือคำซึ่งอยู่ในข้อความนั้นๆ ประกอบกันเป็นข้อความที่ใช้ส่ง sms ในระบบ

4.4 การพิจารณาพื้นที่สีเทา (Uncertain Region)

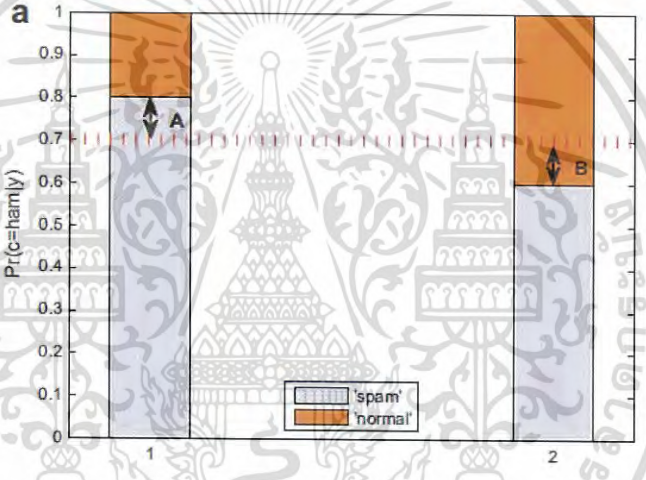
ปกติระบบการกรองข้อความจาก CB สามารถจำแนกผลการทำงานได้เป็น 2 แบบคือ ham และ spam หากกำหนดความน่าจะเป็นของการกรองแทนด้วยการกระจายตัว Pr(c=ham|y) แทนความน่าจะเป็นของข้อความที่อยู่ใน ham region โดย c และ y แทนตัวแปรสุ่มประเภทข้อความและข้อความตามลำดับ กำหนดอัตราส่วนที่ใช้ชี้วัดข้อความนั้นๆ ด้วย O_{post}= Pr(c=ham|y)/Pr(c=spam|y) ถ้า O_{post}>1 แสดงว่าข้อความถูกจัดให้อยู่ใน ham และกรณีอื่นข้อความจะถูกจัดให้อยู่ใน spam เมื่อพิจารณาจุดอ้างอิง (Threshold-base) เพิ่มเพื่อใช้เป็นจุดแบ่งแยก จากการที่ Pr(c=ham|y) มีค่าเข้าใกล้ 1 แล้วข้อความน่าจะถูกจัดอยู่ใน ham และหากมีค่าเข้าใกล้ 0 จะถูกจัดอยู่ใน spam กำหนด $\bar{c} = f(y,h)$ เป็นค่าการกรองของ CB เมื่อ $\bar{c}$ คือเอาท์พุต และ h คือจุดอ้างอิง ดังนั้นเมื่อแทนค่าตัวแปรต่างๆ แล้วตัวกรองสามารถทำงานได้โดยใช้สมการดังนี้

$$\bar{c} = f(y,h) = \begin{cases} ham & if \quad \Pr(c=ham|y) \geq h \\ spam & if \quad \Pr(c=ham|y < h \end{cases} \tag{4.8}$$

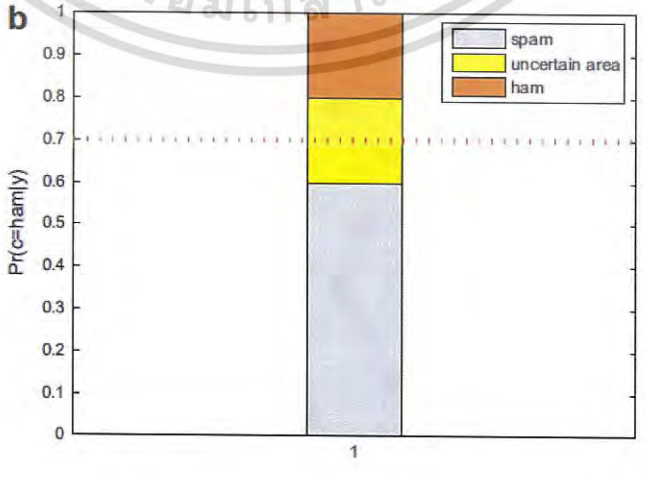
หากกำหนดจุดอ้างอิง (Threshold-base) หรือแทนด้วย h ที่ใช้แบ่งแยกการเป็นข้อความปกติและข้อความขยะที่มีค่า h=0.5 ดังแสดงในรูปที่ 4.10 ซึ่งอาจจะส่งผลทำให้เกิดปัญหาในการจำแนกประเภทข้อความได้เป็น 2 กรณีคือ

- หากกลุ่มข้อมูลทดสอบชุดใหม่ (New Data:ND) ที่เรานำมาทดสอบในครั้งนั้นๆ เป็นกลุ่มข้อมูลที่มีความเป็นปกติ (ham) มากจะทำให้เกิดการ Reject ของ sms ที่เกินจากความเป็นจริง นั่นหมายถึงข้อความที่ปกติจะไม่ถูกนำส่งไปยังผู้รับปลายทางได้ นั่นหมายถึงการทำงานที่ผิดวัตถุประสงค์

- และอีกกรณีหนึ่งคือการที่กลุ่มข้อมูลทดสอบชุดใหม่ (New Data:ND) ที่เรานำมาทดสอบในครั้งนั้นๆ เป็นกลุ่มข้อมูลที่มีความเป็นขยะ (spam) มากเกินไป ก็ย่อมจะส่งผลให้ระบบการกรองข้อความส่งผ่านข้อความขยะไปถึงผู้รับปลายทางได้



รูปที่ 4.10 กรณีที่ 1 h > $\tilde{h}$ และกรณีที่ 2 h < $\tilde{h}$ โดยกำหนดค่าอ้างอิงจริง $\tilde{h}$ แทนด้วยเส้นประ



รูปที่ 4.11 การเพิ่มพื้นที่สีเทาและกำหนดค่าอ้างอิงจริง $\tilde{h}$ แทนด้วยเส้นประ

จากทั้ง 2 กรณี เราสามารถนำวิธีการกรองข้อความไปใช้งานได้ในเบื้องต้น แต่ก็ยังมีจุดอ่อนของการกรองข้อความจากเนื้อหา (CB filtering) นี้ ดังนั้นหากจะนำวิธีการนี้ไปประยุกต์ใช้งานจริง จำเป็นต้องมีการพิจารณากำหนดจุดอ้างอิง (Threshold-base) ให้มีค่าที่เหมาะสม และอาจจะต้องมีการปรับปรุงค่าอ้างอิง เพื่อให้เหมาะสมกับรูปแบบของข้อความต่างๆ ที่มีการเปลี่ยนแปลงไปตามสถานการณ์ในปัจจุบัน

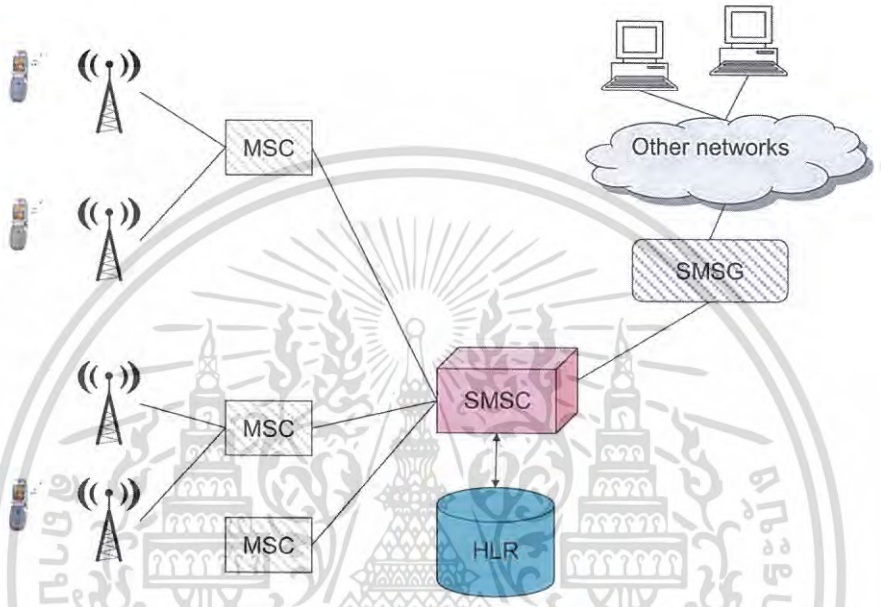
จากจุดอ่อนบางประการของการกรองข้อความจากเนื้อหานั้น จึงมีการพิจารณาจุดอ้างอิง (Threshold) ออกเป็น 2 ช่วงดังแสดงในรูปที่ 4.11 คือ ค่า Upper threshold (จุดอ้างอิงขอบบน) และค่า Lower threshold (จุดอ้างอิงขอบล่าง) ซึ่งจะทำให้สามารถแบ่งข้อความที่เข้ามาในระบบการกรองออกเป็น 3 กลุ่ม (จากเดิมที่แบ่งเป็นแค่ 2 กลุ่ม) คือ กลุ่มเข้าข่าย ham มาก ๆ กลุ่มเข้าข่าย spam มาก ๆ และกลุ่มสุดท้ายก็คือกลุ่มที่อยู่ระหว่าง Upper threshold และ Lower threshold ซึ่งในที่นี้จะถูกเรียกว่า กลุ่มข้อความในพื้นที่สีเทา (Uncertain Region) จากนั้นนำกลุ่มที่ตกอยู่ในพื้นที่สีเทา (Upper and Lower Boundaries) ที่มีโอกาสตัดสินผิดพลาดได้สูงด้วยวิธีการ content-based อย่างเดียวมาเข้าขั้นตอนกรองตามวัตถุประสงค์ของวิทยานิพนธ์นี้ เพราะฉะนั้นประสิทธิภาพของระบบเองจะขึ้นกับความเหมาะสมในการเลือกค่า threshold ทั้ง 2 ค่าเป็นหลัก และก็ขึ้นอยู่กับลักษณะของ spam ที่เกิดขึ้นในระบบระหว่างช่วงหนึ่ง ๆ ซึ่งอาจจะต้องปรับปรุงค่า threshold เป็นระยะตามรูปแบบของ spam ที่เปลี่ยนไป

4.5 การใช้โปรโตคอลการถามตอบ (Challenge-response Protocal)

การถามตอบ (challenge-response) เป็นส่วนหนึ่งของกระบวนการพิสูจน์ตัวตน (Authentication) โดย challenge คือคำถามที่ระบบจะถามมา (ซึ่งอาจจะไม่ซ้ำกันเลยในแต่ละครั้ง) ถ้าเรารู้ข้อมูลลับส่วนตัวของเรา เราก็จะสามารถตอบ challenge นั้นด้วย response ที่เกิดจากผลของ challenge กับรหัสลับของเรารวมกัน โดยที่ไม่ต้องส่งตัวรหัสลับออกไปทางเครือข่ายเลย ผู้ส่งที่ระบบไม่รู้จักจะถูกร้องถามให้พิสูจน์ยืนยันตัวเองก่อน โดยพิมพ์อักษรลอกตามภาพลวดลายที่ปรากฏบนหน้าจอ ซึ่งเรียกว่าเทคนิค CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) CAPTCHA ยืนอยู่บนแนวคิดที่มนุษย์ย่อมต้องมองเห็น และแปลความหมายของภาพลวดลายออก แต่คอมพิวเตอร์ที่ถูกตั้งโปรแกรมให้ส่ง spam จะไม่สามารถมองออกและแปลความหมายได้

วิธีการจำแนกประเภทข้อความที่อยู่ในพื้นที่สีเทาว่าตกอยู่ในช่วงบวกและลบ (False Positive and False Negative) นิยมใช้กระบวนการ CAPTCHA โดยตรวจสอบรูปแบบที่เข้ากันซึ่งผู้ใช้ที่เป็นมนุษย์จะสามารถยืนยันข้อความที่ส่งเข้าไปในระบบ SMSC ได้ ส่วนข้อความขยะจากคอมพิวเตอร์หรือ bot ที่อาจจะสร้าง SMS ขยะจำนวนมากได้แต่ไม่สามารถยืนยันตัวเองและไม่

สามารถตอบจากข้อความภาพเจือจางที่แสดงได้ ดังนั้นจึงกำหนดเมื่อมีการถามตอบที่ถูกต้องแสดงว่ามีความน่าจะเป็นสูงที่ SMS นั้นถูกส่งจากผู้ใช้ รูปแบบสื่อกลางของ CAPTCHA ที่สามารถปรับให้เหมาะสมกับการใช้งานได้ เช่น ภาพ เสียง หรือ ตัวอักษร เป็นต้น โดยจะเรียกวิธีการนี้ว่าการกรอง SMS ที่มีการผสมผสานระหว่าง CB filtering และกระบวนการ CAPTCHA ว่าแบบผสมหรือ Hybrid ดังแสดงในรูปที่ 4.12



รูปที่ 4.12 โครงสร้างการกรองข้อความแบบผสม (Hybrid)

โครงสร้างที่ออกแบบในวิทยานิพนธ์ฉบับนี้แสดงในรูปที่ 4.12 เป็นการจำลองการทำงานของการรับ-ส่ง SMS [16] โดยกำหนดการส่งผ่านในระบบการรับ-ส่งดังนี้

- ผู้ส่ง (S)
- ผู้รับ (R)
- ศูนย์กลางบริการข้อความ (MSC หรือ SMSG)
- ส่วนประกอบอื่นๆ เช่น SMSC, HLR เป็นต้น

โดยกำหนด $y^{h}_{\bar{c}=Type}$ สำหรับ $type \in \{ham, spam\}$ แทนด้วยข้อความที่ถูกกรองซึ่งอยู่ในขนาดของ h ดังนั้นปริมาณข้อความรวมคือ $N_{FilteringOnly} = |y^{h}_{\bar{c}=ham}| \times 6 + |y^{h}_{\bar{c}=spam}| \times 1$ เมื่อ $|.|$

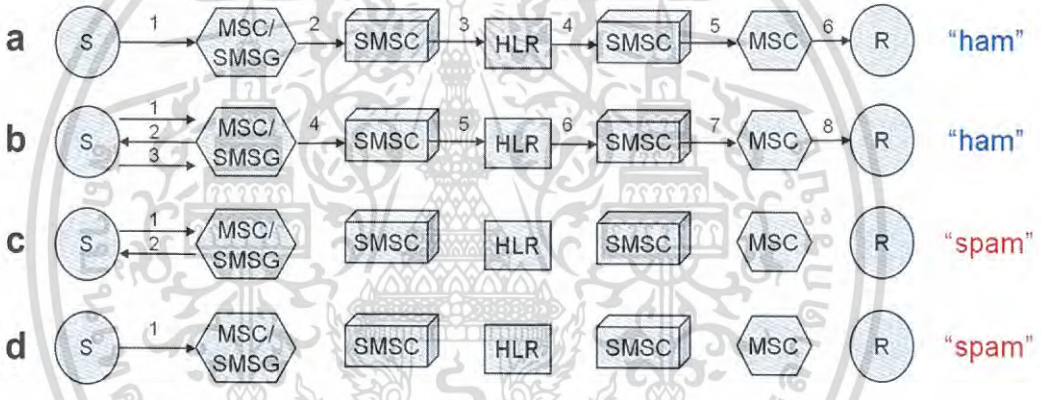
แทนจำนวนนับของ SMS ซึ่งปริมาณทั้งหมดในระบบ เนื่องจาก ham สามารถส่งผ่านไปยังโครงสร้างได้ 6 ส่วน คือ

S --> MSC/SMSG --> SMSC --> HLR --> SMSC --> MSC --> R

และ spam ส่งผ่านได้เพียง 1 ส่วนประกอบเท่านั้นคือ

S --> MSC/SMSG

สำหรับการกรองข้อความรูปแบบผสม (Hybrid) ที่ SMS จะถูกแยกออกเป็น 3 ช่วงโดยใช้ค่าอ้างอิง (threshold) จำนวน 2 ค่า คือ h_1 และ h_2 ดังนั้นจึงสามารถประมาณค่าโดยการนำพารามิเตอร์เพิ่มขึ้นมาพิจารณาเพิ่มเติมอีก คือ N_{un} และ N_{us} ที่อยู่ในพื้นที่สีเทา ซึ่งสามารถแสดงความเป็นไปได้ของเส้นทางการส่งผ่านข้อมูลดังรูปที่ 4.13



รูปที่ 4.13 ความเป็นไปได้ของการส่ง SMS ในรูปแบบ Hybrid ทั้ง 4 กรณี

กรณีที่ 1 (a) SMS ถูกจำแนกเป็น ham โดยระบบจะให้ค่าความน่าจะเป็นสูงกว่าค่าอ้างอิงขอบบน (Upper threshold) และจะถูกส่งไปยังผู้รับปลายทาง ผ่านอุปกรณ์ในระบบโทรศัพท์เคลื่อนที่ต่างๆ มีจำนวนการส่งผ่าน 6 ส่วน คือ

S --> MSC/SMSG --> SMSC --> HLR --> SMSC --> MSC --> R

กรณีที่ 2 (b) SMS ถูกจำแนกอยู่ระหว่างขอบบนและขอบล่าง (Upper – Lower boundary) ในพื้นที่สีเทา (Uncertain region) ซึ่งต้องรอการ Challenge จากระบบ และเมื่อได้รับ

Response หรือผลการตอบที่ถูกต้องจากผู้ใช้งานที่เป็นมนุษย์แล้ว SMS ถูกจำแนกเป็น ham แล้วจะถูกส่งไปยังผู้รับปลายทาง โดยมีการส่งผ่านทั้งหมด 8 ส่วน คือ

S --> MSC/SMSG --> S --> MSC/SMSG --> SMSC --> HLR --> SMSC --> MSC --> R

กำหนดให้ผลรวมการส่งผ่านข้อความในกรณีที่ 2 นี้ด้วย N_{un}

กรณีที่ 3 (c) SMS ถูกจำแนกอยู่ระหว่างขอบบนและขอบล่าง (Upper – Lower boundary) ในพื้นที่สีเทา (Uncertain region) ซึ่งต้องรอการ Challenge จากระบบ แต่เมื่อไม่ได้รับ Response หรือผลการตอบจากผู้ใช้งานที่เป็นมนุษย์ หรืออาจจะเป็น bot ที่สร้างข้อความขยะ ที่อาจจะโจมตี SMSC ระบบการกรองข้อความก็จะไม่ทำการส่ง SMS ไปยังผู้รับปลายทาง ดังนั้นข้อความในลักษณะนี้จึงจัดเป็น spam มีจำนวนการส่งผ่าน 2 ส่วน คือ

S --> MSC/SMSG --> S

กำหนดให้ผลรวมการส่งผ่านข้อความในกรณีที่ 3 นี้ด้วย N_{us}

กรณีที่ 4 (d) ข้อความถูกจำแนกเป็นขยะ spam จากการที่ค่าความน่าจะเป็นมีค่าต่ำกว่า Lower threshold จึงไม่ถูกส่งต่อไปยังผู้รับปลายทาง โดยข้อความดังกล่าวจะถูกคัดออกที่ศูนย์กลางบริการข้อความ (SMSC) จึงมีจำนวนการส่งผ่าน 1 ส่วน คือ

S --> MSC/SMSG

จากทั้ง 4 ของการกรองข้อความ Hybrid สามารถคำนวณปริมาณการส่งผ่านข้อมูลในโครงสร้างระบบการรับ-ส่ง SMS ได้ดังสมการ

$$N_n = | y_{\vec{c}=ham}^{h_2} | \times 6 \qquad (4.9)$$

$$N_{un} = | y^{h_1}_{\bar{c}=ham} \cap y^{h_2}_{\bar{c}=spam} \cap y_{\bar{c}=ham} | \times (1-e_1) \times 8$$
$$+ | y^{h_1}_{\bar{c}=ham} \cap y^{h_2}_{\bar{c}=spam} \cap y_{\bar{c}=spam} | \times e_2 \times 8 \quad (4.10)$$

$$N_{us} = | y^{h_1}_{\bar{c}=ham} \cap y^{h_2}_{\bar{c}=spam} \cap y_{\bar{c}=spam} | \times (1-e_2) \times 2$$
$$+ | y^{h_1}_{\bar{c}=ham} \cap y^{h_2}_{\bar{c}=spam} \cap y_{\bar{c}=ham} | \times e_1 \times 2 \quad (4.11)$$

$$N_{hybrid} = N_n + N_{un} + N_{us} + N_s \quad (4.12)$$

เมื่อ e_1 คือความน่าจะเป็นที่มนุษย์ตอบกลับผิดพลาด

e_2 คือความน่าจะเป็นที่ Spam จาก machine ผ่านการกรองข้อความได้

จากสมการ 4.9 – 4.12 เพื่อให้การพิจารณาง่ายยิ่งขึ้นเราจึงมีการกำหนดค่าความน่าจะเป็นของค่า e_1 และ e_2 มีค่าที่น้อยมาก โดยกำหนดค่า 0.02 และ 0.01 ตามลำดับ ซึ่งอ้างอิงจากรูปแบบจำลองการงานของ Challenge-response [18]

4.6 ROC Curve (Receiver Operating Characteristic Curve)

อีกหนึ่งตัวชี้วัดที่ดีที่ใช้ในการคัดแยกคือ Receiver Operating Characteristic (ROC) [18] โดยในวิทยานิพนธ์ฉบับนี้ได้นำค่าต่างๆ จากผลการวิเคราะห์การกรองข้อความมาเปรียบเทียบกันระหว่าง CB filtering และ Hybrid ซึ่งนำค่าที่อยู่ในระหว่างพื้นที่สีเทา (uncertain region) คือ True Positive (TP), True Negative (TN), False Positive (FP) และ False Negative (FN) ที่มีค่าตั้งแต่ 0 ถึง 1 โดยเปรียบเทียบระหว่าง CB Filtering ที่มี h เป็นค่าอ้างอิงเพียงค่าเดียวและแบบผสมที่มีค่า h_1 และ h_2 ซึ่งประมาณค่าได้จากสมการ

$$Specificity = \frac{TN}{FP + TN} \quad (4.13)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (4.14)$$

บทที่ 5

ผลการทดลองและผลการวิเคราะห์ข้อมูล

ในบทนี้จะได้ทำการแสดงผลการจำลองการทำงานของระบบการกรองข้อความ ทั้งในแบบการกรองข้อความจากเนื้อหา (CB filtering) ที่ได้นำกลุ่มตัวอย่างข้อความ Spam เพื่อใช้ในการเรียนรู้จากบทที่ 4 มาจำลองกับโปรแกรมที่ใช้งานบนคอมพิวเตอร์ซึ่งจะได้ผลของความน่าจะเป็นโดยถูกแบ่งออกเป็น 2 กลุ่มคือ ham (SMS ปกติ) และ Spam (SMS ขยะ) จากนั้นจึงนำผลดังกล่าวไปใช้ในการวิเคราะห์ผลที่คาดว่าจะเกิดขึ้นเมื่อมีการทำงานร่วมกับการรับรองจากมนุษย์ (Hybrid SMS Spam filtering) ในลักษณะความน่าจะเป็นที่เกิดขึ้นจากการเปลี่ยนแปลงค่าของจุดอ้างอิงทั้ง 2 ค่าคือ Upper threshold (จุดอ้างอิงขอบบน) และค่า Lower threshold (จุดอ้างอิงขอบล่าง) และนอกจากนี้ยังนำค่าความน่าจะเป็นที่เกิดขึ้นของการกรองข้อความจากเนื้อหาและการรับรองจากมนุษย์มาเปรียบเทียบถึงแนวโน้มของความถูกต้องที่จะเกิดขึ้น เพื่อให้เห็นถึงประสิทธิภาพในการคัดแยกข้อความ Spam จากการทำงานร่วมกันของวิธีการทั้งสอง โดยผลการทดลองและผลการวิเคราะห์ข้อมูลจะแบ่งออกเป็น 3 ส่วนดังนี้

5.1 ผลการจำลองการทำงานของการกรองข้อความจากเนื้อหา (CB filtering)

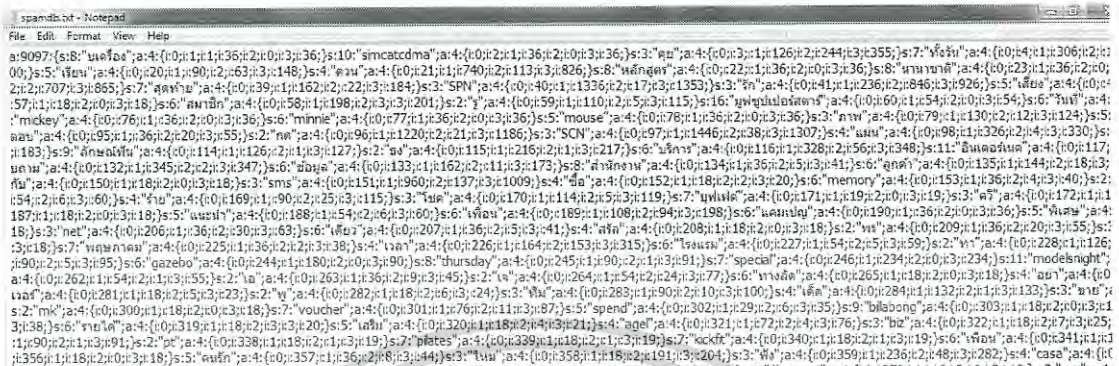
5.2 ผลการวิเคราะห์ข้อมูลของการกรองข้อความแบบผสม (Hybrid)

5.3 ผลการวิเคราะห์ความน่าจะเป็นของค่าความถูกต้องโดยเปรียบเทียบระหว่างการกรองข้อความจากเนื้อหา (CB filtering) และการกรองข้อความแบบผสม (Hybrid)

ทั้งนี้ในบทที่ 5 นี้จะได้แสดงผลของแต่ละขั้นตอนที่ได้ดำเนินการให้เห็นอย่างละเอียดตามรูปแบบที่ได้นำเสนอในวิทยานิพนธ์ฉบับนี้ดังนี้

- กลุ่มข้อความที่นำไปใช้ในการเรียนรู้ ซึ่งได้จากผลสำรวจดังที่แสดงรายละเอียดในบทที่ 4 โดยสรุปแล้วมีจำนวนทั้งหมด 79 คำ ดังแสดงในตารางที่ 4.3 ตัวอย่างคำที่ได้จากผลสำรวจเช่น โบนัส ดวง ดารา ฟรี ลุ้น สนุก สลาก ดูดวง ดูหมอ ด่วน ดาวน์โหลด คลิป คูปอง สิทธิพิเศษ สินค้า สินค้า เชิญชวน เด็ด ชิง ลด พิเศษ เติม เบอร์ โทน สอบถาม เป็นต้น ซึ่งส่วนใหญ่เป็นกลุ่มคำที่เป็นลักษณะการเชิญชวนให้เข้าไปซื้อสินค้าหรือเป็น sms แจ้งโฆษณาชวนเชื่อต่างๆ โดยในการเรียนรู้ของโปรแกรมที่สร้างเพื่อจำลองการทำงานนั้น ฐานข้อมูลจะต้องมีทั้งข้อความที่เป็นรูปแบบปกติ (ham) และข้อความขยะ

(spam) ด้วย โดยข้อมูลดังกล่าวนี้จะถูกจำแนกด้วยมนุษย์และป้อนข้อมูล (ใช้ชื่อไฟล์สำหรับฐานข้อมูลนี้ว่า spamdb.txt) โดยตัวอย่างดังแสดงในรูปที่ 5.1



รูปที่ 5.1 ตัวอย่างฐานข้อมูล spamdb.txt

- ตัวอย่างของข้อความที่จะนำเข้าไปจำแนกว่าเป็น ham หรือ spam ซึ่งมีจำนวน 1,336 ข้อความ ซึ่งมีทั้งภาษาไทยและภาษาอังกฤษผสมกัน (ใช้ชื่อไฟล์สำหรับข้อมูลนี้ว่า sms.txt) โดยนำมาจากระบบโทรศัพท์มือถือ ดังแสดงในรูปที่ 5.2


sms.txt - Notepad
File Edit Format View Help

ช่วงเวลาที่รู้จักกัน ไม่มีวันไหนที่ไม่คิดถึงเธอ
นอนยังครับฝันดีนะคิดถึงนะ
")ภไม่เอาคิดถึโว้ยนะคะ ! กลัวคะไม่อยากได้ยิน ! รู้หละซิว่าเพราะอ"
center.
(1/2)DD7105 26กพ.51เปลี่ยนเวลาเป็นจาก หาดใหญ่ 11.35 ถึง
(1/4)DD7105 on 26/02/08 Change Dep.time from 11.10 to 11.35
(2/2)ดอนเมือง 13.00น.ข้อมูลเพิ่มเติมติดต่อเคาน์เตอร์เช็คอิน
(2/4)Arr.DMK at 13.00.Tel.074-227262,227253DD7105 26 กพ.51
(3/4)เปลี่ยนเวลาออกจาก 11.10 เป็น 11.35 ถึงดอนเมืองเวลา 13.00
(4/4)โทร 074227262/227253
,;*;*;*;*;, *('._.')* >")("< I wish u r Lucky , u r Happy Every Minute, Every Hour, Everyday
..อีกหน่อยจะได้เจอคนไม่จริงหลอกให้เจ็บใจ.
: พึงรู้นะ ยังงี้แบบนี้ ขอบคุณมากนะ
? ส่งข้อความมากวน ส่งมาป่วนหี้ยหัวจัยสัน ส่งข้อความสั้น-สั้น ส่งมา
?Good Night Ja?Good night & Sweet dreams.
0864204556;อ่านข้อความแพนไม่ได้เลยครับ เครื่องมันยังไม่ดีนะครับ
0864206286;CDMAบริการส่งฟรี50sms/เตือนส่งฟรีทุกเครือข่ายด้วยนะ/เพชร
0864413329;ความรักหน้าตาเป็นไงไม่เค
0864503917;เค้านอนไม่หลับอา เอ๋อ เค้าจะรอนะ รอวันที่พลอยรักเค้า จะรอ.
0864605869;คิดถึงน้องแอมจัง รักนะจ๊ะ
0864605869;เค้าเหงาจังเรยย์ คิดถึงตัวเองจัง
0864605869;โค-ตะ-ระ งางเลยพึงงะ งานยังไม่เสร็จ น้องแก้วงวงอะเป่า
0864605869;เราก้อหลับในเช่นกัน ได้ยินเสียงSMSตกใจสดุ้งวับเลย
0864607088;แม่ขี้น้อย พี่หมูไม่ได้ยินเสียงตั้ง 5 รัน รู้ไมจะบ้าตายนะค่
"0864607088?คิดถึงน0ภแม่ขี้น้อยอ่างจะทำให้ขาปล่ะนี่โภคิศิศ4 เป็นห่วงเค้อ"
0864703807;บอกโตมกันว่าอย่าพวแฟนเราเหียวมากแรงนะเป็นห่วงทั้งสองคนแหละ
0864706189;กู๊ดไนท์ จุ๊บ จุ๊บ นะค้า ลูกหมู นอนหลับฝันดีนะ ฝันถึงพ่อหมูด้วยนะจ๊ะ
0864709435;พี่เจม ถึงบ้านยัง ขับรถกลับบ้านดีๆนะคะ น้องเป็นห่วงคะ
0864709435;รักมากนะไอ้หมาตัวโต พี่พูดด้วยอารมร์หรือจากความรู้สึกคะ
0864709435;หลับฝันดี ฝันถึงด้วยนะ แล้วยาอ่อร่อให้มากเมียหึงนะโว้ยยยยย
0864714711;พี่อี้มีอีกเบอร์ 086-4714711 โทรฟรี 08.01 - 20.00 น./ โต๊ก
0864802840;คิดถึงก็ตอบ ไม่ชอบก็ del
252250ยิงราษฎร บริเวณบ้านทุ่งยาว ม.6 ต./อ.โคกโพธิ์ ปน.
27!แล้ว!คิดถึงแกจัง<
Nok New Schedule!!DD7105/26Feb08dep. HDY11.35 -arr.DMK13.00 hrs. More info.please contact counter check-in
nuu2519@thaimail.com If you are the user of this email, pls.visit www.job-passport.com/mail.php now.
phone i notcall &notspeaking msn and not goto68 internet i n't sl
Ruk cher tarta papa kitthung nay mark
RUKnaGOODnight
Shouldn't reserve the transportation Now!. Please wait the suggestion form the other. Jar ...
SL1:Alarm AIS STM1-E Ch.3 U7 NodeBachao-A(To U14 Saiburi-A)Time22:50 Eff.1800Area NWT & Loop PTN-YLA-NW
SORN 086 -7259682
sorry bell ter lamkan gor speak ma tongtong di no put off the tele
st is lucky suscess kor hai test tid na bell name gig j_m_ satr??p
status : : sleeping -.-zzz
Test
THAN NWE ST009 YONG HAO KNITTING LTD PART 114/28-29 CHIDVANA RD = MAESOD TAK 63110
Thank
Thank จ้า
TJPG.AJภแฮปปี้ต่อโป# "นาทีแรก2บ.ถัดไป1บ." ใช้ได้ถึง30กย.51 สนใจโปรอ
User : tanapak Password : F2S4G0 ขอบคุณที่ใช้บริการค่ะ
User : wasupoj98 Password : X6P7A5 ขอบคุณที่ใช้บริการคะ
wakeup
Wishing u g?Ld time,g?Ld helth,g?Ld cheer ,very happy happy happy and hapiies in ?Lur HBD,,,P.S love u away eiei
You have a new MMS from +66851239340.Your message key is MSG.3.3837
You Sending SMSgig to 0864802840 Complete!!
You Sending SMSgig to 0871175804 Complete!!
ก็ทำอยู่พูดไปก็ไม่ฟังและไม่เชื่อผมมันแลวจะมีใครมาฟังและเชื่อ
กลับจากพิตโลกหรอ ขับรถดี ๆ นะจ๊า ถ่างวงก็หากาแฟทานก่อนนะ เป็นห่วง
กลับมาคิดทำไม บลาๆๆ ไม่บนละ ไปนอนดีกา 555+
กแล้วนี่นา ฮะ ฮะ ฝัน ดี นะ
กอจะลืมทุกสิ่งทุกอย่างที่เป็นเรื่องของเรา
ก่อนตัดสินใจอะไรลงไป คิดถึงหัวใจคนทางนี้บ้างนะ... รักนะ
กันไปเดินเล่นปะ
การกระทำกับคำพูดมันขัดกันทำให้ฉันมองเธอว่าเป็นเรื่องกลัวฉันไม่รับผิดชอ
การโกหกกันบางครั้งมันก็รู้สึกดีแต่บางครั้งมันก็เสียความรู้สึก
การคิดอาจต้องใช้ เหตุผลแต่การคิดถึง ใครบางคน เหตุผลคงไม่มี
การบ้านเสร็จมั้ยคะ

รูปที่ 5.2 ตัวอย่างข้อความ sms.txt

- การทำงานของ Pre-processing ซึ่งเป็นกระบวนที่ถือได้ว่ามีความสำคัญอีกกระบวนการหนึ่งเนื่องจากเราต้องการตรวจสอบคำหลายๆคำที่อยู่ในข้อความ และโดยเฉพาะโครงสร้างของภาษาไทยที่มีการเขียนค่อนข้างซับซ้อนกว่าภาษาอังกฤษ วิทยานิพนธ์ฉบับนี้ใช้วิธีการตัดคำแบบยาวที่สุด (longest matching) ซึ่งเป็นการค้นหาคำเริ่มจากตัวอักษรซ้ายสุดของข้อความนั้นไปยังตัวอักษรถัดไปจนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรม (ชื่อไฟล์ว่า sw.txt) ดังแสดงตัวอย่างในรูปที่ 5.3

รูปที่ 5.3 ตัวอย่างข้อความที่ใช้เป็นพจนานุกรมใน sw.txt

สำหรับการตัดตัวอย่างของข้อความที่ถูกตัดออกมาเป็นคำ แสดงตัวอย่างได้ดังตารางที่ 5.1 คำในวิทยานิพนธ์ฉบับนี้ได้เรียกใช้ไฟล์ swath.exe ซึ่งเป็น Application ที่ถูกสร้างขึ้นมาใช้แยกคำ โดยโปรแกรมจะเป็นฟังก์ชั่นๆ หนึ่งถูกเรียกไปทำงานทีละข้อความ ดังนี้

- เรียกไฟล์ข้อความ sms.txt มาใช้งาน
- ทำการแบ่งข้อความใน sms.txt ออกเป็นทีละบรรทัด
- ตัดสัญลักษณ์ ตัวเลข ออกจากข้อความ
- ตัดคำโดยเรียกใช้ฟังก์ชั่น swath.exe
- นำคำที่ตัดมาใส่เครื่องหมายคั่น (|) เพื่อแยกคำ
- นำคำไปตรวจสอบแนบโน้มความน่าจะเป็น spam ต่อไป

ตารางที่ 5.1 ตัวอย่างข้อความที่ถูกตัดคำ

คำที่ถูกตัด
ยกเว้น|ซิ|มว๊าวซ์|ที่|ยังคง|มี|โป|ร|โม|ชั่น|นี้|เติม|บาท|อยู่|ได้|นาน|ปี|
รับ|ปีใหม่|ด้วย|โป|ร|โม|ชั่น|นาฬิกา|สุด|หรู|วันนี้|ม|ค|โทร|
โป|ร|โม|ชั่น|วันนี้|มค|ที่|บูธรมย์รวินท์|เมือง|ทอง|t|
โป|ร|โม|ชั่น|double|a|big|thank|สำหรับ|ลูกค้า|ที่|ซื้อ|กระดาษ|double|a|ก|
cat|โป|ร|โม|ชั่น|ราย|เดือน|หมดสิ้น|เดือน|นี้|ไม่|แน่ใจ|ว่า|จะ|ขยาย|หรือ|ไม่|
chana|สวัสดี|ปีใหม่|โป|ร|โม|ชั่น|สำหรับ|ลูกค้า|ทุก|ท่าน|ให้|ทันที|
fe|โป|ร|โม|ชั่น|ซื้อ|ค|วอน|ตัม|ชิ้น|ฟรี|ชิ้น|รวม|ได้|ชิ้น|พร้อม|i|d|แถม|ฟรี|ค|วอน|ตัมส|ปา|อาบน้ำ|
happy|new|years|good|health|โป|ร|โม|ชั่น|พิเศษ|ถึง|ม|ค|เท่า|น|
hinet|by|cat|ขยาย|โป|ร|โม|ชั่น|ถึง|มิ|ย|โดย|ส่วน|การตลาด|ขต|ตต|
u|bar|ballantine|s|ต้อนรับ|ปีใหม่|กับ|โป|ร|โม|ชั่น|ตอบแทน|ลูกค้า|กิน|เท|
catcdma|โป|ร|โม|ชั่น|ใหม่|บาท|เริ่ม|เวลา|ถึง|นอก|เวลา|ราคา|
glife|พบ|โป|ร|โม|ชั่น|พิเศษ|พร้อม|ของแถม|ได้|แล้ว|วันนี้|ม|ค|ที่|สาขา|ศรี|ราชา|นะ|คะ|
glife|พบ|โป|ร|โม|ชั่น|พิเศษ|พร้อม|ของแถม|ได้|แล้ว|วันนี้|ม|ค|ที่|สาขา|หาด|ใหญ่|นะ|คะ|
glife|โป|ร|โม|ชั่น|พิเศษ|พร้อม|ของแถม|ที่|สาขา|ศรี|ราชา|ตั้งแต่|วันที่|ม|ค|นี้|
กิน|ข้าว|ได้|แล้ว|นะ|คับ|โป|ร|โม|ชั่น|พิเศษ|ถ้า|คุณ|โท|หา|เรา|ตอนนี้|ลุ้น|ข้าว|กล่อง|ฟรี|
ขยาย|เวลา|การ|ลงทะเบียน|สมัคร|โป|ร|โม|ชั่น|kpack|free|ถึง|วันที่|ม|ค|
ขอม|อบ|โป|ร|โม|ชั่น|ปี|นี้|ให้|คุณ|เต็มไปด้วย|เวลา|แห่ง|ความ|สุข|ตลอด|ปี|นะ|ครั|
ขอ|แก้ไข|โป|ร|โม|ชั่น|ตรุษจีน|จาก|clarks|แรง|ขึ้น|กว่า|เดิม|ลด|ทุก|คู่|ทั่ว|ประ|เท|
ขอ|แจ้ง|รายการ|โป|ร|โม|ชั่|นอ|อรัลบี|จาก|ส่วนกลาง|เพิ่มเติม|รายการ|เล่น|ได้|ตั|
ข่าว|ดี|โป|ร|โม|ชั่นน๊อคเอาท์|set|ซื้อ|คลอ|แคป|กระปุก|โสม|พลัส|กระปุก|

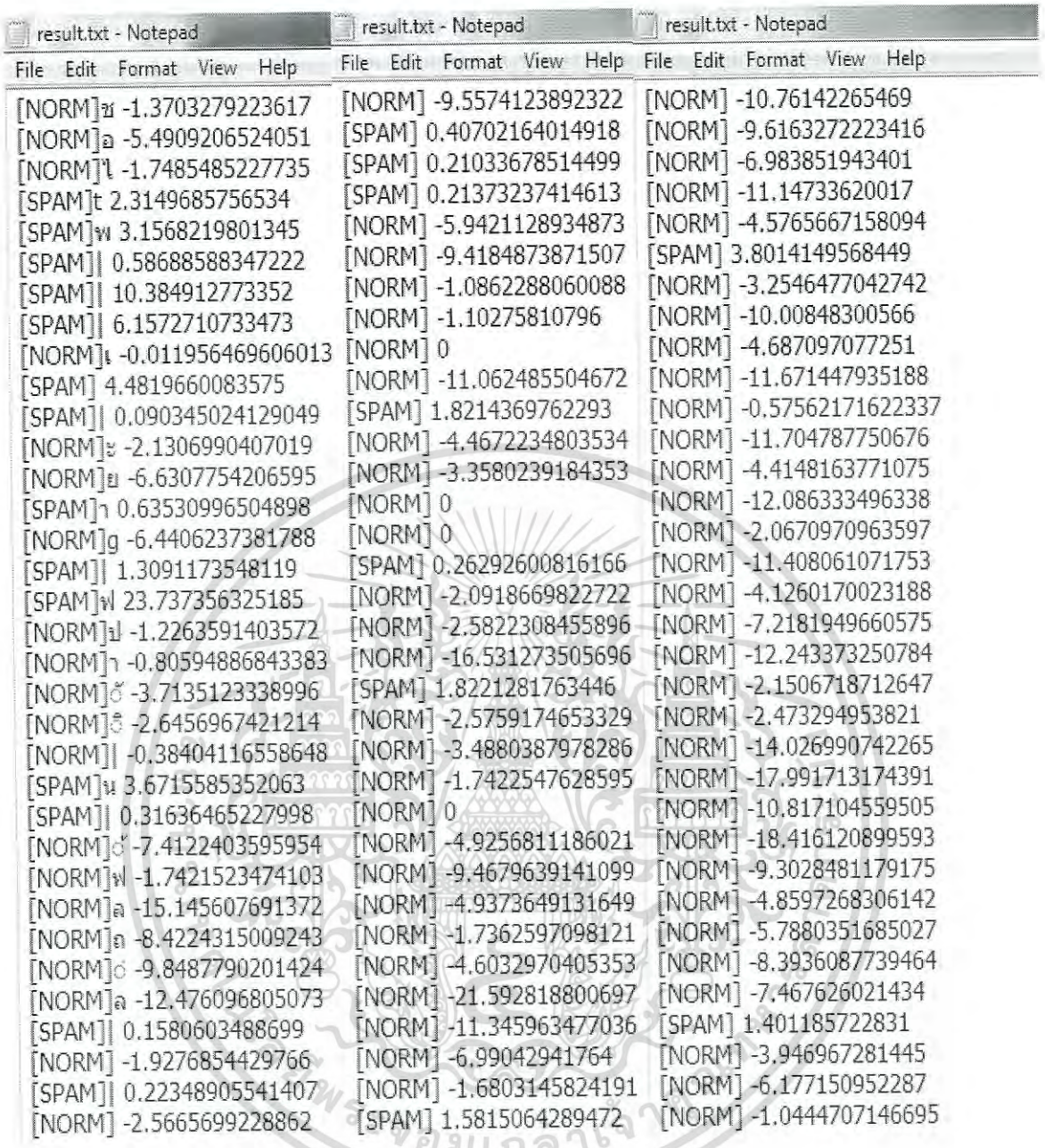
ตารางที่ 5.1 ตัวอย่างข้อความที่ถูกตัดคำ (ต่อ)

ข่าว|ด่วน|โป|ร|โม|ชั่น|ต่อ|double|point|cancun|มี|ต่อ|ถึง|มค|นี้|รีบ|ลุย|ด่วน|
คน|ที่|ขาย|ชุด|โป|ร|โม|ชั่น|พิเศษ|sub|dr|เดือน|นี้|เมื่อ|เปิด|ออ|เดอร์|แล้ว|สามารถ|
ด่วน|ปิด|ex|วันนี้|ส|ม|ค|วัน|เท่านั้น|จะ|ได้|คะแนน|เท่า|ของ|โป|ร|โม|ชั่น|แคน|คู|น|
นี้|เอ|ลอง|เซ่|ยัง|มี|โป|ร|โม|ชั่น|ช่วง|ปีใหม่|นี้|เพียง|ท่าน|ซื้อ|ชุด|เพชร|ติ|นั่ม|มูล|
น|ใช้|งาน|ตาม|โป|ร|โม|ชั่น|ที่|ระบุ|ไว้|ด้วย|ค่ะ|ราย|เดือน|บ|โทร|ฟรี|ทุก|เครือ|ข|
น้ำ|ทับทิม|หมด|โป|ร|โม|ชั่น|พรุ่งนี้|นะ|คะ|฿กิฟฟา|รี|น|
น้ำ|ทับทิม|หมด|โป|ร|โม|ชั่น|วันที่|นี้|นะ|คะ|฿กิฟฟา|รี|น|
ปีใหม่|ขอ|ท่าน|และ|ครอบครัว|มี|ความ|สุข|เช่นเดียวกัน|จาก|ใจ|ขาว|ทอง|โป|ร|โม|ชั่น|ชุม|
ฝาก|ประชาสัมพันธ์|โป|ร|โม|ชั่น|สำหรับ|เทศกาล|วา|เลน|ไทน์|สำหรับ|คู่รัก|ที่|ม|
ร|รับ|โป|ร|โม|ชั่น|ของ|ทาง|โครงการ|เป็น|ที่|เรียบร้อย|แล้ว|ซึ่ง|ท่าน|สามารถ|ติดต่อ|
ร|รับ|โป|ร|โม|ชั่น|ของ|ทาง|โครงการ|เป็น|ที่|เรียบร้อย|แล้ว|ซึ่ง|ท่าน|สามารถ|ติดต่อ|
รับ|โป|ร|โม|ชั่น|โดน|ใจ|เพียง|แจ้ง|เปิด|บัตร|โทร|แล้ว|ใช้จ่าย|ผ่าน|บัตร|tmb|
อ|โร|ม่า|กรุ๊ป|มี|โป|ร|โม|ชั่น|พิเศษ|เครื่อง|ชงกาแฟ|จาก|ประเทศ|สเปน|ลด|วั|
เซ็น|จู|รี่|โกลด์|โป|ร|โม|ชั่น|ตรุษจีน|ปกติ|บาท|พิเศษ|กรัม|ละ|บาท|
เดือน|หน้า|พร้อม|แจ้ง|โป|ร|โม|ชั่น|ใน|การ|ออม|เงิน|งวด|ต่อไป|ด่วน|ค่ะ|ขวัญเมือง|ไทย|
เริ่ม|แล้ว|โป|ร|โม|ชั่น|ท่องเที่ยว|ต่าง|ประเทศ|สำหรับ|สมาชิก|ใหม่|ง่าย|แสน|ง่าย|
เสริม|สุข|ต้อนรับ|ตรุษจีน|ด้วย|โป|ร|โม|ชั่น|สุด|คุ้ม|จาก|clarks|ซื้อ|คู่|ลด|ซี้|
โทร|มา|เร็ว|มี|เรื่อง|ถาม|ไม่|โทร|ไป|เพราะ|หมดเวลา|โป|ร|โม|ชั่น|ต้อง|เสีย|เงิน|เพิ่ม|
โป|ร|โม|ชั่น|hinet|ต่อไป|ถึง|มิ|ย|surin|
โป|ร|โม|ชั่น|งาน|สัมมนา|cpf|turbo|รับ|กุ้ง|ภายใน|มค|โอน|เงิน|ภายใน|มค|
โป|ร|โม|ชั่น|ชุด|ของขวัญ|ม|ค|และ|คูปอง|ลด|แชมพู|บาท|ม|ค|นี้|เท่านั้น|ค่ะ|
โป|ร|โม|ชั่น|ต้อนรับ|ปีใหม่|เลือก|รับ|โบนัส|กัน|ง่าย|ตั้งแต่|ที่|ww|
โป|ร|โม|ชั่นน๊อคเอาท์|ยก|ที่|ซื้อ|โสม|c|c|คลอ|แคป|ซาน|ซี|โต้|โสม|c|c|
โป|ร|โม|ชั่น|ประจำเดือน|ชุด|ที่|พลัส|ยาสีฟัน|บาท|คะแนน|ชุด|ท|
โป|ร|โม|ชั่น|ปีใหม่|ที่นา|รา|กุล|set|a|บาท|pp|mixer|ขวด|อาหาร|
คุณ|คือ|สมาชิก|ที่|ผ่าน|โป|ร|โม|ชั่น|เกาหลี|ขอ|เชิญ|ร่วม|ถ่าย|ภาพ|และ|โชว์|ตัว|ใน|
ขยาย|เวลา|การ|ลง|ทะ|บียน|สมัคร|โป|ร|โม|ชั่น|kpack|free|ถึง|วันที่|ม|
บ|ลูกหลาน|ฟรี|ชิง|รางวัล|lcd|นิ้ว|สร้อย|ทอง|ประธาน|ไพ|
จาก|วัน|อนุมัติ|รับ|ฟรี|คะแนน|สะสม|scb|rewards|คะแนน|
โดย|มี|ยอด|ซื้อ|ครบ|บาท|รับ|ฟรี|ชุด|กล่อง|อเนกประสงค์|กล่อง|มู|
จาก|วัน|อนุมัติ|รับ|ฟรี|คะแนน|สะสม|scb|rewards|คะแนน|
นี่|รู้สึก|ว่า|promotion|ที่|น|ใช้|นะ|มัน|โทร|ฟรี|แค่|ใน|เครือข่าย|ais|นะ|
พี่|แบงค์|คะ|ทาน|ข้าว|ให้|อร่อย|น๊ะคะ|ช่วง|เย็น|ไม่|มี|โป|ร|โทร|ฟรี|
ไป|ตลาด|ได้|ผัก|ฟรี|มา|เลย|ทำ|เมี่ยง|ผัก|กิน|ก็|อร่อย|ดี|พรุ่งนี้|จะ|กิน|ต่อ|
ห้องเช่า|เตียง|พร้อม|ที่|นอน|ตู้|พัดลม|น้ำ|ฟรี|จอด|รถ|สะดวก|ติดต่อ|

ตารางที่ 5.1 ตัวอย่างข้อความที่ถูกตัดคำ (ต่อ)

ยูร|โทร|ฟรี|นาฬิกา|ว่าง|ก็|โทร|มา|ภายใน|เวลา|นะ|
นางเอก|ขี้|ร้อง|อะ|พี่|ตูน|ส่ง|ข้อความ|ฟรี|นะ|มะ|ต้อง|ส่ง|กลับ|มา|อา|ผ่า|
มค|กพช้อป|ผ่าน|บัตร|ทุก|บ|รับ|ฟรี|บัตร|ของขวัญ|บ|ตาม|เงื่อนไข|บริษัท|
ม|ค|ร่วม|ฉลอง|ครบ|รอบ|ปี|memory|pub|ฟรี|คอนเสิร์ต|บอดี้ส|แล|ม|ฟรีบาร์|
happy|new|year|ขอให้|สมาชิก|ชมรม|กิน|ฟรี|มี|เกียรติ|ทุก|ท่าน|มี|ความ|สุข|กาย|สุ|
ยกเลิก|ส่ง|ฟรี|ออก|ตั๋ว|ภายใน|ครับ|รบกวน|ด้วย|ครับ|เก่ง|
มค|เชิญ|ร่วมงาน|amway|expo|เมือง|ทอง|ฟรี|กิจกรรม|ลุ้น|รางวัล|ล้าน|
ให้|มี|สติ|แข็งแรง|มั่งคั่ง|อลังการ|ให้|มี|พลัง|ได้|ฟรี|ไม่|มี|วัน|หมด|ตั๋|
protvs|ม|ค|คิด|ค่า|ติดตั้ง|บาท|กรณี|ต้องการ|ฟรี|ค่า|ติดตั้ง|ต้อง|คีย์|ภาย|
rac|เพชรบูรณ์|จัด|สัม|นา|ผู้ปกครอง|หัวข้อ|เส้นทาง|การ|เข้า|มหา|ลัย|ฟรี|แต่|ต้อง|จอง|
countdown|ที่|สตาร์เวลล์|บาหลี|รับ|เบียร์|ฟรี|ขวด|โทร|
toyota|ธนบุรี|เชิญ|ช|ฟรี|รายการ|โทร|ลด|อะไหล่|วันนี้|ถึง|
กศน|บางกอก|น้อย|สอน|หลากหลาย|อาชีพ|ฟรี|รับ|ถึง|ม|ค|t|
กรุณา|คลิก|ที่|link|นี้|เพื่อ|ดาวน์|โหลด|โปรแกรม|ลบ|และ|ป้องกัน|ไวรัส|ฟรี|จาก|เอ|
ก|ถุง|รับ|ฟรี|ทันที|ถัง|ปูน|อย่าง|ดี|ใบ|ที่|ร้าน|โฮ|มมาร์ท|โฮ|มเอ๊กเพ|รส|และ|ร้|
ขยาย|เวลา|สิทธิ์|พิเศษ|ฟรี|ปากกา|parker|เมื่อ|ซื้อ|น้ำมัน|ปลา|ถึง|มค|นี้|
ข่าว|ดี|รักษา|ยอด|ก่อน|มค|นี้|รับ|hrt|ฟรี|รักษา|ยอด|แถม|แถม|จำนวน|จำกัด|

- เมื่อนำคำที่ได้จากการตัดคำไปเปรียบเทียบความน่าจะเป็นขยะกับฐานข้อมูลที่เก็บไว้ และนำทุกคำมารวมกันในข้อความนั้นๆ ดังสมการที่ 4.7 แล้วก็จะได้ค่า $p(C|y_i)$ ของแต่ละข้อความ โดยค่าตั้งค่าอ้างอิงหรือ h=0.5 ซึ่งก็จะสามารถแยกได้ว่าข้อความใดเป็น ham หรือ spam โดยตัวอย่างของเอาท์พุตของ sms ที่ได้จากการจำลองการทำงานของโปรแกรมที่สร้างขึ้นจะแสดงว่าข้อความนั้นเป็น ham หรือ spam ซึ่งจะเป็นลำดับของ Index ในแต่ละข้อความและนำ Index นั้นไปเทียบกับข้อมูล sms ที่นำเข้าไปทดสอบ ดังแสดงในรูปที่ 5.4 พร้อมกันนี้ตัวอย่างข้อความที่ได้จำแนกแล้วเป็น ham แสดงตัวอย่างในตารางที่ 5.2 และตัวอย่างข้อความที่เป็น spam แสดงในตางรางที่ 5.3



รูปที่ 5.4 ตัวอย่างเอาท์พุตจากการจำลองการทำงาน

ตารางที่ 5.2 ตัวอย่างข้อความที่เป็น ham (ข้อความปกติ)

ประชุมรับมอบข้อปฏิบัติเร่งด่วน วันนี้2ก.ย.51เวลา14.00น.ห้องPOC
จัดสเปคคอมออฟฟิศให้พี่หน่อยจะสั่งวันนี้ด่วนหรือติดต่อกลับพี่แนว
คุณมณฑล (โปทีม)จะซื้อTotalใหม่ 1 ตัว โทรกลับด่วนค่ะ
ขอสำเนาบัตรนักศึกษา ส่งให้พี่ด้วยครับ ด่วนเจ้า (พี่นัท)

ตารางที่ 5.2 ตัวอย่างข้อความที่เป็น ham (ข้อความปกติ) (ต่อ)

กรุณาติดต่อชำระค่างวดรถยนต์ TOYOTA VIOS กับทางธนาคารธนชาตด่วน
ค่ากระแสไฟฟ้าของท่านครบกำหนดแล้ว โปรดติดต่อชำระเงินโดยด่วนภายใน 2กย51
ค่าไฟของท่านครบกำหนดแล้ว โปรดติดต่อชำระเงินโดยด่วน ภายใน 1กย51
ตลอดเวลา โปรโมชั่นวันนี้คิดถึงหนึ่งนาที หอมแก้มฟรีหนึ่ง ครั้ง ...
ทัศนี แพงคำฮัก CTB RC ขอใบสมรส ด่วนๆ จ้า
ฐานวัฒน์ แสรกำรุ่งเกียติ CTB RC ขอใบเปลี่ยนชื่อ และ นามสกุล ครับด่วน
กรุณาแจ้งคุณนรินทร์ มะธิตะถังชำระค่างวดรถภายในวันที่15/09/51ด่วนเพื่อป
โทรกลับคุณสิทธิชัย 0814253389 จะให้ทำผังซอยโรงพยาบาลพระราม 2 ด่วนค่ะ
ด่วนเนอะอ้ายไม่มีเงินใช้
ทดสอบกำลังเรียนส่งเอสเอ็มเอสฟรีจ้า*-*จากออยเล็กผู้น่ารัก
กลับมาที่ตึก 70 ปี ด่วน
ทดสอบกำลังเรียนส่งเอสเอ็มเอสฟรีจ้า
5พรรคร่วมฯถกด่วนบ้านสุวัจน์เผยให้สิทธิ์พปช.เลือกนายกฯ-สมพงษ์ยังเหนียม
พี่บอย ลิเวอร์ชนะ มาเลี้ยงไก่น้องๆด่วนนนนนน 5555+
เรียนคุณรุจิรากรุณาชำระค่างวดรถยนต์ที่ค้างชำระอยู่ 2งวดด่วนคะภายในวันท
โทรศัพท์โทรฟรีอยู่ในกระเป๋าเครื่องสำอางบนโต๊ะ
น้ำท่วมน้ำป่าไหลหลากในพื้นที่โปรดแจ้งสำนักฯ3โดยด่วนเพื่อรายงานกรมฯทราบ
ด่วน!ศาลฎีกาออกหมายจับทักษิณ-พจมาณ ให้มาฟังคำตัดสินคดีรัชดา21ตค.14น.
เติ่งๆ มีเรื่องแล้ว โทรกลับตอนนี้ด่วน
แจ้งเพื่อทราบขณะนี้อินเตอร์เน็ตใช้งานไม่ได้กำลังแก้ไขอย่างเร่งด่วนครับ
ว่าแต่ไปดูแผนที่หรือยังว่าบ้านพี่อยู่ทางทิศไหนของลพบุรี รีบหาคำตอบด่วน
ม.พายัพรับนักศึกษาใหม่ มีทุนฟรีถึง10พ.ย.T.053851478#481www.payap.ac.th
คิดถึงทุกเวลา อยากโทรหาทุกนาที ถ้าทักษิณให้โทรฟรี ทุกนาทีมีแต่เทอ
ร.ร.ตรีมิตรขอแจ้ง นร.ในความดูแลของท่านไม่ผ่านเกณฑ์การประเมินติดต่อด่วน
ด่วน!สมชายประชุมครม.ที่ดอนเมือง-ให้ตร.เครียพท.หาช่องแถลงนโยบายให้ได้
ด่วน!ตร.ยิงแก๊สน้ำตาสลายการชุมนุมพธม.หน้ารัฐสภา-มีคนเจ็บหลายคน
กรุณาชำระยอดค้างสินเชื่อเงินผ่อน Capital OK ภายในวันนี้ด่วน
นายกฯเรียก!ผบ.ทุกเหล่าทัพประชุมด่วน14.00น.ที่บก.สูงสุดห้องประชุมชั้น4
นักเรียนติด ร ติดต่อครูภารดี ด่วน โทร 0894961381
มีปัญหาผลการเรียน ติดต่อครูดิสมัยด่วน
โหรส.ว.ฟันธง!ดวงสมชายส่อหลุดเก้าอี้นายกฯ เชื่อต้องออก-ยุบสภาก่อน15ธ.ค.
อยู่ที่ไหนมารายงานตัวด่วนขณะนี้ไม่มี เจ้าหน้าที่เวร สก.บางรัก/คุณลอย
ฟรี!บ่ายนี้ผลสลากรางวัลที่1เลขท้าย2และ3ตัวจะส่งตรงถึงคุณทางsmsโทร1113
ข่าวด่วน:เกิดไฟไหม้ผับใหญ่ ระหว่างเอกมัยซอย9และ11มีนักท่องเที่ยวเสียชี
คุณมี 1 ข้อความใหม่ เมื่อ 20/11/51,18:24น. ฟังข้อความเสียงกด*99 ฟรี
[ฝ้าย]โกรธนะ แต่ก็ขอบใจมากจ๊ะ เพลงฮิตเกาหลี *452827621756611
ก็มันไม่ใช่อ่ะ เปงเพื่อนกันอ่ะดีแร้ว *
ส่งช้าไปมั้ย นอนยัง ?!? พึ่งท่องเสดค่ะ รั๊กสุดขอบโลกเลย > < จุ๊บส์~!!*
ปะออน "ฝันหวานๆนอ คืนนี้หนาวบ้า* ^^
พรุ่งนี้เอาสายเกียร์ออโต้2เส้นไปให้พ่อด่วน
เหอะน่า อย่าบ่นขอร้อง* เมิงนิ 555

ตารางที่ 5.2 ตัวอย่างข้อความที่เป็น ham (ข้อความปกติ) (ต่อ)

ส่งช้าไปมั้ย นอนยัง ?!? พึ่งท่องเสดค่ะ รั๊กสุดขอบโลกเลย > < จุ๊บส์~!!*
เยลลี่ฟรีอร่อยจัง ขอบคุณนะคะ
ส่งช้าไปมั้ย นอนยัง ?!? พึ่งท่องเสดค่ะ รั๊กสุดขอบโลกเลย > < จุ๊บส์~!!*
โทรหาบอมย์ด่วน
[นิอัซ]ไม่เห็นออนเอ็มเลย พี่วัน เพลงฮิตเกาหลี *452827621756611
์เองเหรอค่ะ งง *?*
โรงเรียน...ฝันดีนะ *.*
ไม่ได้ชื่อเหมือน = =*
เอาเปนว่า รีบนอนดีกว่านะ* เจอกัลพุ่งนี้ขี๊ะ
ฝันดีนะเทอ!รักเทอมากนะ รักกัลบ้างป่าว?~รักกัลไปนานๆนะ*คิดถึงมากมาย
พรุ่งนี้เอาสายเกียร์ออโต้2เส้นไปให้พ่อด่วน
ส่งช้าไปมั้ย นอนยัง ?!? พึ่งท่องเสดค่ะ รั๊กสุดขอบโลกเลย > < จุ๊บส์~!!*
ทำมัยม่ารับอะ โกดเค้าหรอ ?? ฝันดีน๊า า ค๊า ที่ รัก ของ ปา โล มา า*
สุขสันต์วันเกิดนะมีความสุขมากๆ เป็นเด็กดีนะ ดูแลกันและกันนะ
ที่รักของน้องเอพักผ่อนบ้างนะคะ เป็นห่วงมากๆ รักและคิดถึงพี่ชินที่สุด
0864484247;ทำไมกวนแบบนี้เนี่ย ไม่เข้าใจ ปลาทอง ทำมั่งแล้วจะรู้สึกนะ
ที่รักถึงบ้านแล้วโทรหาด้วยอย่ากินเยอะไม่ไหวก็พอนุกห่วงและรักแอนนะครับ
จูงมือก้าวไปด้วยกันข้ามภูผา เมฆาตระหง่าน
ก็สิ่งที่นิตทำไง เหตุผล
ไม่รับจะไปนอนที่ห้องนก
เราแต่งงานกันนะเมียรักติ๊กมากรออยู่นะ
ไม่ได้หวังให้กิ่งกลับมารักพี่เข้าใจดีหมดเวลาแก้ตัวของพี่แล้วแต่อยาก
0864484247;บอกให้ไปนอนก่อน ดึก ๆเข้าห้องน้ำก็โทรมา ปลาทองลำบากใจ
ถ้าเตียงที่บ้านดร..มีพี่โอ๋นอนอยู่ข้างอย่างเมื่อกี้ก็ดีซินะ
ถึงหรือยังคะ รถติดสุดๆ ๆ พึ่งถึงคะ
ที่รักเมื่อกี้ฟังวิทยุระวังตำรวจเข้าอย่าลืมนะถึงบ้านโทรด้วยนะห่วงแอน
ฝันดีจ้าคุณพี่..
กบจะให้เมียพี่โทรแทนกบว่าพี่จะรับมั๊ย
ข้อความนี้ส่งเพื่อยืนยันว่าผมรักผู้หญิงคนนี้จริงๆโปรดอย่าถามถึงเหตผล
สุขสันต์วันเกิดขอให้มีความสุขนะ<พี่แป้ม>
Good night.
สุขสันต์วันเกิดครับขอให้มีความสุขและขอให้น่ารักแบบนี้ตลอดไปอย่าให้ใ
ใคถาม
ครมาจีบขอมีพี่จีบคนเดียวอิอิ ครับไม่หวานแต่จริงใจกับหนิงครับ H.B.D
ช้าดิโว้ยย,ไอ้ป๊า
ถึงแล้วฝันดี
Call me back now!!
nuttcg2@gmail.com
happy birth day นะคะพี่ชาย ขอให้หล่อๆรวยๆสาวรักสาวษหลงนะคะ
เบอร์ก๊อฟ 0847924594
คิดถึงจัง นอนหลับฝันดีนะ(ไม่ส่งSMSมาหาบ้างเลย)หรือไม่รักกันแล้ว

ตารางที่ 5.2 ตัวอย่างข้อความที่เป็น ham (ข้อความปกติ) (ต่อ)

มีความสุขมากนะในวันเกินปีนี้มีเราแต่เราก็ยังจะจำและรักไปตลอดรักนะ24ชม
ไอซ์ถ้าการที่ทำแบบนี้มันดีที่สุดแล้วดาวจะไม่โทรหาดูแลตัวเองดีๆล่ะ
ตัว แพรถอนชื่อ ไม่ไปลาวแล้วนะ เปลืองตังค์
แนนโทรกลับด่วน Kosol
ที่ผ่านมานึกว่าบุญทาน
0932148835 นส ศรัญญา ไพศาลวิทย์
ที่รักขา ก๊ัดขนาด ง่วงนอนมั่ก อุอุ ไค่เติ้งหาเด้อ ไค่ได้โตยยย Love Lov
e ^_^
สอบตรง 4 คณะ ธรรมศาสตร์ (ลำปาง) ดูที่ www.pec9.com จาก PEC Online
ในสัญญาที่ทำไว้มียอดชำระ 48 งวด เริ่มตั้งแต่ ธค48 นับเอาเองว่าหมดยัง
หนิงเบื่อว่ะขี้เกียจรอ
ฉุกเฉิน ค่าโทรหมด! รบกวนโทรกลับด้วยที่ 0850490846
An2beem@thaimail.com
ฉุกเฉิน ค่าโทรหมด! รบกวนโทรกลับด้วยที่ 0874569892
พี่ต้องเชื่อผม ผมพูดความจริงเพราะผมรักพี่นะ
ขอละหมาดก่อนนะแล้วจะโทรกลับ
STOP
สงสัยคืนนี้แน่
คิดถึงก็ตอบ ไม่ขอบก็ Deleteที่บ้านมีอีก จะส่งจนกว่าจะคิดถึง
2130กย.ยิงกำนันเสียชีวิต.ม1บ.สะพานไม้แก่นตก.ต.สะพานไม้แก่น.อ.จะนะ.จ.สข
เรียน ผสค.ขอนแก่น,อุดรฯ,นครราชสีมา ขอทราบระยะทางสนง.เศรษฐกิจการเกษตร
0864608076;วันนี้อากาศร้อนเป็นบ้า ถ
้าได้ดูเธอแก้ผ้าฉันคงหายซ่าไปเยอะ
You Sending SMSgig to 0872471128 Complete!!
!-L
ว.0ทท.1ว.28บร.4ก.ย.9โมงเช้าชั้น15ทท.แต่งเครื่องแบบ เสนอผลงาน ฝ.ละ5นาที
กกต.มติเอกฉันท์เสนอยุบพรรคพลังประชาชน
ได้รับรายการสั่งซื้อแล้ว ยอดโอน 7410 บาทค่ะ
พี่กุ้งคะก้อยรบกวนจ่ายค่าไฟของนางแสงดารา แซ่ตั้ง ให้ด้วยนะคะโอนแล้ว730
ฉุกเฉิน ค่าโทรหมด! รบกวนโทรกลับด้วยที่ 0804727098
พักผ่อนเยอะๆนะจะได้ไม่เพลีย ส่วนเค้าตอนนี้ปวดท้องและเพลียด้วย คิดถึ
CB1 AT M FL.
หมูขอตังหน่อย
20000นายบอกจะคืนให้.ยืมอ้อนมันหักจาก100000เก่มี2000ก็เพิ่มให้นายเอง
คอร์ทสยังไม่ได้รับชำระค่างวดจากท่านกรุณาชำระด่วนขออภัยหากชำระแล้ว
021030 ปิดระบบชั่วคราวตั้งแต่วันที่ 1ก.ย.51เป็นต้นไปเนื่องจากทำการย้าย
มากินข้าวที่บ้านไม่ต้องเปลืองเงินและซื้อตับไว้แล้ว
ท่านได้ขาย (บาท) T-CASH 150,000.00
02-1293894 เบอร์แฟกซ์ค่ะ
ผัวคงไม่ว่าเพราะไม่ใช่เป็นคนอื่นโยคงต้องอธิบายมันนานซีเรียสนะ

ตารางที่ 5.3 ตัวอย่างข้อความที่เป็น spam (ข้อความขยะ)

ท่านได้สมัครSMSอ.ลักษณ์3บ./SMSฟรี15วันหมดอายุ16/11/2008สอบถาม025026767
ฟรี!SMSทุกกลเม็ดเคล็ดลับเติมเต็มรักให้ชีวิตคู่สมหวังยืนยาวโทร025020469
หมอกฤษฎ์ คอมเฟิร์มแม่นทั้งวงการดารามาคอมเฟิร์มดวงคุณทางSMSฟรี!โทร1113
ฟรี!smsผลฟุตบอลสรุปครึ่งแรก,ครึ่งหลังสมัครรับบริการด่วน!โทร02-5020477
อย่าพลาด!ชมJamesBondฟรีก่อนใคร10000ที่ 4 พย1ทุ่ม@SFจองวันนี้ที่SFca
โปรพิเศษสมาชิกฟิตเนส 1ปีฟรีที่แคลิฟอร์เนีย ว้าว25ท่านแรกโทร027910000
อีก5วันจะสิ้นสุดเวลารับข่าวฟรี3ด,ใช้ต่อค่าบริการ29บ/ด,ยกเลิกกด*779ค่ะ
Mazdaน้ำมันเครื่องลด30% เช็คฟรี4ระบบพร้อมส่วนลดอะไหล่-15มค. 02-6619880
ออนไลน์ใหม่ล่าสุดทดลองฟรี7วันที่ www.ultrarichteam.comครับ / perawas
บริการนี้ฟรี30วัน(ปกติ29บ./ด)ยกเลิก พิมพ์ B813ส่ง4849090
จำหน่าย Battery Notebook ทุกยี่ห้อราคาถูก ส่งฟรีถึงบ้าน 0-2849-2611
Big Bonus! รับปี 2009 ฟรีส่วนลดซื้อบ้าน 20,000 บ.ลงทะเบียนวันนี้ที่
อยากดูโหลดเลย คลิปดาราสาวและเพื่อนๆในวงการ ฟรี7วัน กด*482539600โทรออก
รับฟรี!! วิธีอยู่แบบไม่มีหนี้ตลอดชีวิต สนใจ กด*48253320012 โทรออก
รับฟรี!! สุดยอดวิธีมีเงินเหลือเก็บ สนใจ กด*48253320012 โทรออก
รับฟรี!! วิธีอยู่แบบไม่มีหนี้ตลอดชีวิต สนใจ กด*48253320012 โทรออก
hi-Q Songคืออะไร? โหลดฟรี พัฒนาการของเพลงเสียงใสกิ๊งระดับ CD บนมือถ
ฟรี!SMS10เทคนิคเติมเต็มรักกับเคล็ดลับพิชิตใจคู่รักง่ายๆแค่โทร025020469
รับฟรี! สารพัดวิธีสวยทันตา ในแบบชมพู่ อารยา สนใจ กด*48259070011โทรออก
รับฟรี! สารพัดวิธีสวยทันตา ในแบบชมพู่ อารยา สนใจ กด*48259070011โทรออก
้องไปสอบที่ ยโส ส่ง SMS ฟรีที่ เว็บของ CDMA นะ ห้อง VAS แหล่งบริการเสร
บแอร์ ห้องนอนใหญ่ ฟรี Urban Sathorn 02-869-7788 เริ่ม 4.29 ล้าน
0864458669;คนที่ใช้มือถือ cat สามารถลงทะเบียนเพื่อใช้สิทธิ์ส่ง sms ฟรี
0864607088;ส่วน Sim โปรโทรฟรี 300 มีนะ ราคาพิเศษ 250 บ สำหรับเบอร์
แจกฟรีนาฬิกาแขวนมูลค่า1,500บาท300ชิ้นสุดท้ายที่ดูโฆษณาโมบิลเดลแวคซ์
สิทธิ์พิเศษเฉพาะสมาชิก Pantip ฟรีน้ำหอม FCUK 100 ml มูลค่า 2100 บาท
12-13พย.ร่วมฉลองเทศกาลลอยกระทงเปิดเหล้ารับกระทงฟรี1 ใบพร้อมร่วมสนุก
ทดลองบริการโทรศัพท์ระหว่างประเทศผ่าน NetTalk ฟรี 100บ. นาน 30 วัน (
คอร์สพิเศษ!90วันเปลี่ยนชีวิตสู่ธุรกิจทำเงินที่มั่นคงโทร.0866825974 ฟรี
ลดทั้งร้าน055717771โนตบุ๊คฟรีแรม2G#แฮนดี้ไดร์2G=199#0%KTC / สยามคอม กพ
ใช้เฟิร์สช้อยส์ซื้อ ACER, HP,Compaq, Vaio ที่เอเน็ต รับฟรี Handy Drive
ใช้เฟิร์สช้อยส์ซื้อคอมพิวเตอร์ ที่ไฮเทคคอมพิวเตอร์ รับฟรี Handy Drive
บริการSahapun1ยกเลิกพิมพ์C813ส่งไป4809099(ฟรี)
ครบกำหนดทดลองใช้ฟรีบริการ INN Hot News (29บ/เดือน) จะต่ออายุบริการใ
ับฟรี LCD TV + Home Theatre 02-869-7788
ชาโตว์ อินทาวน์ รัชดา17จองวันนี้ เราจ่ายให้คุณฟรีทุกเงื่อนไข02-2776776
โทร 1188ลุ้นรับผลิตภัณฑ์ Missha ส่งถึงบ้านฟรี
ถอยมาสด้าBT50วันนี้เลือกรับโปรโมชั่นMotor Expoฟรีน้ำมัน1แสนบาทหรือผ
่อนเพียง5999/เดือนหรือดอกเบี้ย0%ฟรีประกันภัยโทร026619880
โปรพิเศษสมาชิกฟิตเนส 1ปีฟรีที่แคลิฟอร์เนีย ว้าว25ท่านแรกโทร027910000

ตารางที่ 5.3 ตัวอย่างข้อความที่เป็น spam (ข้อความขยะ) (ต่อ)

รับฟรี! ความจริงเกี่ยวกับผู้ชายที่คุณยังไม่รู้ สนใจ กด*48253320013 โทร
ลางกรุงให้ข้อมูลซื้อขายเดอะซิ้ดทุกทำเล ฟรีตลอดงาน 027899955
เปิดตัวฟอร์ดเอสเคปใหม่พบบอลลูนยักษ์ครั้งแรกในภูเก็ตรับฟรีคูปองน้ำมัน
DBเครื่องละเท่าไร+ดูฟรี 3เดือนรับประกันเท่าไรdanai415@yahoo.com
ฟรี ที่02-8805500หรือ0863249977 ช่วง08.30-17.00น. ภายใน7-22 พ.ย.นี้
รับข้อเสนอMotorExpoก่อนใคร ถอยMazdaBT50ดอกเบี้ย0%ฟรีประกันโทร026619880
ซื้ออายุร่าพิงค์เลดี้4ขวดแถมD-tocควบคุมน้ำหนักฟรี1ขวด/หลิน086-5125030
ามาเราก็ดีใจแล้ว พาคนที่ท่านรักไปวัดไขมันได้ฟรีนะคะ ขอบคุณค่ะ/แตง+นุ้ย
0864458669;ผู้ที่ใช้เลขหมาย cat สามารถใช้สิทธิ์ส่ง sms ฟรี 50/เดือน อย
เข้าไปสมัครใช้ฟรีที่ http://www.catconference.com/th/
เปิดบริการ ระบบการส่ง SMS แล้ว สำหรับห้อง 603 เท่านั้น บริการฟรีชั่วคร
[Gibsa]ถ้าโทฟรี โทรหาเหมียวด้วย เพลงใหม่เกาหลี*452827621756611
NaraiPizzeriaฉลอง40ปีพิซซ่าลด40%โชว์SMSฟรีVoucherตลอดพย.51โทร026780555
0864209671;มันเป็นของcatลองสมัครที่เวปtttonline.netดูสินี่ก็ฟรี
ทำประกันสุขภาพที่ธ.กสิกรไทย วันนี้ รับของสมนาคุณฟรี ถึง 31 ธ.ค. 51
โฉมใหม่ www.tollD.com โทรต่ำสุด 80สต สมัครสมาชิกโทรฟรี 1 ชม/02-3001113
ผ่อนโทรศัพท์มือถือที่ TG Fone 1 เครื่อง (เฉพาะรุ่น) ฟรีอีก 1 เครื่อ
หมอกฤษฎ์ คอนเฟิร์มแม่นรับดวงฟรี7วัน(3บ/SMS)โทร*299*001#สอบถาม025026767
โชว์SMSในงานฉลองเปิดศูนย์ฯ21-23พย.ชั้นใต้ดินThe Mall รับของสมนาคุณฟรี
ล้างรถฟรีเมื่อแวะชมแคลิฟอร์เนียฟิตเนส!ดิ อเวนิว พัทยาสาย2 โทร038399999
อย่าลืมกด *551# แล้วโทรออกน่ะค่ะ จะได้โทรฟรีในเครือข่ายทรู
คืนนี้ ฟรีคอนเสิร์ต * ตั๊กแตน ชลดา * @สภาดินโทร 0863676676
อยากดูโหลดเลย!!คลิปดาราสาวและเพื่อนๆในวงการ ฟรี7วัน กด*482539600โทรออก
นิก ชวนคุณเข้ามาร่วมสนุกใน 12Frenz สมัครสมาชิกฟรี! สนุกกันทุกที่ ไม่มี
วันพ่อนี้สมัครประกันลูกกตัญญูที่ธ.กสิกรไทย ฟรีของสมนาคุณถึง 31 ธ.ค. 51
EXT พิเศษ! 200 ท่านแรกโชว์ sms รับของพรีเมี่ยม ฟรี ที่ Booth ลงทะเบ
H2O Plus โชว์ SMS รับฟรี Sample และมอบคะแนนคูณ 3 เมื่อซื้อสินค้าตั้
H2O+เชิญนวดหน้าฟรีที่Mallงามวงศ์วาน28/11-31/12/51สำรองที่T.02-5500386
H2O+เชิญนวดหน้าฟรี ที่Ro.อุดรธานี 28/11-31/12/51 สำรองที่T.042-343809
H2O+เชิญนวดหน้าฟรีที่Ro.เชียงใหม่ 28/11-31/12/51สำรองที่T.053-279-815
H2O+เชิญนวดหน้าฟรี ที่Ro.อุดรธานี 28/11-31/12/51 สำรองที่T.042-343809
สั่งปูนTPIวันนี้1พ่วงรับปูนซ่อมเอนกประสงค์M600ฟรี10ถุงทันทีมีจำกัด/ศทม
ORLAN ร่วมฉลองเปิด Beauty Hall Mall งามวงศ์วาน เพียงโชว์ SMS รับฟรี
ฟิตเนส1เดือนฟรีที่แคลิฟอร์เนียว้าว50ท่านแรก โทรภายใน15นาทีนี้027910000
พร้อมอยู่ ฟรีspecial package กว่าแสนบาท!! เริ่ม 2.29 ล้าน 02 789 1
มิตซูบิชิเชิญชวนลูกค้าทุกท่าน เข้ารับการตรวจสภาพรถยนต์ ฟรี และเปลี่
ฟรี!บ่ายนี้ผลสลากรางวัลที่1เลขท้าย2และ3ตัวจะส่งตรงถึงคุณทางsmsโทร1113
ศูนย์ฯนนทบุรี สาธิตบริการ CAT2Call Plus 8-12ธค.51ทดลองโทรฟรีในงาน สอบถ
ลุ้นSonyVaioฟรี!เมื่อโหลดทรูโทนรักสามเศร้า/พริกไทยโทร*192002290338นะคะ
บินวันนี้ ฟรีค่าธรรมเนียมน้ำมัน จองด่วนวันนี้-14 ธ.ค.ที่airasia.com
แม่นมาก!ฟรีSMSดวงรายวันคอนเฟิร์มทุกเรื่องโดยหมอกฤษณ์คอนเฟิร์มโทร1113

ตารางที่ 5.3 ตัวอย่างข้อความที่เป็น spam (ข้อความขยะ) (ต่อ)

COCKPITจัดฟรีคอนเสิร์ต Jennifer Kim&Mr.Sax man โชว์,เกมส์,รางวัล,อา
0864613288;149บ/ดโทรฟรีทุกเครือข่ายทั้งวัน ที่Mshopศก/Jsmtelecom.com
ฟรีPie@PizzaHutจ่ายด้วยบัตรเดบิต+คูปอง(รับ@KBank)-31ธ.ค.028888888#02
เช็คดวง29พย.-5ธค.!พิสูจน์ความแม่นกับอ.กฤษฎ์โทร1900190069คอนเฟิร์ม
เช็คดวง6-12ธค.!พิสูจน์ความแม่นกับอ.กฤษฎ์โทร1900190059คอนเฟิร์ม
ter3 13.00น 20ธค.นี้@Paragon Cineplexจองด่วน022972299รหัสจอง107634
ลุ้นทองทุกวัน!และรับSMSน่ารักๆสมัครฟรี!กด *298*289# (5บ/ว) 028927322
จองด่วน! ฮิโรชิมา-ฟุกุโอกะ 5วัน 47100 บ. ฟ้าไทยฮอลิเดย์ส โทร.1771 กด 3
ใช้น้ำมันและทำบุญ เวลา 14.00 ใต้ตึกอธิการ ลุ้นรับจักรยาน ฟรี
เปิดบัตรเครดิตบิ๊กซี02-6673684 กด 3 ช้อปครบ3000รอบบัญชีแรกฟรีชุดผ้านวม
น ฟรี! สูงสุด 1 ล้านบาท ด่วน! ช้าหมด โทร.02-617-6900
i-clubลุ้นบัตรชมภาพยนตร์โรแมนติคคอมมิดี้MAMMA MIA!วิวาห์ในฝันท่ามกล
กด*321*Pin4หลัก*เลข4หลักสุดท้ายเบอร์ผิด*Ref 9หลัก#โทรออกฟังฟรี*915667
ด่วน!ประกาศสถานการณ์ฉุกเฉินเฉพาะกทม.สนใจกด*48259080011โทรออก(ฟรี15วัน)
พันธ์มิตรเรียกชุมนุม ด่วนนนนน
นำSMSนี้มาแลกรับของสมนาคุณฟรีที่บูธเฟิร์สช้อยส์ 5-6ก.ย. ร.ร.ทวินโลตัส
สมัครฟรี!รับส่วนลดสินค้าชั้นนำกว่า50%สนใจกด*48253320011โทรออก ได้ทุกคน
ฟรี!ส่วนลดสินค้าแบรนดังมากมายสนใจกด*48253320011โทรออก(ได้ทุกคน)
(Rotring) ซื้อ Waterman ทุกรุ่น ยกเว้น Hermisphere รับฟรีทันที ปากก
ฟรี!!ทอง3เส้น เล่นฟรี!!3วัน ทุกสัปดาห์ สมัครกด*4825971 แล้วโทรออก
ฟรี!อัพเดทเทรนด์แฟชั่นมาแรง ที่นี่!!ก่อนใครสนใจกด*48259070011โทรออก
แม่นจริงๆ!SMSฟันธงดวงการงาน,การเงิน,ความรักคุณโดยอ.ลักษณ์ฟรี!โทร1113
ฟรี!เทคนิคแต่งตัวดูดีมีสไตล์,รวมสูตรลับความงามสนใจกด*48259070011โทรออก
นำSMSรับของแถมฟรีที่บูธเฟิร์สช้อยส์12-14 กย.งานคอมพิวเตอร์ร.ร.สยามธานี
รับฟรี! ทีเด็ดไม้ตาย มัดใจแฟน สนใจกด *4825944 โทรออก
โปรโมชั่นมือถือเป็นเหมาจ่าย 129บาท ในเครือข่ายทรูมูฟโทรฟรี ตี5-5โมงเย็
ฟรี รายงานผลพรีเมียร์ลีกอังกฤษ สด แบบลูกต่อลูก รับผลกด*4825927โทรออก
ASIMAR ราคาต่ำBV1.27 เก็งแจกวอร์ฟรีลดหนี้สินต่อทุนสูง ต้าน1.20 บ
ฟรี!!ทอง3เส้น เล่นฟรี!!3วัน ทุกสัปดาห์ สมัครกด*4825971 แล้วโทรออก
ฟรี!ผลบอลนัดแดงเดือดและพรีเมียร์ลีกทุกคู่สดแบบลูกต่อลูกโทร025020435
ฟรี!เที่ยวเกาหลี/วิรันดารีสอร์ท/โน๊ตบุ๊คAcerAspireone/กด*4825310โทรออก
i-clubลุ้นรับบัตรi-mobileO:IC DeliveryConcert4พบกับCalories BlahBla
น พร้อมลุ้นรับทอง 1 บาทเมื่อผ่อน 10,000 บ.ขึ้นไป วันนี้-30 พ.ย. ตาม
2/2 ลุ้นรางวัลสัมผัสตัวจริงโรนัลดินโญ่ ดูรายละเอียดในอะซีฟ ต.ค.51
Happy ใจดี อนุญาติให้คิดถึงฟรีค่ะ
น พร้อมลุ้นรับทอง 1 บาทเมื่อผ่อน 10,000 บ.ขึ้นไป วันนี้-30 พ.ย. ตาม
ลุ้นทอง5บาท!กดเลย *298*289# พร้อมรับSMSใสๆ ส่งให้แฟน(5บ/ว) 028927322
รับฟรี!ส่วนลด16-50%EssenceSkll75ml สนใจ กด*48253320011โทรออก(ได้ทุกคน)
ภาพหลุด 2 เกย์เสื้อฟ้าจุ๊บกันในผับเรื่องถึงทางบ้านมาเคลียร์ด่วน
รับฟรี!! วิธีอยู่แบบไม่มีหนี้ตลอดชีวิต สนใจ กด*48253320012 โทรออก
รับฟรี! 108 วิธี สวัสดีความสวย สนใจ กด*48259070011โทรออก

ตารางที่ 5.3 ตัวอย่างข้อความที่เป็น spam (ข้อความขยะ) (ต่อ)

รับฟรี! สารพัดวิธีสวยทันตา สนใจ กด*48259070011โทรออก
รับฟรี!ส่วนลด16-50%EssenceSkll75ml สนใจ กด*48253320011โทรออก(ได้ทุกคน)
ร่วมสมัครVIHOK-DARA โทรที่ DTAC:*19828300 AIS:*453030111ฟรี1เดือน
ฟรี!SMS10เทคนิคเติมเต็มรักกับเคล็ดลับพิชิตใจคู่รักง่ายๆแค่โทร025020469
i-clubลุ้นร่วมกิจกรรมK-pop dance workshopที่Fit Oxygen studioบ้านสี
ลม19ตค.นี้&ซื้อi-mobileรับบัตรทดลองเล่นfitnessฟรี2000ท่านแรกT.02502
พิเศษสุดจริงๆส่งงานFCมากยิ่งได้มากสอบถามเพิ่มเติมด่วน053412299#221แน๊ต
โหลดไม่อั้นๆๆ19บ/เดือน UnlimitedPackสมัครกด *4912 โทรออก ถาม029755577
การโหลดringtone 1.กดปุ่มOK 2.เลือก 3.เมนู เลือก6ไปที่URLกด/ 5.พิมพ์htt
โชว์SMSรับฟรีปากกาที่นาโนสแควร์อุดรฯ,แนะนำลูกค้าใหม่ลด5-15%,0%12เดือน
บินวันนี้ ฟรีค่าธรรมเนียมน้ำมัน จองด่วนวันนี้-14 ธ.ค.ที่airasia.com
COCKPITจัดฟรีคอนเสิร์ต Jennifer Kim&Mr.Sax man โชว์,เกมส์,รางวัล,อา
หารมากมาย @Central Airport 16.30-19.30น. 13ธค.นี้จองด่วน0863031777
สมาชิกค็อกพิทชมฟรีคอนเสิร์ตJennifer Kim 16.30น หรือหนังThe Transpor
ter3 13.00น 20ธค.นี้@Paragon Cineplexจองด่วน022972299รหัสจอง95919
คืนนี้!บอล5คู่ทีเด็ดโดยบอลสเต็ปดอทเนทมีทีมใดบ้างฟังด่วน!โทร1900190043
ภาพถ่ายกิจกรรมสัปดาห์สะพานโหลดได้ที่ \\console\sw\photo
โชว์SMSร่วมกิจกรรมในงานรับของสมนาคุณฟรี ศักดิ์ชัยโซลูชั่น 5-13 ธค.นี้
เม้าท์!ข่าวลือข่าวฮอตให้คุณอัพเดททุกวันsmsซุบซิบดาราฟรี15วัน025020430
ท้ายด่วน เพียงแวะชมรับบัตร Spa ฟรี 2,200 Urban Sathorn 02-869-7788
แม่นมาก!ฟรีSMSดวงรายวันคอนเฟิร์มทุกเรื่องโดยหมอกฤษณ์คอนเฟิร์มโทร1113
กด*48903082222 เพียงเล่นเกมลุ้นบินไปเกาหลีฟรี!!สอบถาม025026767(3บ/SMS)
ฟรี!smsดวงคุณโดยหมอกฤษฎ์ คอนเฟิร์ม,อ.ลักษณ์ ฟันธง,ปู โลกเบี้ยวโทร1113
ขัดผิวหรืออโรมา2ชม.800บ.แถมนวดศีรษะฟรี สอบถามJustRelaxSpa 029485446-7
ดูดวงสดกับหมอดูชื่อดัง!ตลอดธ.ค.ลุ้นรับทองคำมูลค่า10,000บ.โทร1900190033
ฟรี!! รับโปรโมชั่นโรงแรมทั่วไทยฟรี7วัน กด*48259390011แล้วโทรออก
โชว์บัตรเฟิร์สช้อยส์รับของขวัญฟรี Commart 3-6 ก.ค.ศูนย์ประชุมกาญจนา มข
ับมาที่i-club2008@hotmail.comลุ้นรับi-mobile101ถึง15ก.ค.นี้T.025028
คลิบเซ็กซี่สุดๆจากน้องๆพริตตี้/โคโยตี้ ฟรี7วัน กด*4825911โทรออก
สุดคุ้ม!ใช้ฟรี15วัน โมโนลูกทุ่งฮิต+ผลสลาก โทร*45223111110/027302424
ลูกค้า Krungsri SMS Banking ร่วมรายการ Krungsri Yellow Points ฟรีทั
ายใน2รอบบัญชี รับฟรี กระเป๋า Buddy Bag มูลค่า 1,590 บาท โทร 02-627-
ายใน2รอบบัญชี รับฟรี กระเป๋า Buddy Bag มูลค่า 1,590 บาท โทร 02-627-
ลูกค้า Krungsri SMS Banking ร่วมรายการ Krungsri Yellow Points ฟรีทั
สมัครด่วน! Telesales อบรม 22-23 กค 51 รับแค่ 15 ท่านเท่านั้น...TN
แม่นๆพยากรณ์สดทีมอ.ลักษณ์ เช็คดวงชะตาโทร.1900222405ลุ้นรับพระตรีมูรติ
ฟรี!! รับโปรโมชั่นโรงแรมทั่วไทยฟรี7วัน กด*48259390011แล้วโทรออก
รับSMSกิ๊กๆไว้ส่งให้แฟน+ลุ้นทองสมัครฟรี!กด *298*289# (5บ/ว) 028927322
ฟรี15วันรับเลข3เซียนดังรับประกันแม่นทุกงวดกด*48259400011โทรออก(39บ./ด)
ฟรีกระเป๋ามูลค่า350บ.ที่บูธTMBMoneyExpo 8-11พค.51_โชว์บัตร Ready Cash
้คุณมากกว่าใคร ดอกเบี้ย 0% 1 ปี ฟรีโอน+จดจำนอง โทร 1375

ตารางที่ 5.3 ตัวอย่างข้อความที่เป็น spam (ข้อความขยะ) (ต่อ)

นาตาลี ช๊อค!ภราดรชุกลูกสาวสนใจกด*48259890011ฟรี15วันโทร021003555_39บ/ด
ความแตก!ลูกภราดรโผล่สนใจกด*48259890011ฟรี15วันโทร021003555_39บ/ด
ปู โลกเบี้ยว ดูไพ่ยิปซีแม่นๆแค่โทร.*4958 (9บ/นาที)ลุ้นดูดวงสด/2076888
รับโปรโมชั่นโรงแรมทั่วไทย ฟรี7วัน กด*48259390011โทรออก info021003555
ลุ้นรับดูดวงสดกับปู โลกเบี้ยว โทร.1900222262 ดูไพ่ยิปซี ยูเรเนี่ยนsms
ฟรี!บัตรกำนัลเทสโก้ 200 บาทเมื่อเบิกสินเชื่อขั้นต่ำ 50,000 บาท(มีต่อ)
ด่วน!แจกทองทุกสัปดาห์หนักรวม 5 บาท กด *7654 โทรออก ถาม029388070
คลิปเด็ด! ขาวสวยหมวย X รับฟรี15วันสนใจ กด *4825911 โทรออก39บ./ด
ะดาษทิชชู limited edition พร้อมส่ง sms ลุ้นรางวัลใหญ่ วันนี้ - 31 ก
(พหลโยธินรังสิต) "แคมเปญพิเศษ!ซื้อบ้านแถมรถJazz"ฟรีดาวน์0%ดอกเบี้ยพ
ิเศษลุ้นรับรางวัลเครื่องไฟฟ้าฟรีทุกชม.และสนุกกับคอนเสิร์ตAF4/31 พ.ค
ฟรี!พบกับคลิ๊ดลับใหม่เทคนิคชั้นเชิงให้ชีวิตคู่ยิ่งเร้าใจโทร02-5020469
เคล็ดลับเรื่องรัก ที่คุณควรรู้ สมัครกด*4825915 แล้วโทรออก รับฟรี 15วัน
ฟรี!เคล็ดลับมัดใจคนรักเติมไฟรักของคุณให้ร้อนแรงอยากรู้โทร.02-5020469
กลวิธีพิชิตใจคู่รัก ให้อยู่หมัด กด*4825915 แล้วโทรออก ฟรี 15วัน
สดทุกภาพเสียง! รับMMSข่าวด่วนINNฟรี15วัน โทร*452288811 สอบถาม027302424
ฟรี!! รับโปรโมชั่นโรงแรมทั่วไทยฟรี7วัน กด*48259390011แล้วโทรออก
ฟังข่าวบันเทิงจากปากดารา โทร*4946นาทีละ2บ.ฟรีรายเดือน15วัน/027302500
รับฟรี! บัตรกำนัลเทสโก้โลตัส มูลค่าสูงสุดถึง 2,000 บาท (มีต่อ)
สุดคุ้ม!ใช้ฟรี15วัน โมโนลูกทุ่งฮิต+ผลสลาก โทร*45223111110/027302424
ส่ง SMS ผ่านเว็ปไซต์ไปยังมือถือทุกค่าย ฟรี..ที่ www.cat4sms.com

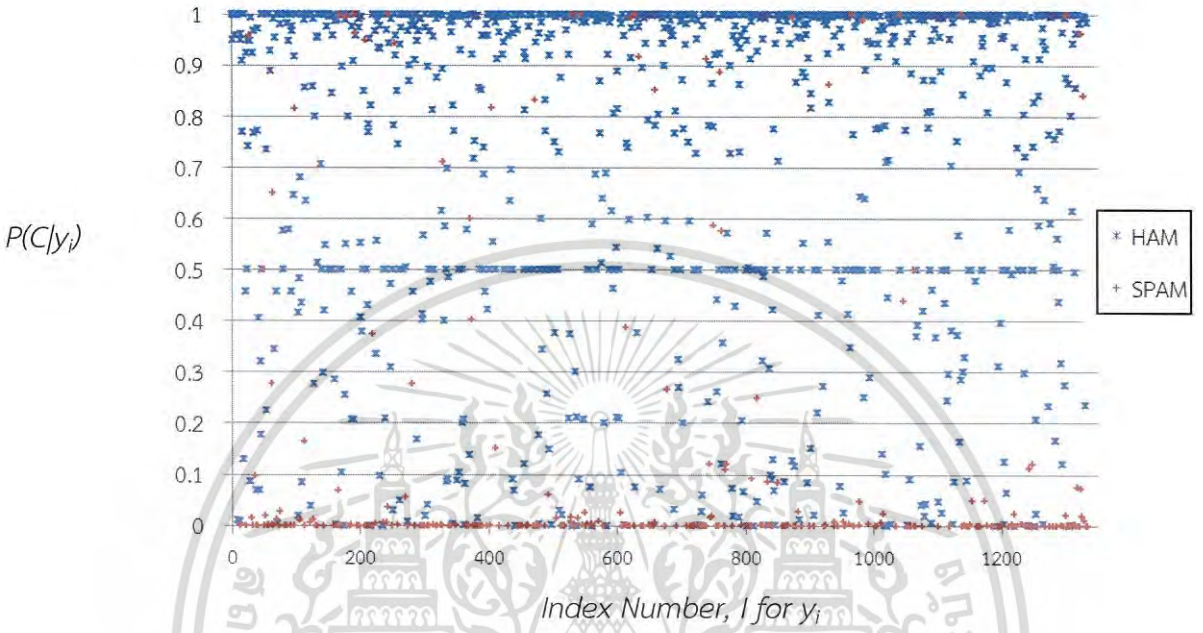
- เมื่อได้ผลลัพธ์จากการจำลองการทำงานของ Content-based filtering ในส่วนนี้แล้วจึงได้นำไป Plot ในกราฟเพื่อแสดงให้เห็นถึงการกระจายตัวของรูปแบบ sms ที่นำมาทดสอบ และได้กำหนดให้ผลลัพธ์ที่ได้ในส่วนนี้เป็นค่า ham หรือ spam เพื่อนำไปใช้ในการวิเคราะห์ร่วมกับการทำงาน Challenge-respone ซึ่งจะแสดงผลได้ดังส่วนที่จะอธิบายต่อไปนี้

5.1 ผลการจำลองการทำงานของการกรองข้อความจากเนื้อหา (CB filtering)

การทดลองการทำงานของตัวกรองเนื้อหา (CB Filtering) ในวิทยานิพนธ์นี้ได้จากการจำลองการทำงานจากโปรแกรมที่สร้างขึ้นมาและติดตั้งบนคอมพิวเตอร์แทนการติดตั้งที่ SMSC หรือ SMS Gateway โดยมีข้อมูลจำนวน 2 ชุดคือ ชุดข้อมูลฝึกสอน (Training Data : TD) จำนวน 1,250 ข้อความ ที่นำตัวอย่างข้อความ Spam ไปฝึกให้มีการเรียนรู้โดยชุดตัวอย่างคำที่เป็น SMS ขยะที่ได้มาจากผลการสำรวจของงานวิจัย และชุดข้อมูลทดสอบชุดใหม่ (New Data : ND) จำนวน 1,336

ข้อความ ที่ผสมระหว่างภาษาไทยและอังกฤษซึ่งนำมาจากระบบบริการ CAT CDMA ของบริษัท กสท โทรคมนาคม จำกัด (มหาชน) ที่ให้บริการรับ-ส่ง SMS อยู่ในปัจจุบัน

เมื่อระบบทำการคัดกรองเนื้อหาของ SMS แล้วและได้ค่า P(c=ham|y) หรือความน่าจะเป็นของ SMS ที่เป็นปกติและขยะ Pr(c=spam|y) ซึ่งสามารถนำมาแสดงได้ดังรูปที่ 5.5



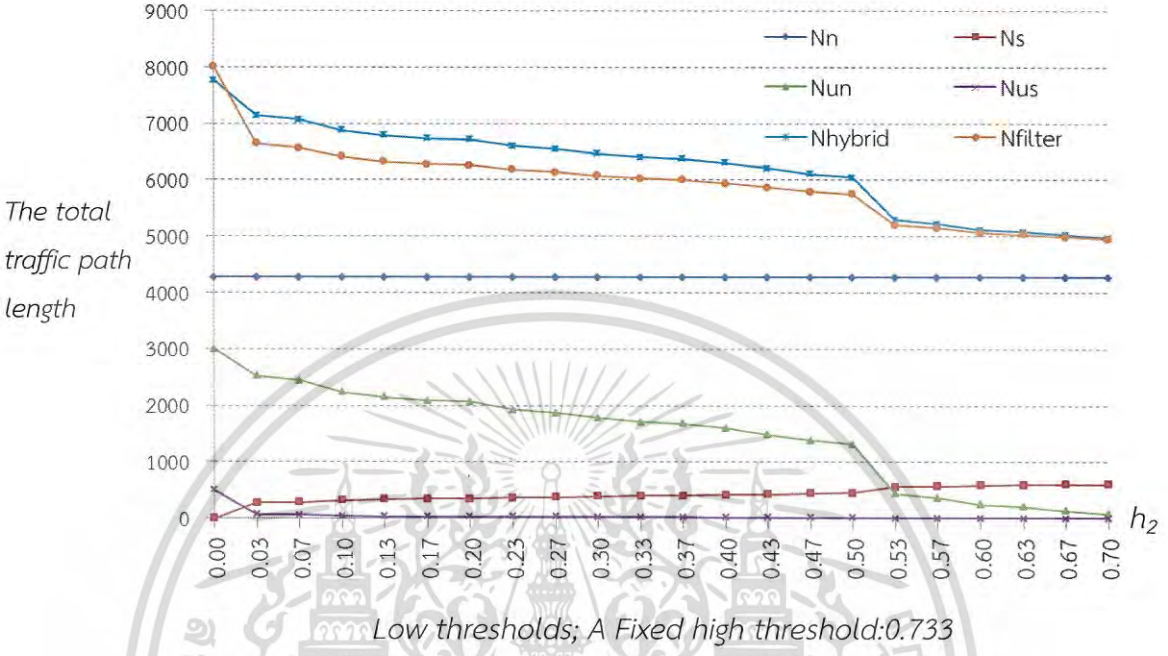
รูปที่ 5.5 ผลการจำลองการทำงานการกรองข้อความจากเนื้อหา (CB Filtering)

จากรูปที่ 5.5 แสดงผลการจำลองการที่ได้จากการกรองข้อความจากเนื้อหา โดยแกน x แสดงลำดับโดยการสุ่มของข้อความ SMS ที่ได้เข้านำมาทดสอบจำนวน 1,336 ตัวอย่าง และแกน y แสดงความน่าจะเป็นของ SMS ที่ถูกคัดกรอง หรือ Pr(c=ham|y) โดยมีค่าตั้งแต่ 0 ถึง 1 ซึ่งจะเห็นได้ว่าความน่าจะเป็นของ SMS นั้นมีการกระจายตัวทั้งข้อความปกติ (ham) และข้อความขยะ (Spam) และสามารถคัดแยกข้อความปกติออกเป็น 82.2% และข้อความขยะ 17.8%

5.2 ผลการวิเคราะห์ข้อมูลของการกรองข้อความแบบผสม (Hybrid)

จากผลการจำลองการกรองข้อความจากเนื้อนั้น อาจจะมี SMS บางส่วนที่โปรแกรมไม่สามารถแยกแยะว่าเป็นข้อความปกติ หรือข้อความขยะได้ เนื่องจากจุดอ้างอิง (Threshold) ที่ไม่เหมาะสมก็อาจจะทำให้ SMS นั้นถูกคัดแยกผิดประเภทได้ ดังนั้นจึงได้มีการนำค่าของความน่าจะเป็น Pr(c=ham|y) และ Pr(c=spam|y) มาวิเคราะห์ร่วมกับการรับรองจากมนุษย์ เพื่อหาแนวโน้มของความน่าจะเป็นในการคัดแยกข้อความให้มีความถูกต้องมากขึ้น โดยการพิจารณาจากการเปลี่ยนแปลง

Upper threshold (จุดอ้างอิงขอบบน : h_1) และค่า Lower threshold (จุดอ้างอิงขอบล่าง : h_2) และพารามิเตอร์อื่นๆ คือ กำหนด e_1=0.02 และ e_2=0.01 ซึ่งเกี่ยวข้องดังที่กล่าวมาในบทที่แล้วดังนี้



รูปที่ 5.6 ผลการวิเคราะห์การจำลองการทำงานของการกรองข้อความแบบผสม (Hybrid) โดยกำหนด h_1 = 0.733 และเปลี่ยนแปลง h_2 ตั้งแต่ 0 - 0.7

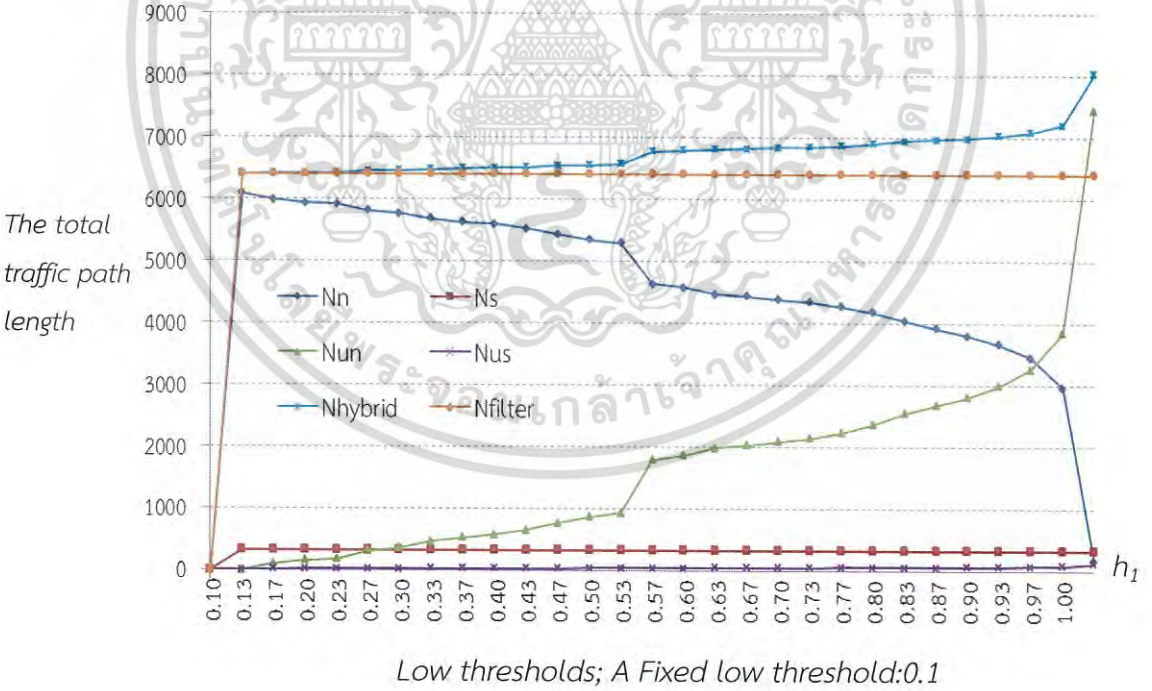
จากรูปที่ 5.6 ผลรวมของ Traffic path length ในการกรองข้อความแบบผสมนั้นจะอธิบายตัวอักษรที่กำกับในแต่ละส่วนเพิ่มเติมดังนี้

- N_n = Ham
- N_s = Spam
- N_{un} = Ham in uncertain region
- N_{us} = Spam in uncertain region
- $N_{Hybrid} = N_{un} + N_{us} + N_s$
- $N_{Filtering}$ = CB Filtering only

จากรูปที่ 5.6 แสดง Traffic path ของระบบการรับ-ส่ง SMS ในการกรองข้อความแบบผสม (Hybrid) โดยแกน x แทนค่าของ Low threshold (h_2) ที่มีค่าตั้งแต่ 0 - 0.7 และค่า High

threshold (h_1) มีค่าคงที่ที่ 0.733 และแกน y แทนผลรวมของ Traffic path length เมื่อ h_2 มีค่าเปลี่ยนแปลงไปนั้น จะเห็นได้ว่าเมื่อค่า h_2 มีค่าเพิ่มขึ้นจะส่งผลให้ผลรวมของ Traffic path length ลดลง ทั้งในส่วนของค่าในเขตพื้นที่สีเทา (Uncertain region) จะมีพารามิเตอร์ที่บ่งชี้การทำงานคือ N_{un} , N_{us} และ N_{Hybrid} เนื่องจาก h_2 นั้นเป็น Lower threshold (จุดอ้างอิงขอบล่าง) ที่บ่งบอกถึง SMS ที่มีแนวโน้มจะเป็น Spam ทำให้ไม่สามารถมีการส่งผ่านเข้าไปในโครงข่ายของ SMSC ได้จึงทำให้ผลรวม Traffic path length ลดลง ซึ่งก็สอดคล้องกับสมมุติฐานจากสมการที่ได้กำหนด

หากพิจารณาถึงการเปลี่ยนแปลงของ h_2 จะเห็นได้ว่าการกำหนดค่า threshold ที่เหมาะสมนั้นมีความสำคัญมาก เช่น หากกำหนด h_2 ที่ค่าต่ำเกินไปก็จะส่งผลให้ SMS นั้นต้องผ่านกระบวนการรับรองจากมนุษย์ที่มาเกินความจำเป็น ซึ่งอาจจะส่งผลต่อเวลาในกระบวนบวนการ-ส่ง SMS ที่ Delay ขึ้นได้ ดังนั้นหากจะนำไปประยุกต์ใช้งานควรศึกษาถึงธรรมชาติหรือคุณลักษณะของข้อความว่ามีกลุ่มคำประเภทใดที่เข้ามาในระบบ รวมถึงคำศัพท์ใหม่ๆ ที่จะต้องเข้าสู่กระบวนการเรียนรู้ และปรับปรุงค่าต่างๆ ให้เหมาะสมกับสภาวการณ์อยู่อย่างต่อเนื่องเพื่อให้การทำงานเกิดประสิทธิภาพสูงสุดตามวัตถุประสงค์ของวิทยานิพนธ์



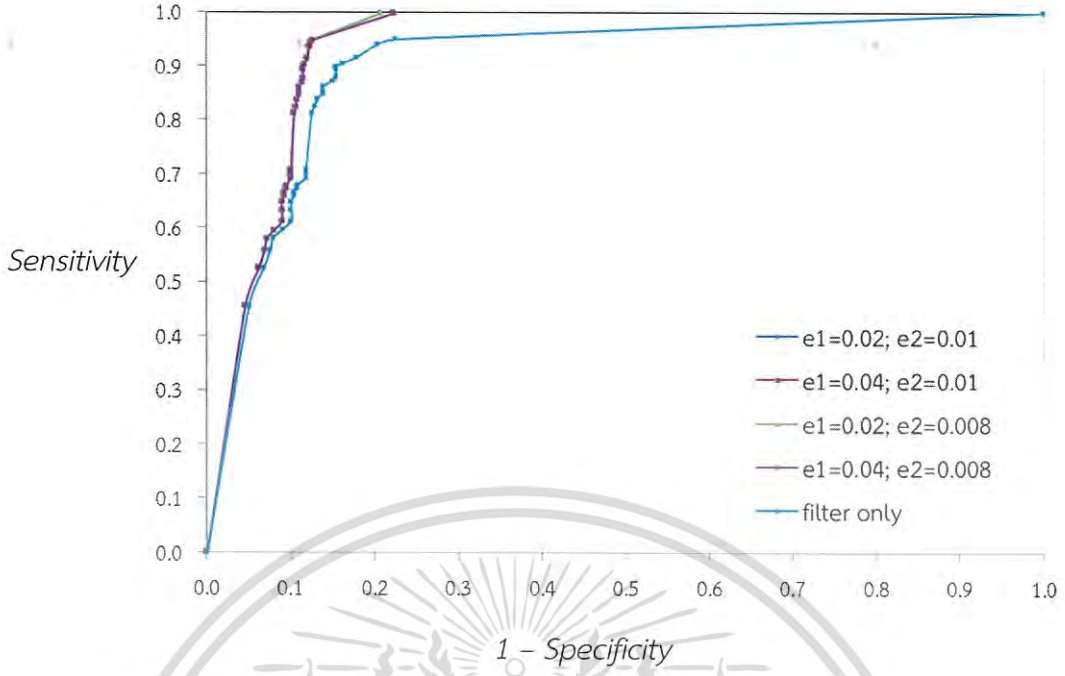
รูปที่ 5.7 ผลการวิเคราะห์การจำลองการทำงานของการกรองข้อความแบบผสม (Hybrid) โดยเปลี่ยนแปลง h_1 ตั้งแต่ 0.1 - 1.0 และกำหนด h_2 = 0.1

จากรูปที่ 5.7 แสดง Traffic path ของระบบการรับ-ส่ง SMS ในการกรองข้อความแบบผสม (Hybrid) โดยแกน x แทนค่าของ High threshold (h_1) ที่มีการเปลี่ยนแปลงตั้งแต่ 0.1 – 1.0 และแกน y แทนผลรวมของ Traffic path length โดยกำหนดค่า Low threshold (h_2) มีค่าเท่ากับ 0.1 จากกราฟจะเห็นได้ว่าเมื่อ h_1 มีค่าเพิ่มขึ้นจะส่งผลให้ N_n ในระบบลดลงเนื่องจากค่า h_1 จะเป็นจุดที่อ้างอิงในการบ่งชี้ SMS นั้นๆ ว่าเป็น ham และส่งผลให้ผลรวมของ Traffic path length ที่อยู่ในพื้นที่สีเทา (Uncertain region) มีค่าเพิ่มมากขึ้นเพื่อรอผลการ Challenge - response มาคัดแยกอีกครั้ง โดยจะเห็นได้ว่าทั้ง N_{un} , N_{us} และ N_{Hybrid} ต่างก็มีผลรวมของ Traffic path length เพิ่มขึ้นทั้ง 3 ตัวซึ่งก็สอดสอดคล้องกับความเป็นจริงที่น่าจะเกิดขึ้นในระบบของ SMSC

5.3 ผลการวิเคราะห์ความน่าจะเป็นของค่าความถูกต้องโดยเปรียบเทียบระหว่างการกรองข้อความจากเนื้อหา (CB filtering) และการกรองข้อความแบบผสม (Hybrid)

จากผลการวิเคราะห์ในส่วนที่ผ่านมาจะเห็นถึงผลรวมของ Traffic path length ที่เกิดขึ้นในส่วนต่างๆ ของระบบการกรองข้อความ และเราจะนำผลที่ได้นี้มาวิเคราะห์เปรียบเทียบเทียบถึงแนวโน้มความถูกต้องของาการกรองข้อความทั้ง 2 วิธี เพื่อแสดงให้เห็นถึงข้อดี ข้อเสียของการกรองในแต่ละวิธี โดยแสดงในลักษณะของกราฟ และตารางที่แสดงตัวเลขให้เห็นอย่างเด่นชัด

จาก ROC Curve ที่อธิบายในบทที่ 4 เราจึงนำค่าของผลรวมของ Traffic path length ที่เกิดขึ้นมาร่วมคำนวณ และทดสอบปรับเปลี่ยนค่าพารามิเตอร์อื่นๆ ประกอบ คือค่า e_1 และ e_2 เพื่อศึกษาแนวโน้มของการกำหนดความน่าจะเป็นของการการคัดแยก SMS ที่ผิดพลาดจากทั้งมนุษย์และโปรแกรมการเรียนรู้เอง โดยกำหนดให้มีความแตกต่างกัน 4 ค่าคือ e_1 = { 0.02, 0.04} และ e_2 = { 0.008, 0.01} ซึ่งจะได้ผลดังแสดงในรูปที่ 5.8



รูปที่ 5.8 ROC Curve

จากรูปที่ 5.8 แสดง ROC Curve ที่มีเกิดจากการเปลี่ยนแปลงค่าอ้างอิงจาก 0 ถึง 1 โดยแกน x แทนค่าของ 1 – Specificity และแกน y แทนค่าของ Sensitivity จะเห็นได้ว่าเส้นกราฟที่แสดงแทนการทำงานการกรองแบบผสม (Hybrid) มีประสิทธิภาพในการจำลองการทำงานที่ดีกว่า ซึ่งในรายละเอียดของค่าความคลาดเคลื่อน e_1 และ e_2 ที่ค่าแตกต่างกันก็จะทำให้ผลลัพธ์มีค่าแตกต่างกันด้วยเช่นกัน และจากกราฟ ROC Curve เราสามารถพิจารณาในอีกรูปแบบหนึ่งที่เข้าใจได้ง่ายคือการพิจารณาพื้นที่ใต้ ROC Curve หรือ Area Under the ROC Curve (AUC) [18] ที่บ่งบอกถึงความน่าจะเป็นของพื้นที่ที่มากกว่าจะสามารถคัดแยกข้อความที่ถูกต้องมากกว่า เราสามารถทำการหาค่าเฉลี่ยพื้นที่ใต้ ROC Curve ดังแสดงในตารางที่ 5.4

ตารางที่ 5.4 แสดงการเปรียบเทียบ AUC

Method	Ratio (e_1)	Ratio (e_2)	AUC
Filter only	-	-	0.9022
Hybrid	0.02	0.01	0.9351
Hybrid	0.04	0.01	0.9344
Hybrid	0.02	0.008	0.9354
Hybrid	0.04	0.008	0.9347

จากข้อมูลที่แสดงในตารางที่ 5.4 จะเห็นได้ว่า AUC ที่แทนการทำงานการกรองข้อความแบบผสม (Hybrid) นั้นมีค่ามากกว่าการทำงานจากการกรองแบบเนื้อหา (CB filtering) ซึ่งมีค่า AUC เท่ากับ 0.9022 โดยมีค่าแตกต่างสูงสุดเท่ากับ 0.0332 เมื่อเปรียบเทียบกับการกรองแบบ Hybrid ที่มีค่า AUC เท่ากับ 0.9354 ที่ค่า e_1=0.02 และ e_2=0,008 และเมื่อพิจารณาในส่วนของพารามิเตอร์ที่ประกอบการทำงาน e_1 และ e_2 จะเห็นได้ว่า เมื่อค่าของทั้ง e_1 และ e_2 มีค่าลดลงจะทำให้ AUC มีค่าเพิ่มมากขึ้น นั่นหมายถึงหากค่าความคลาดเคลื่อนลดลงก็จะทำให้การกรองข้อความมีการทำงานที่มีความน่าจะเป็นที่ถูกต้องเพิ่มขึ้นด้วยเช่นกัน

บทที่ 6

สรุปผลการวิจัยและข้อเสนอแนะ

บริการข้อความสั้นหรือ Short Message Service (SMS) ถือได้ว่ามีประโยชน์และได้รับความนิยมในการใช้งานในปัจจุบันเป็นอย่างมาก โดยเฉพาะการเป็นอีกหนึ่งสื่อที่มีประสิทธิภาพในการทำการตลาดไปยังผู้บริโภค และเข้าถึงกลุ่มเป้าหมายได้ครั้งละจำนวนมาก จากการที่ผู้ใช้บริการโทรศัพท์เคลื่อนที่ที่มีการยายตัวอย่างต่อเนื่อง โดยจะใช้วิธีการส่ง SMS แบบ Short Message Peer-to-Peer Protocal (SMPP Protocal) ที่สามารถส่งจากหนึ่งผู้ส่งไปยังหลายผู้รับ (One-to-many message) ที่สามารถส่งผ่านระบบอินเทอร์เน็ตผ่าน SMSC ไปยังผู้รับปลายทาง แต่อย่างไรก็ดีปัญหาจากการที่มีการส่ง SMS จำนวนมากเกินไปผ่านเข้ามาที่ SMSC ทำให้เกิดความล่าช้าในการทำงาน หรืออาจจะทำให้ SMSC หยุดการทำงานไปเลย โดยส่วนหนึ่งของ SMS นั้นอาจจะเป็นข้อความขยะหรือ Spam ที่เป็นข้อความที่ไม่ได้เรียกร้องให้ส่ง" (unsolicited message) จำพวกโฆษณาที่ก่อให้เกิดความรำคาญและอาจสร้างปัญหาการล่อลวงให้เสียทรัพย์สิน ดังนั้น SMS ที่เข้ามาในระบบโทรศัพท์เคลื่อนหากมีการคัดกรองข้อความที่เป็นประโยชน์แก่ผู้รับแล้ว ก็ย่อมจะทำให้เกิดการทำงานของ SMSC ที่มีประสิทธิภาพสูงสุด

ในสถานการณ์ส่วนใหญ่ผู้ส่งข้อความจะเป็นผู้จ่ายค่าบริการ แต่ยังมีกรณีที่ผู้รับข้อความเปิด international roaming ไปต่างประเทศและใช้งานเครือข่ายที่คิดค่าบริการในการรับข้อความ หากมีการปล่อยให้ spam ส่งออกไปก็จะทำให้ผู้รับข้อความต้องเสียค่าบริการโดยไม่จำเป็นได้ อย่างไรก็ดีในกรณีที่ผู้รับไม่ต้องเสียค่าบริการก็ตาม การที่ปล่อยให้มีข้อมูล spam จำนวนมากส่งผ่านโครงข่ายของผู้ให้บริการจนถึงระดับที่กระทบกับคุณภาพการให้บริการ ตัวอย่างเช่น ผู้ใช้งานปกติจากโทรศัพท์มือถือไม่สามารถใช้บริการส่งข้อความตามปกติได้ หรือทำให้ระบบส่งข้อความไปถึงผู้รับปลายทางได้ช้า ก็จะทำให้คุณภาพบริการของให้บริการลดลงส่งผลให้ผู้ใช้บริการหันไปใช้บริการกับรายอื่นได้ ดังนั้น spam ส่งผลกระทบได้ทั้งกับผู้ใช้บริการ และผู้ให้บริการ (Mobile operator)

วิทยานิพนธ์ฉบับนี้จึงได้นำเสนอวิธีการกรองข้อความสั้นในระบบรับ-ส่งข้อความ ที่เป็นการทำงานร่วมกันระหว่างการกรองจากเนื้อหาข้อความ (CB filtering) และแบบผสม (Hybrid) โดยทำงานร่วมกับการรับรองจากมนุษย์ เพื่อคัดแยกข้อความที่มีลักษณะเป็นขยะหรือ Spam แนวคิดในการทำงานแบ่งเป็น 2 ขั้นตอน คือ ขั้นตอนแรกคัดแยกว่าเป็น spam หรือ ham โดยอาศัยค่า threshold แต่แทนที่จะมีค่า threshold เดียว ในวิทยานิพนธ์ฉบับนี้ใช้ threshold สองค่า คือ ค่า threshold บน และค่า threshold ล่าง ซึ่งจะแบ่งข้อความเป็น 3 กลุ่ม (จากเดิมที่แบ่งเพียง 2 กลุ่ม) คือ กลุ่มเข้าข่าย ham มาก ๆ กลุ่มเข้าข่าย spam มาก และกลุ่มในพื้นที่สีเทา จากนั้นนำกลุ่มที่ตกอยู่

ในพื้นที่สีเทา (Uncertain region) ที่มีโอกาสตัดสินผิดพลาดได้สูงด้วยวิธีการ content-based อย่างเดียวมาเข้าขั้นตอนกรองแบบที่สอง ทำงานร่วมกับการรับรองจากมนุษย์ (Human Intervention) โดยวิธีการถามตอบ (Challenge-response) ซึ่งอาจจะใช้กระบวนการ (Completely Automated Public Turing Computer and Humans : CAPTCHA) จึงเป็นอีกวิธีการที่นำเสนอซึ่งในวิทยานิพนธ์จะเรียกอีกอย่างว่า Hybrid เพื่อให้เกิดความน่าจะเป็นในการคัดแยก SMS ที่ถูกต้องมากยิ่งขึ้น เพราะฉะนั้นประสิทธิภาพของระบบเองจะขึ้นกับความเหมาะสมในการเลือกค่า threshold (h_1 และ h_2) ทั้ง 2 ค่าเป็นหลัก และก็ขึ้นอยู่กับลักษณะของ spam ที่เกิดขึ้นในระบบระหว่างช่วงหนึ่ง ๆ ซึ่งอาจจะต้องปรับปรุงค่า threshold รวมถึงการปรับปรุงข้อความขยะที่ใช้ในการเรียนรู้เป็นระยะตามรูปแบบของ spam ที่เปลี่ยนไป

เอกสารอ้างอิง

[1] He P., Sun Y., Zheng W. and Wen X. "Filtering short message spam of group sending using CAPTCHA." In : Workshop on knowledge discovery and data mining., 2008, Pp. 558–61

[2] Gordon V. Cormack, Jose Maria Gomez Hidalgo and Enrique Puertas Sanz. "Content base SMS spam filtering."in Proc. ACM Symposium on document engineering, 2006, Pp.114-122

[3] Androutsopoulos I., Koutsias J., Chandrinos K. and Spyropoulos CD. "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages." In Proc. 23rd annual international ACM SIGIR conference on research and development in information retrieval., New York, USA, 2000, Pp. 160–167

[4] Shirali-Shahreza S., Movaghar A. "An anti-SMS-spam using CAPTCHA." In Proc. of the 2008 ISECS international colloquium on computing, communication, control and management., IEEE Computer Society, Washington, DC, USA, 2008, Pp. 318–321

[5] Giovanni Camponovo, Davide Cerutti."The spam issue in mobile business a comparative regulatory overview" in Proc. of the Third International Conference on Mobile Business, M-Business 2004

[6] ผศ. ดร. กานดา รุณนะพงศา, นางสาวปโยธร อุราธรรมกุล."การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่"ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น,2006

[7] Nivet C., Prarinya Sa. and Phayung M. "A comparative study on feature weight in Thai document categorization framework" งานวิจัยภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า พระนครเหนือ, หน้า 257-266

[8] Mita K., Mukesh A. "Automatic text classification:A technical review" in Proc. Of International Journal of Computer Application, India, 2011, Pp. 37-40

[9] Vidhya K., Aghila G. "A survey of naive bayes machine learning approach in text document classification" in Proc. Of International Journal of Computer Science and Information Security, India, 2010, Pp. 206-211

[10] Lingling Yuan "An improved naive bayes text classification algorithm in Chinese information processing" in Proc. Of 3rd International symposium on computer science and Computational technology (ISCSCT), Jiaozuo, China, 2010, Pp. 267-269

[11] Andrew Mc. Callum, Kamal Nigam "A comparison of event model for naive bayes text classification" Pittsburgh, PA

[12] ผศ.สุนีย์ สัมมาทัต, นายกฤษฎา เหล็กดี และนายพิษณุ ทองขาว. "การประยุกต์ใช้ตัวแบบเบย์สำหรับวิเคราะห์ความเสี่ยงการถอนรายวิชาแคลคูลัสสำหรับวิศวกร 2" งานวิจัยคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร วิทยาเขตพระนครเหนือ, 2553

[13] ชลธิศา พลทองมาก, พุธษดี ศิริแสงตระกูล. "การวิเคราะห์ความเสี่ยงการเป็นโรคไวรัสตับอักเสบซี โดยต้นไม้การตัดสินใจและทฤษฎีเบย์เซี่ยน" งานวิจัยภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 2553

[14] Roman R., Zhou J. and Lopez J. "An anti-spam scheme using pre-challenges" Computer Communications, 2006, 29(15), Pp. 2739–49

[15] Chaiyaporn K., Nont B. "Short message service filtering for Thai and English language on mobile phone network" in Proc. The 5th National conference on computing and information technology, Bangkok, 2009, Pp.436-442

[16] Prieto AG., Cosenza R. and Stadler R. "Policy-based congestion management for an SMS Gateway" in Proc. the fifth IEEE international workshop, 2004

[17] Yan J, El Ahmad AS. "Usability of CAPTCHAs or usability issues in CAPTCHA design" in Proc. ACM the 4th symposium on usable privacy and security, New York, USA, 2008, Pp. 44–52

[18] Ji Won Yoon, Hyoungshick Kim and Jun Ho Huh. "Hybrid spam filtering for mobile communication" computers & security, 2010, Pp. 446–459

ภาคผนวก

ภาคผนวก ก.

1. งานวิจัยที่เกี่ยวข้อง

อำนาจ ละมัยกลาง, สุวิพล สิทธิชีวภาค "การวิเคราะห์รูปแบบการกรองข้อความสั้น จากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่", ตีพิมพ์ใน วิศวสารลาดกระบัง, ปีที่ 29, ฉบับที่ 4, ธันวาคม 2555, หน้า 13-18

ISSN 0125-1724

LADKRABANG ENGINEERING JOURNAL

วิศวสารลาดกระบัง

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang

ปีที่ 29 ฉบับที่ 4 ธันวาคม 2555

บทความวิชาการ

บทความวิจัย

http://www.kmitl.ac.th/lej

การวิเคราะห์รูปแบบการกรองข้อความสั้นจากเนื้อหาร่วมกับการรับรองจากมนุษย์สำหรับการสื่อสารโทรศัพท์เคลื่อนที่

Analysis of Content Base & Human Intervention SMS Spam Filtering Model for Mobile Communication

อำนาจ ละมัยกลาง สุวิพล สิทธิชีวภาค
สาขาวิชาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

บริการข้อความสั้นหรือ Short Message Service (SMS) ได้รับความนิยมในการใช้งานเป็นอย่างมากในช่วงหลายปีที่ผ่านมาจึงทำให้จำนวนข้อความขยะเพิ่มขึ้นซึ่งส่งผลกระทบต่อประสิทธิภาพการทำงานของระบบศูนย์กลางบริการข้อความสั้นหรือ Short Message Service Center (SMSC) อย่างไรก็ดีเราสามารถควบคุมจำนวนข้อความขยะได้ด้วยระบบการกรองข้อความ บทความฉบับนี้จึงนำเสนอรูปแบบการกรองข้อความสั้นที่ผสมผสานระหว่างการตรวจสอบเนื้อหาและการรับรองจากมนุษย์ (CAPCHA) ของข้อความที่ไม่ได้ถูกจำแนกสถานะ หากไม่มีผลการตอบกลับแสดงว่าข้อความสั้นถูกส่งจากแหล่งที่สร้างข้อความขยะดังนั้นข้อความสั้นจึงไม่ถูกจัดส่งไปยังผู้รับปลายทาง เนื้อหาของบทความฉบับนี้จะอธิบายรูปแบบและกระบวนการทำงานที่สามารถจำแนกข้อความขยะ ผลการวิเคราะห์ทำงานของรูปแบบดังกล่าวพบว่ามีความน่าจะเป็นที่จะคัดแยกถูกต้อง 0.9354 ในขณะที่รูปแบบที่ทำการตรวจสอบเนื้อหาเพียงอย่างเดียวมีความน่าจะเป็นที่จะคัดแยกได้ถูกต้อง 0.9022 ดังนั้นรูปแบบที่นำเสนอนี้จะส่งผลให้ SMSC สามารถทำงานได้อย่างเต็มประสิทธิภาพมากยิ่งขึ้น

คำสำคัญ: ข้อความสั้น, ข้อความขยะ, ระบบศูนย์กลางบริการข้อความสั้น, การกรองข้อความ, การรับรองจากมนุษย์

Abstract

Short Message Service (SMS) has been the most popular means of mobile communication in recent years and hence the spam is an increasing threat to Short Message Service Center (SMSC) efficiency. The spam threat can be controlled through efficient and robust SMS filtering systems. In this paper we present new model that is a combination of content-base (CB) filtering and human intervention (CAPCHA). A message, that has been classified as uncertain by CB filtering, is further checked by sending a challenge to the message sender. An automated spam generator is unlikely to send back a correct response, in which case, the message is classified as spam and don't deliver to recipients. Based on this formulation, results show that our framework achieved a higher accuracy of 0.9354 comparing to those of content-based filtering at 0.9022 consequently, promoted efficiency of SMSC operation.

Keywords : SMS, Spam, SMSC, SMS Filtering, Human intervention

1.บทนำ

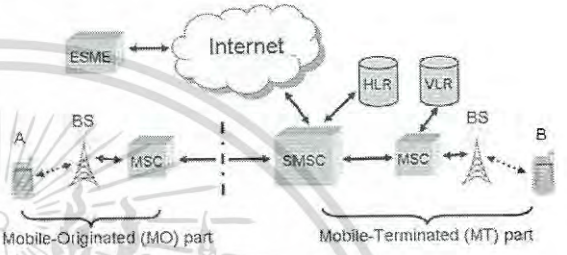
บริการส่งข้อความสั้นหรือ Short Message Service (SMS) รวมถึงบริการข้อความสื่อหรือ Multimedia Message Service (MMS) ได้รับความนิยมและเป็นที่แพร่หลายในการใช้บริการเป็นอย่างมากบนเครือข่ายโทรศัพท์เคลื่อนที่ (Mobile - Communication) เช่น เป็นเครื่องมือในการสื่อสารการตลาด, เป็นช่องทางในการตลาดแบบทางตรง, สร้างธุรกิจบริการเสริมหรือ Value Added Service (VAS)

เนื่องจากข้อความสั้นมีข้อดีหลายอย่าง เช่น เป็นสื่อที่มีประสิทธิภาพ, ค่าใช้จ่ายต่อข้อความมีแนวโน้มลดลง, สามารถกระตุ้นการรับรู้ได้ทันที ข้อความสั้นต่างๆ ที่ใช้ในบริการนี้อาจจะมีส่วนที่ถูกจัดกลุ่มเป็นข้อความขยะ (SMS Spam) ปะปนเข้ามาซึ่งจากการศึกษาในประเทศเกาหลีใต้และญี่ปุ่นนั้นพบว่ามีข้อความขยะสูงถึง 50% ของการใช้งาน ซึ่งส่งผลกระทบต่อประสิทธิ ภาพของศูนย์กลางการรับส่งข้อความหรือ Short Message Service Center (SMSC) ที่ต้องรับภาระการทำงานเกินความจำเป็น

Spam SMS คือข้อความสั้นที่ไม่ได้เรียกร้องให้ส่งหรือ Unsolicited Message ที่ก่อให้เกิดความรำคาญแก่ผู้ใช้และอาจสร้างปัญหาการล่อลวงให้เสียทรัพย์สินทางโทรศัพท์มือถือ เช่น ข้อความโฆษณาขายสินค้า-บริการ การหลอกล่อให้ผู้รับทำกิจกรรมบางประเภทที่เกิดความเสียหาย ข้อความหลอกลวง (Phishing) เป็นต้น หรือ Spammer อาจจะใช้ Robot Software เข้ามาช่วยในการส่ง Spam SMS ที่มีลักษณะการส่งครั้งละหลายข้อความและหลายปลายทางในครั้งเดียว ปัจจุบันได้มีมาตรการป้องกันต่างๆ เช่น การลงทะเบียนไม่ขอรับข้อความโฆษณาจากผู้ให้บริการ การใช้ Software กรองที่เครื่องโทรศัพท์ การใช้Software กรองที่ฝั่งเซิร์ฟเวอร์ เป็นต้น

การแก้ปัญหาด้วยซอฟท์แวร์การกรองข้อความขยะบนเครือข่ายโทรศัพท์เคลื่อนที่ที่ฝั่งเซิร์ฟเวอร์มีหลายหลายวิธี เช่น Bogofilter, DMC, LR, SVM [1] ซึ่งล้วนแต่มีการพัฒนาต่อเนื่องจากพื้นฐานการกรองอีเมลขยะ(E-Mail Spam Filtering) คือ การตรวจเนื้อหาและการจำแนก SMS นั้นว่าเป็นขยะหรือไม่ สำหรับงานวิจัยชิ้นนี้ได้นำวิธีการ Naive Bayesian [2] ที่มีการตรวจจับคำหรือวลีสำคัญซึ่งเป็นส่วนประกอบหนึ่ง SMS และเสนอรูปแบบการทำงานร่วมกันระหว่างการกรองข้อความสั้นจากเนื้อหา (Content-base) กับการรับรองของมนุษย์ที่ได้จากการถามตอบ (Challenge-response) ซึ่งเป็นส่วนหนึ่งของกลไกลอัตโนมัติที่มีวัตถุประสงค์เพื่อความปลอดภัยเนื่องจากการโจมตีจะใช้สิ่งที่เรียกว่า "บอตส์" (bots) ที่สร้างขึ้นจากคอมพิวเตอร์ แต่คอมพิวเตอร์ไม่สามารถแก้ปัญหาการทดสอบด้วย CAPTCHA ได้ ซึ่งมนุษย์เท่านั้นที่เพ่งดูกราฟฟิคและแกะตัวอักษรออกมาเพื่อพิมพ์ยืนยันรับรองเหล่า SMS นั้นๆ จากนั้นจึงนามน่าจะเป็นของการ SMS จาก CB ที่ได้มา ซึ่งอยู่ในรูปแบบความน่าจะเป็นไปวิเคราะห์แนวโน้มของการส่งผ่านข้อมูล (Traffic Path) ในกรณีต่างๆ ที่จะเกิดขึ้นได้ตามสมมติฐาน ภายใต้การจำลองการทำงานจากกลุ่มตัวอย่างของ SMS ที่ได้จากระบบโทรศัพท์เคลื่อนที่ที่ใช้งานจริงว่ารูปแบบการกรองข้อความที่นำเสนอนี้มีช่วยเพิ่มการกรองข้อความให้มีความถูกต้องมากขึ้น

2. ระบบการรับ-ส่งข้อความ



รูปที่ 1 โครงสร้างการรับ-ส่งข้อความ

รูปที่ 1 แสดงพื้นฐานของการรับ-ส่งข้อความในระบบโทรศัพท์เคลื่อนที่ประกอบด้วย 2 ส่วนสำคัญคือ

1.ผู้ส่ง (Mobile Originating : MO) ซึ่งรวมถึงเครื่องโทรศัพท์มือถือที่ใช้ส่ง, สถานีฐาน (Base Station : BS) และชุมสายโทรศัพท์มือถือ (Mobile Switching Center : MSC) ที่ทำหน้าที่ค้นหาเส้นทางและเชื่อมต่อสัญญาณต้นทาง-ปลายทาง

2. ผู้รับ (Mobile Terminating : MT) ซึ่งรวมสถานีฐาน (BS) และชุมสายโทรศัพท์มือถือ (MSC) ในส่วนปลายทาง นอกจากนี้ยังมีส่วนที่สำคัญคือ SMSC ที่ควบคุมระบบการรับส่ง SMS ทั้งหมดโดยจะได้รับข้อมูลตำแหน่งของผู้รับจาก Home Location Register (HLR) และ Visitor Location Register (VLR)

นอกจากนี้ระบบ SMS ยังสามารถรองรับการส่งอยู่อีกประเภทคือ External Short Message Entities (ESMEs) ที่เป็นการส่งจากหนึ่งผู้ส่งไปยังหลายผู้รับ (One-to-many message) ซึ่งได้ถูกนำมาใช้งานอย่างแพร่หลายในด้านการตลาดและบันเทิงเพราะคุ้มค่าจากการส่ง SMS ไปยังกลุ่มเป้าหมายได้ครั้งละจำนวนมาก เรียกการเชื่อมต่อแบบนี้ว่า Short Message Peer-to-Peer Protocal (SMPP Protocal) ซึ่งสามารถใช้งานผ่าน Internet ที่อาจจะถูก bot ใช้งานได้อย่างง่ายด้วยเช่นกัน

SMPP เป็น Protocal มาตรฐานในการรับ-ส่งข้อมูล SMS, MMS หรือ Push Message ภายในระบบ

โทรศัพท์เคลื่อนที่ซึ่งประกอบด้วย Protocal Description Unit (PDU) 2 ส่วนสำคัญดังตารางที่ 1 คือ ส่วนที่ 1 PDU Header ที่ใช้ระบุความยาวชนิดและลำดับของข้อความ ส่วนที่ 2 PDU Body ใช้บรรจุข้อมูลที่ต้องการส่ง เช่น เนื้อความ, ลิงค์สำหรับ Push Message เป็นต้น

ตารางที่ 1 รูปแบบของ SMPP PDU

SMPP PDU				
PDU Header (mandatory)				PDU Body (Optional)
Command length	Command id	Command status	Sequence number	PDU Body
4 octets	4 octets	4 octets	4 octets	Length*
4 octets	Command Length – 4			

Length*= (Command Length value - 16) octets

3. การกรองข้อความจากเนื้อหาร่วมกับการรับรองจาก มนุษย์ (Hybrid:การกรองแบบผสม)

รูปที่ 2 รูปแบบการกรองข้อความแบบผสม CB และ CAPCHA

3.1 การกรองข้อความจากเนื้อหา (CB Filtering)

การจำแนกประเภทข้อความด้วยวิธีการตรวจสอบเนื้อหาหรือ Content-Base (CB) นิยมใช้ Naive Bayesian ที่เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพเหมาะสมกับกรณีของเซตตัวอย่างมีจำนวนมากและมี Attribute ของตัวอย่างที่ไม่ขึ้นต่อกัน มีการนำไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification), การวินิจฉัย (Diagnois) ซึ่งพบว่าใช้งานได้ดี

กำหนดให้ตัวแปรสุ่ม y แทนกลุ่มข้อมูลข้อความ ที่มี Attribute ทั้งหมด n ตัว สามารถหาค่าความน่าจะเป็นของข้อความปกติแทนด้วย *ham* หรือ $Pr(c=ham|y)$ เขียนได้ดังนี้

$$p_r(c = ham \mid y) = \frac{p(ham)}{p(y)} \prod_i^n p(w_i = ham \mid y) \quad (1)$$

โดยที่ c คือความน่าจะเป็นของการจำแนกประเภทข้อความ

w_i คือลำดับของคำในข้อความ

ในทำนองเดียวกันความน่าจะเป็นของข้อความที่มีโอกาสเป็น SMS ขยะแทนด้วย *spam* หรือ $Pr(c=spam|y)$ เขียนได้ดังนี้

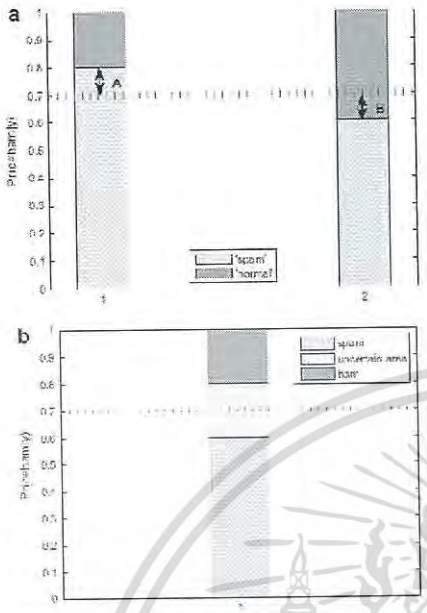
$$p_r(c = spam \mid y) = \frac{p(spam)}{p(y)} \prod_i^n p(w_i = spam \mid y) \quad (2)$$

3.2 การพิจารณาพื้นที่สีเทา (Uncertain Region)

ปกติระบบการกรองจาก CB สามารถจำแนกผลการทำงานได้เป็น 2 แบบคือ *ham* และ *spam* หากกำหนดความน่าจะเป็นของการกรองแทนด้วยการกระจายตัว $Pr(c=ham|y)$ แทนความน่าจะเป็นของข้อความที่อยู่ใน *ham region* โดย c และ y แทนตัวแปรสุ่มประเภทข้อความและข้อความตามลำดับ กำหนดอัตราส่วนที่ใช้ชี้วัดข้อความนั้นๆ ด้วย $O_{post} = Pr(c=ham|y)/Pr(c=spam|y)$ ถ้า $O_{post}>1$ แสดงว่าข้อความถูกจัดให้อยู่ใน *ham* และกรณีอื่นข้อความจะถูกจัดให้อยู่ใน *spam* เมื่อพิจารณาจุดอ้างอิง (Threshold-base) เพิ่มเพื่อใช้เป็นจุดแบ่งแยก จากกการที่ $Pr(c=ham|y)$ มีค่าเข้าใกล้ 1 แล้วข้อความน่าจะถูกจัดอยู่ใน *ham* และหากมีค่าเข้าใกล้ 0 จะถูกจัดอยู่ใน *spam* กำหนด $\bar{c} = f(y,h)$ เป็นค่าการกรองของ CB เมื่อ $\bar{c}$ คือเอาท์พุต และ h คือจุดอ้างอิง ดังนั้นเมื่อแทนค่าตัวแปรต่างๆ แล้วตัวกรองสามารถทำงานได้โดยใช้สมการดังนี้

$$\bar{c} = f(y,h) = \begin{cases} ham & if \quad Pr(c=ham|y) \geq h \\ spam & if \quad Pr(c=ham|y) < h \end{cases} \quad (3)$$

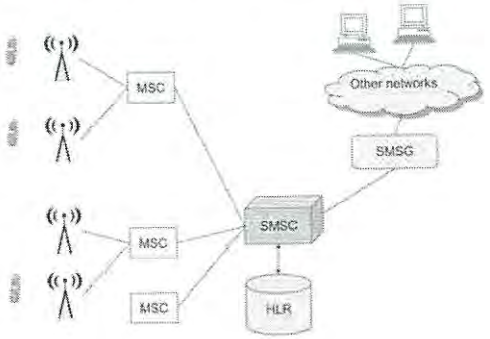
หากกำหนดจุดแบ่งแยกที่มีค่า $h=0.5$ อาจจะส่งผลทำให้เกิดปัญหาในการจำแนกประเภทข้อความได้ ดังนั้นเราจึงได้มีการนำขอบบนและขอบล่าง (Upper and Lower Boundaries) มาใช้ช่วยพิจารณาเพื่อแก้ไขจุดอ่อนของตัวกรอง CB ด้วยพื้นที่สีเทา (Uncertain Region) ดังแสดงในรูปที่ 3b



รูปที่ 3 (a) กรณีที่ 1 $h > \tilde{h}$ และกรณีที่ 2 $h < \tilde{h}$ โดยกำหนดค่าอ้างอิงจริง $\tilde{h}$ แทนด้วยเส้นประ (b) ปรับปรุงโดยการเพิ่มพื้นที่สีเทาและกำหนดค่าอ้างอิงจริง $\tilde{h}$ แทนด้วยเส้นประ

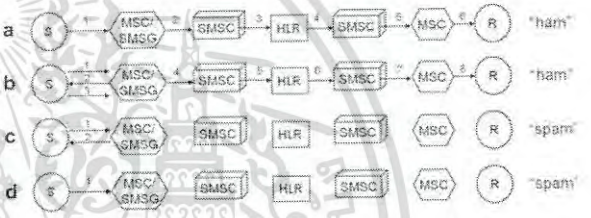
3.3 โปรโตคอลการถามตอบ:Challenge-response protocol

วิธีการจำแนกประเภทข้อความที่อยู่ในพื้นที่สีเทาว่าตกอยู่ในช่วงบวกและลบ (False Positive and False Negative) นิยมใช้กระบวนการ CAPTCHA [3] โดยตรวจสอบรูปแบบที่เข้ากันซึ่งผู้ใช้สามารถยืนยันได้ ส่วน SMS ขยะจากคอมพิวเตอร์หรือ bot อาจสร้าง SMS ขยะจำนวนมากได้แต่ไม่สามารถยืนยันตัวเองและตอบจากข้อความภาพเงื่อนไขที่แสดงได้ ดังนั้นจึงกำหนดเมื่อมีการถามตอบที่ถูกต้องแสดงว่ามีความน่าจะเป็นสูงที่ SMSนั้นถูกส่งจากผู้ใช้ รูปแบบสื่อกลางของ CAPTCHA ที่สามารถปรับให้เหมาะสมกับการใช้งานได้ เช่น ภาพ เสียง หรือ ตัวอักษร เป็นต้น โดยจะเรียกวิธีการนี้ว่าการกรอง SMS แบบผสม (Hybrid)



รูปที่ 4 โครงสร้างการกรองข้อความแบบผสม (Hybrid)

โครงสร้างที่ออกแบบดังแสดงในรูปที่ 4 จำลองการทำงานโดยกำหนดผู้ส่ง (S), ผู้รับ (R), ศูนย์กลางบริการข้อความ (MSC หรือ SMSG) และส่วนประกอบอื่นๆ เช่น SMSC, HLR เป็นต้นโดยกำหนด $y^{h}_{c=Type}$ สำหรับ $type \in \{ham, spam\}$ แทนด้วยข้อความที่ถูกกรองซึ่งอยู่ในเทอมของ h ดังนั้นปริมาณข้อความรวมคือ $N_{FilteringOnly} = |y^{h}_{c=ham}| \times 6 + |y^{h}_{c=spam}| \times 1$ เมื่อ $|.|$ แทนจำนวนนับของ SMS ซึ่งปริมาณทั้งหมดในระบบ เนื่องจาก *ham* สามารถส่งผ่านไปยังโครงสร้างได้ 6 ส่วน คือ (S-MSC/SMSG-SMSC-HLR-SMSC-MSC-R) และ *spam* ส่งผ่านได้เพียง 1 ส่วนประกอบเท่านั้นคือ (S-MSC/SMSG)



รูปที่ 5 ความเป็นไปได้ของการส่ง SMS ในรูปแบบผสมทั้ง 4 กรณี

สำหรับรูปแบบผสม (Hybride) ที่ SMS ถูกแยกออกเป็น 3 ช่วงโดยใช้ 2 ค่า คือ h_1 และ h_2 จึงประมาณค่าพารามิเตอร์เพิ่มขึ้นมาอีก คือ N_{sa} และ N_{sc} ที่อยู่ในพื้นที่สีเทา ซึ่งสามารถแสดงความเป็นไปได้ของเส้นทางการส่งผ่านข้อมูลดังรูปที่ 5

กรณีที่ 1 (a) SMS ถูกจำแนกเป็น *ham* โดยให้ค่าความน่าจะเป็นสูงกว่าค่าอ้างอิงขอบบน จะถูกส่งไปยังผู้รับปลายทางผ่านอุปกรณ์ต่างๆ มีจำนวนการส่งผ่าน 6 ส่วน คือ S-MSC/SMSG-SMSC-HLR-SMSC-MSC-R

กรณีที่ 2 (b) SMS ถูกแยกอยู่ระหว่างขอบบนและขอบล่างในพื้นที่สีเทา เมื่อได้รับผลตอบที่ถูกต้องแล้ว SMS ถูกจำแนกเป็น *ham* มีการส่งผ่านทั้งหมด 8 ส่วน คือ S-MSC/SMSG-S-MSC/SMSG-SMSC-HLR-SMSC-MSC-R

กรณีที่ 3 (c) SMS ถูกแยกอยู่ระหว่างขอบบนและขอบล่างในพื้นที่สีเทา แต่ไม่ได้รับผลตอบที่ถูกต้อง แล้วข้อความถูกแยกเป็น *spam* มีจำนวนการส่งผ่าน 2 ส่วน คือ S-MSC/SMSG-S

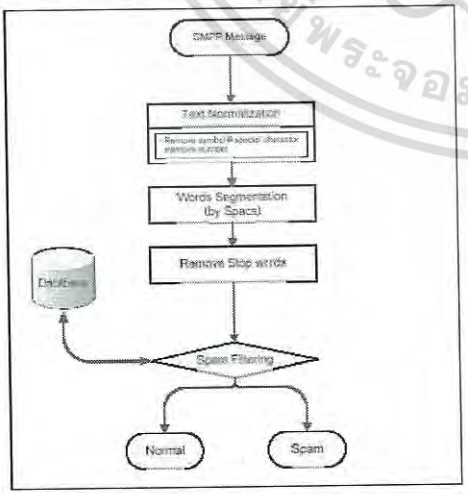
กรณีที่ 4 (d) ข้อความถูกจำแนกเป็นขยะ *spam* ไม่ถูกส่งต่อโดยคัดออกที่ศูนย์กลางบริการข้อความจึงมีจำนวนการส่งผ่าน 1 ส่วน คือ S-MSC/SMSG สามารถคำนวณปริมาณข้อมูลในโครงสร้างดังสมการ

$$
\begin{aligned}
N_n &= |\, y^{h_1}_{c=ham} \,| \times 6 \\
N_{un} &= |\, y^{h_1}_{c=ham} \cap y^{h_2}_{c=spam} \cap y_{c=ham} \,| \times (1-e_1) \times 8 \\
&\quad + |\, y^{h_1}_{c=ham} \cap y^{h_2}_{c=spam} \cap y_{c=spam} \,| \times e_2 \times 8 \\
N_{us} &= |\, y^{h_1}_{c=ham} \cap y^{h_2}_{c=spam} \cap y_{c=spam} \,| \times (1-e_2) \times 2 \\
&\quad + |\, y^{h_1}_{c=ham} \cap y^{h_2}_{c=spam} \cap y_{c=ham} \,| \times e_1 \times 2 \\
N_{hybrid} &= N_n + N_{un} + N_{us} + N_s
\end{aligned}
\qquad (4)
$$

เมื่อ e_1 คือความน่าจะเป็นที่มีผลการตอบสนองกลับอย่างถูกต้อง (ส่งจากผู้ใช้) และ e_2 คือความน่าจะเป็นของขยะที่ถูกส่งจากคอมพิวเตอร์ที่อยู่ในพื้นที่สีเทา

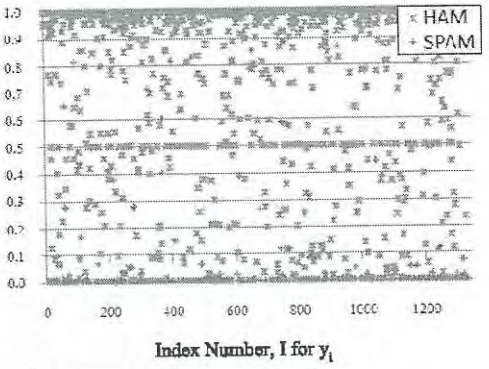
4. การทดสอบและการวิเคราะห์ผลการทดลอง

การทดลองการทำงานของตัวกรองเนื้อหา (CB Filtering) ในบทความฉบับนี้ได้จากการจำลองการทำงานจากโปรแกรมที่สร้างขึ้นมาและติดตั้งบนคอมพิวเตอร์แทนการติดตั้งที่ SMSC หรือ SMS Gateway โดยมีข้อมูลจำนวน 2 ชุดคือ ชุดข้อมูลฝึกสอน (Training Data:TD) ที่นำตัวอย่างข้อความขยะไปฝึกให้มีการเรียนรู้โดยชุดตัวอย่างคำที่เป็น SMS ขยะที่ได้มาจากผลการสำรวจของงานวิจัย [4] และชุดข้อมูลทดสอบชุดใหม่ (New Data:ND) ที่ผสมระหว่างภาษาไทยและอังกฤษซึ่งนำมาจากระบบบริการ CAT CDMA ของบริษัท กสท โทรคมนาคม จำกัด (มหาชน) ที่ให้บริการอยู่ในปัจจุบัน โดยมีขั้นตอนการทำงานดังรูปที่ 6



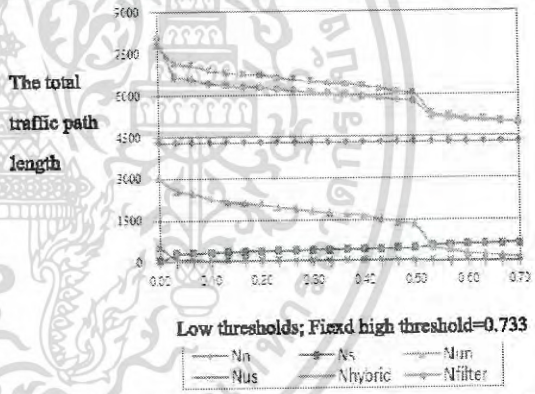
รูปที่ 6 ขั้นตอนการทำงานระบบ CB SMS Filtering

เมื่อระบบทำการคัดกรองเนื้อหาของ SMS แล้วและได้ค่า $Pr(c=ham|y)$ หรือความน่าจะเป็นของ SMS ที่เป็นปกติและขยะซึ่งสามารถนำมาแสดงได้ดังรูปที่ 7

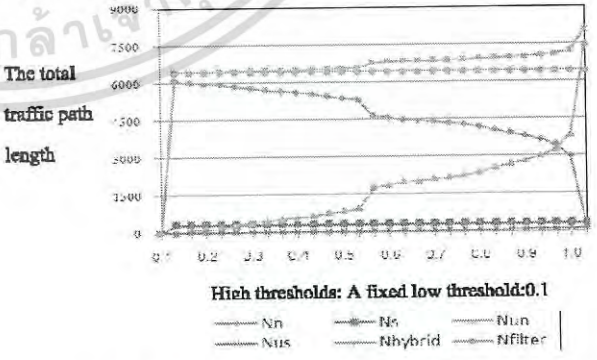


รูปที่ 7) การกระจายตัวของความน่าจะเป็น SMS

ผลจำลองการทำงาน CB Filtering พบว่ามี SMS ที่เป็นปกติ 82.2% และ SMS ขยะ 17.8% หากนำค่าความน่าจะเป็นของ SMS มาพิจารณาโดยใช้วิธีการรับรองรองจากมนุษย์มาเกี่ยวใช้ในการอ้างอิง(Threshold) จะถูกแบ่งออกเป็น 2 ช่วง ดังนั้นเราจึงนำค่า $Pr(c=ham|y)$ ไปใช้เพื่อจำลองผลการวิเคราะห์ดังรูปแบบที่เสนอในบทความฉบับนี้



รูปที่ 8 Traffic Path เมื่อ h_2 คงที่ และ h_1 เปลี่ยนแปลงจาก 0 ถึง 0.733

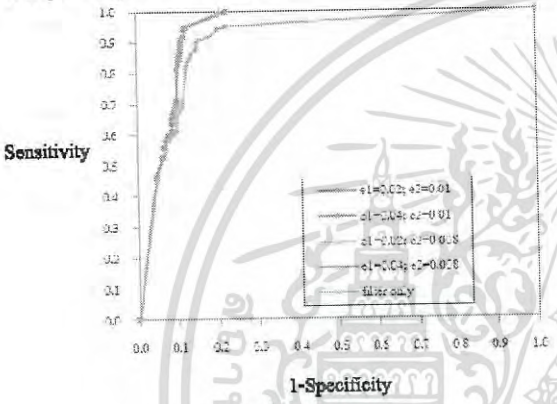


รูปที่ 9 Traffic Path เมื่อ h_1 คงที่ และ h_2 เปลี่ยนแปลงจาก 0.1 ถึง 1

รูปที่ 8 และ 9 แสดงผลของรวมของ Traffic path ที่มีการกำหนดค่า $h_1 h_2$ จากกราฟแสดงให้เห็นว่าการ Traffic usage ลดลงเมื่อค่า h_1 มีค่าเพิ่มมากขึ้นซึ่งก็สอดคล้องกับความเป็นจริงที่เกิดขึ้นคือ h_1 ที่มากขึ้นนั้น SMS จะถูกแยก

เป็น *spam* และส่งผลให้ถูกนำไปสู่รับรองจึงทำให้จำนวน Traffic path ลดลง

อีกตัวชี้วัดหนึ่งเพื่อแบ่งแยก SMS คือ Receiver Operating Characteristic (ROC) Curve ที่ได้นำค่าซึ่งตกอยู่ในพื้นที่สีเทามาเปรียบเทียบ คือ True Positive (TP), True Negative (TN), False Positive (FP) และ False Negative (FN) ที่มีค่าตั้งแต่ 0 ถึง 1 โดยเปรียบเทียบระหว่าง CB Filtering ที่มี h เป็นค่าอ้างอิงเพียงค่าเดียวและแบบผสมที่มีค่า h_1 และ h_2 มาช่วยคัดแยก SMS



รูปที่ 10 ROC Curve

รูปที่ 10 กราฟแสดง ROC Curve ที่บอกถึงแนวโน้มความถูกต้องของการคัดแยก SMS โดยที่แกน x แทนค่าของ 1-*Specificity* และแกน y แทนค่าของ *Sensitivity* ที่เปรียบเทียบผลจากการทำงานของระบบ CB Filtering และ Hybride ซึ่งประมาณค่าจากการสมการ

$$Specificity = \frac{TN}{FP+TN} \quad and \quad sensitivity = \frac{TP}{TP+FN} \qquad (5)$$

พบว่าการกรองแบบผสมนั้นมีค่าความถูกต้องที่สูงกว่าแบบ CB Filtering และเมื่อค่า e_1 และ e_2 มีค่าน้อยลงส่งผลให้เส้น ROC มีค่าเพิ่มขึ้น ซึ่งเมื่อพิจารณาพื้นที่ได้ ROC Curve หรือ Area Under the ROC Curve (AUC) ที่บ่งบอกความน่าจะเป็นของการคัดแยก SMS ที่จะมีค่าเป็นปกติมากกว่า SMS ขยะ

ตารางที่ 2 การเปรียบเทียบค่า AUC

Method	Ratio (e_1)	Ratio (e_2)	AUC
Filter only	-	-	0.9022
Hybrid	0.02	0.008	0.9354
Hybrid	0.04	0.008	0.9347
Hybrid	0.02	0.01	0.9351
Hybrid	0.04	0.01	0.9344

จากตารางที่จะเห็นได้ว่า AUC ของรูปแบบผสมนั้นมีค่า 0.9354 ซึ่งมากกว่าแบบ CB Filtering ที่ 0.9022 โดยเป็นการเพิ่มความน่าจะเป็นที่ระบบจะกรอง SMS ปกติได้ถูกต้องมากขึ้น ค่าของ e_1 และ e_2 ที่เพิ่มขึ้นส่งผลให้ AUC มีค่าลดลง

5. สรุปผลการทดลอง

บทความนี้เสนอวิธีการกรอง SMS แบบผสมระหว่าง CB Filtering และการรับรองจากมนุษย์ (CHAPCHA) จากรูปแบบภาพเงื่อนไขที่ Bot ไม่สามารถตอบได้ โดยวิเคราะห์ผลการจำลองการทำงานจากจำนวนการส่งผ่านข้อมูลแล้วนำมาหาค่าความน่าจะเป็นของ SMS ปกติอย่างเพื่อแสดงให้เห็นว่าการกรอง SMS ที่นำเสนอนี้ช่วยให้สามารถกรอง SMS ได้ถูกต้องมากกว่าการนำ SMS มาการกรองแบบ CB Filtering เพียงอย่างเดียว ซึ่งเป็นการลดภาระการทำงานในส่วนของข้อมูลขยะที่ไม่จำเป็นให้ทำงานได้อย่างเต็มประสิทธิภาพมากขึ้น ทั้งนี้ค่าต่างๆ ที่จะนำมาใช้งานจะต้องคำนึงถึงค่าอ้างอิง h , h_1 , h_2 , e_1 , e_2 ที่เหมาะสมเนื่องจากจะส่งผลกระทบต่อจำนวนของ SMS ที่ถูกนำมาเข้าสู่กระบวนการรับรองจากมนุษย์เพิ่มเติมมากเกินความจำเป็นซึ่งอุปกรณ์ก็จะต้องทำงานได้อย่างรวดเร็วและถูกต้อง จึงจะสามารถช่วยลดภาระการทำงานของ SMSC ได้ตามวัตถุประสงค์

6. เอกสารอ้างอิง

[1] G. Hidalgo, J. Maria, E. Sanz, Gacia, "Content base SMS spam filtering," ACM, Symposium on Document engineering, pp.114-122, 2006.

[2] Androutsopoulos I., "An Evaluation of Naive Bayesian Anti-Spam Filtering, Proc of the workshop on Machine Learning in the New Information Age," Bacelona, Spain, pp.9-17, 2000.

[3] S. Shirali-Shahreza, A. Movaghar, "An anti-spam using CAPTCHA," IEEE Trans. Computer Society, pp. 318-321, 2008.

[4] N. Boonnitiprasert, C. Khemmapatapan, "SMS Filtering for Thai & EnglishLanguage on Mobile Phone Network," NCCIT. The 5th National Conference on Computing and Information Technology, Bangkok, pp.436-442, 2009.

ประวัติผู้เขียน

ชื่อ-นามสกุล นายอำนาจ ละมัยกลาง

วัน เดือน ปีเกิด 13 พฤษภาคม 2524 ที่จังหวัดนครราชสีมา

ที่อยู่ 9 หมู่ที่ 7 ตำบลธารปราสาท อำเภอโนนสูง จังหวัดนครราชสีมา 30420

ประวัติการศึกษา 2548 วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมอิเล็กทรอนิกส์และโทรคมนาคม (เกียรตินิยมอันดับ1) มหาวิทยาลัยเทคโนโลยีราชมงคลอิสาน วิทยาเขตภาคตะวันออกเฉียงเหนือ นครราชสีมา

ประสบการณ์การทำงาน

พ.ศ.2549-2550 วิศวกร แผนก HGSA Test Engineer

บริษัท ซีเกท เทคโนโลยี (ประเทศไทย) จำกัด

พ.ศ. 2550-ปัจจุบัน วิศวกร ระดับ 6 สังกัดส่วนธุรกิจบริการเสริมอินเทอร์เน็ต

ฝ่ายพัฒนาผลิตภัณฑ์อินเทอร์เน็ตและโทรศัพท์

บริษัท กสท โทรคมนาคม จำกัด (มหาชน)