Analyzing User Reviews in Thai Language toward Features in Application

***Abstract*—The abstract goes here. limited to the maximum of 6 pages of A4 form in PDF format. an abstract of about 100 words. The authors' names and affiliations, postal addresses, telephones, fax numbers and e-mail addresses must be omitted.**

I. Introduction

ในปัจจุบันเรามีการใช้งานอุปกรณ์อิเล็กทรอนิกส์กันมากขึ้น โดยเฉพาะอุปกรณ์พวก smart phone และ tablet จนเราอาจสามารถเรียกได้มันว่าเป็นอวัยวะส่วนหนึ่งของเรา ดังนั้นจึงเป็นเหตุให้มีการพัฒนาโปรแกรมสำหรับใช้งานบนอุปกรณ์พวกนี้ขึ้นเป็นจำนวนมาก

และในการพัฒนาโปรแกรมหนึ่งให้ติดตลาดการใช้งาน ไม่ใช่เพียงแควาเราพัฒนาโปรแกรมตามความพึงพอใจของเราเพียงอย่างเดียวเท่านั้น แต่เราต้องดูกระแสตอบรับของผู้ใช้งานโปรแกรมนั้น ๆ ด้วย ว่าพวกเขามีความรู้สึกอย่างไรกับโปรแกรมที่เราพัฒนาขึ้นมา มิเช่นนั้นโปรแกรมที่เราพัฒนามานั้นอาจจะไม่มีใครใช้มันเลยก็เป็นไปได้ โดยมีการสำรวจขอมูลคำถามที่นักพัฒนาต้องการทราบ ซึ่งพบว่านักพัฒนาต้องการทราบ "What parts of a software product are most used and/or loved by customer?" สูงเป็นอับดับที่สอง [1]

ดังนั้นใน app store ต่าง ๆ จึงได้มีช่องทางสำหรับให้ผู้ใช้งานมาแสดงถึงปัญหา ความคิดเห็นและให้คะแนนโดยรวมเกี่ยวกับโปรแกรมที่ใช้นั้น ๆ เพื่อให้ผู้ใช้งานคนอื่นที่สนใจ รวมถึงเจ้าของโปรแกรมนั้น ๆ ได้รับทราบถึงปัญหาหรือคำแนะนำจากผู้ใช้งานคนอื่น ๆ แต่ทั้งนี้ทั้งนั้นความคิดเห็นของผู้ใช้งานในโปรแกรมนั้น ๆ อาจจะมีจำนวนมากจนทำให้เราไม่สามารถที่จะวิเคราะห์ความคิดเห็นได้ด้วยตนเองทั้งหมด หรืออาจจะทำได้แต่ใช้เวลาที่นานจนทำให้โปรแกรมนั้นอัพเดทหรือปรับปรุงแก้ไขปัญหาที่เกิดได้ไม่ทัน อีกทั้งความคิดเห็นของผู้ใช้งานบางคนอาจจะไม่มีประโยชน์ต่อการวิเคราะห์ข้อมูล (เช่น การบอกว่า "ดี" เพียงอย่างเดียว เราไม่สามารถทราบได้ว่าคำว่าดีที่ว่าหมายถึงอะไร) และ rate ที่ผู้ใช้งานให้มานั้นสามารถบอกได้เพียงแค่ภาพรวมของโปรแกรมเท่านั้น ไม่สามารถแจกแจงได้ว่าส่วนไหนที่ผู้ใช้ชอบหรือไม่ชอบ

ดวยปัญหาที่กล่าวข้างต้นทำให้เกิดงานวิจัยที่ใช้ในการวิเคราะห์ความคิดเห็นของผู้ใช้งานโปรแกรมเหล่านี้เป็นจำนวนมาก โดยมีเป้าหมายในการช่วยลดภาระการวิเคราะห์ความคิดเห็นของผู้ใช้งาน ไม่ว่าจะเป็นการหาข้อมูลที่มีสาระประโยชน์จากความคิดเห็นทั้งหมด หรือการสกัดเอาคำสำคัญของความคิดเห็นเหล่านั้นขึ้นมาเพื่อจับกลุ่มของแสดงหัวข้อที่ผู้ใช้งานกล่าวถึง

แต่ด้วยในขณะนี้งานวิจัยที่มียังไม่สามารถนำมาใช้กับการวิเคราะห์ความคิดเห็นภาษาไทยได้ ซึ่งงานวิจัยนี้จึงมีจุดประสงค์ในการพัฒนากระบวนการและแนวคิดในการวิเคราะห์ความคิดเห็นที่เป็นภาษาไทยเพื่อเป็นแนวทางในการวิเคราะห์ขอมูลได้อย่างมีประสิทธิภาพมากขึ้น โดยในหัวข้อ II จะกล่าวถึงทฤษฎีที่มีความสำคัญเกี่ยวกับงานวิจัยนี้, หัวข้อ III กล่าวถึงงานวิจัยที่เกี่ยวข้อง, หัวข้อ IV กระบวนการในการวิจัย, หัวข้อ V ผลลัพธ์ของการวิจัย และหัวข้อ VI จะเป็นการสรุปงานวิจัยนี้

II. Background

เนื่องจากในการประมวลผลข้อความต่าง ๆ เราจำเป็นต้องทราบถึงคำแต่ละคำในข้อความนั้น เพื่อที่เราจะสามารถนำคำเหล่านั้นมาพิจารณาได้ จึงทำให้เกิดกระบวนการการตัดคำขึ้นมา แต่ทั้งนี้ทั้งนั้นการตัดคำเพียงอย่างเดียว อาจจะไม่เพียงพอต่อการนำคำต่าง ๆ ที่ได้ไปประมวลผล เนื่องจากคำบางคำมีความหมายและหน้าที่ของคำที่ต่างกัน จึงเป็นเหตุให้นอกจากการตัดคำเพียงอย่างเดียวนั้นไม่สามารถแยกแยะความหมายและหน้าที่ของคำได้ ทำให้ตองมีกระบวนการหาหน้าที่ของคำเหล่านั้นขึ้นมาด้วย

ซึ่งในงานวิจัยนี้มีความต้องการที่จะวิเคราะห์หาหัวข้อที่แสดงถึงทัศนคติของความคิดเห็นของผู้ใช้งานโปรแกรมบนอุปกรณ์ smart device ทำให้งานวิจัยนี้เกี่ยวข้องกับการประมวลผลขอความ ซึ่งต้องมีการตัดคำและหาหน้าที่ของคำเหล่านั้น และนอกจากนั้นงานวิจัยนี้ยังต้องหาหัวขอของความคิดเห็น และวิเคราะห์ทัศนคติของความคิดเห็นเหล่านั้นด้วย

ดังนั้นงานวิจัยนี้จึงมีทฤษฎีที่เกี่ยวข้องอยู่ 4 ส่วนคือ การตัดคำ การหาหน้าที่ของคำ การหาหัวข้อของข้อความต่าง ๆ และการวิเคราะห์ทัศนคติของประโยค

A. Word Segmentation

การตัดคำในประโยคภาษาอังกฤษนั้นเราสามารถทำได้โดยไม่ยากเนื่องจากเรามีจะใช้ช่องว่างในการแบ่งคำหรือใช้ '.' ในการจบประโยค หรือใช้ '?' ในการจบประโยคสำหรับคำถาม แต่สำหรับประโยคภาษาไทยนั้นจะมีลักษณะที่คล้ายกับภาษาจีนและญี่ปุ่นตรงที่เราไม่มีการเว้นวรรคคำแต่ละคำในประโยค ทำให้การหาคำภาษาไทยมีความลำบากมากกวาภาษาอังกฤษ [2]

โดยแนวคิดสำหรับการตัดคำในปัจจุบันนั้นมีหลายแนวคิด เช่น

1) Longest Matching [3]: วิธีการนี้เป็นการนำสายอักขระทั้งสายมาเปรียบเทียบกับคำที่อยู่ในพจนานุกรม (lexicon) โดยถ้ามีคำตรงกับใน lexicon วิธีนี้ก็จะนำคำที่ได้ออกมาสายอักขระ แต่ถาเทียบแล้วไม่พบใน lexicon วิธีนี้จะใช้วิธีตัดตัวอักษรตัวสุดท้าย

ของสายอักขระออกไป 1 ตัวอักษร แล้วทำการเทียบกับ lexicon ใหม่ โดยจะทำอย่างนี้จนจบทั้งสายอักขระ

ตัวอย่างคำที่ได้จากวิธีนี้ เช่น "ฉันนั่งตากลมอยู่ริมตลิ่ง" จะได้ "ฉัน นั่ง ตาก ลม อยู่ ริม ตลิ่ง"

2) Maximal Matching [4]: วิธีการนี้เป็นการทดลองตัดคำที่มีโอกาสเกิดขึ้นได้ทุกรูปแบบก่อน จากนั้นจะทำการเลือกรูปแบบที่มีจำนวนคำที่น้อยที่สุดออกมา แต่ถ้าเกิดรูปแบบที่มีคำน้อยที่สุดมีหลายรูปแบบ วิธีการนี้ก็จะนำวิธีการ Longest Matching เข้ามาช่วยในการตัดสินใจ

ตัวอย่างคำที่ได้จากวิธีการนี้ เช่น "ไปหามเหสี" ซึ่งมีโอกาสได้คำว่า "ไป หาม เห สี" และ "ไป หา มเหสี" โดยวิธีการนี้จะเลือกคำว่า "ไป หา มเหสี" ออกมา

3) Probabilistic Model [5]: วิธีการนี้จะมีการใช้ข้อมูลเชิงสถิติการเกิดคำและหน้าที่ของคำ เข้ามาช่วยในการหาโอกาสของคำที่จะเกิดขึ้นในประโยคที่มากที่สุด ซึ่งวิธีการนี้จำเป็นต้องมี corpus ที่มีข้อมูลของคำและหน้าที่ของคำที่ถูกต้อง เพื่อนำมาใช้ในการคำนวณสถิติของคำที่จะเกิดขึ้น

4) Feature-based Approach [2]: วิธีการนี้จะมีการพิจารณาจากบริบทและการเกิดร่วมกันของคำ มาตัดสินใจในการตัดคำนั้น ๆ

เช่น คำว่า "มากว่า" ถ้าคำที่มาต่อท้ายคำนี้เป็นตัวเลข วิธีนี้จะตัดสินใจว่าควรจะได้คำว่า "มา กว่า"

ซึ่งในงานวิจัยนี้ผู้วิจัยเลือกใช้วิธีการตัดคำแบบ Longest Matching โดยใช้โปรแกรม LexTo [6] ซึ่งเป็นโปรแกรมที่ถูกพัฒนาโดย National Electronics and Computer Technology Center (NECTEC) ซึ่งรองรับในการตัดคำด้วยวิธีการ Longest Matching

B. Part of Speech

หน้าที่ของคำเป็นสิ่งที่ใช้กำหนดชนิดของคำในข้อความนั้น ๆ โดยชนิดของคำสามารถแบ่งได้ 8 ประเภทใหญ่ คือ 1. noun, 2. pronoun, 3. verb, 4. adverb, 5. adjective, 6. preposition, 7. conjunction, 8. interjection

โดยการหาหน้าที่ของคำนั้นเป็นส่วนหนึ่งในการประมวลผลภาษาธรรมชาติ ซึ่งจำเป็นต้องมี corpus สำหรับการเรียนรู้หน้าที่ของคำ เพื่อที่เราจะสามารถตอบได้ว่าคำที่มีอยู่ในข้อความนั้นเป็นคำชนิดอะไร และ corpus ของภาษาไทยที่สามารถพบเห็นได้บ่อย คือ NAiST [7] corpus ซึ่งเป็นคลังข้อมูลที่พัฒนาโดยคณะวิศกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ และ ORCHID [8] corpus ที่พัฒนาโดย NECTEC

ในการหาหน้าที่ของคำนั้นมีเครื่องมือที่ช่วยในการค้นหาอยู่มากมายตัวอย่างเช่น Nature Language ToolKit (NLTK) [9] เป็นเครื่องมือที่ใช้ในการประมวลภาษาธรรมชาติภาษาอังกฤษ ซึ่งรองรับการเพิ่ม corpus อื่นๆ นอกจากที่มีอยู่ในระบบอยู่แล้ว และ RDRPOStagger [10] เป็นเครื่องมือสำหรับการหาหน้าที่ของคำโดยเฉพาะ โดยมี corpus 7 ภาษา รวมถึงภาษาไทย (ซึ่งใช้ orchid เป็น corpus)

C. Topic Modeling

Topic Modeling เป็นวิธีการจัดกลุ่มหัวข้อของประโยคที่เรากำลังพิจารณาอยู่ โดยการคาดเดาความน่าจะเป็นของคำที่จะเกิดในกลุ่มของหัวข้อนั้น ๆ ซึ่งจะมีวิธีการหาหัวข้อได้ 2 แบบหลัก ๆ คือ

1) Aspect and Sentiment Unification Model (ASUM) [11]: เป็นวิธีการหาหัวข้อของประโยค โดยมีหลักการว่า 1 ประโยค จะมีผู้กล่าวจะกล่าวถึงหัวข้อเพียงหัวข้อเดียว

2) Latent Dirichlet Allocation (LDA) [12]: มีแนวคิดที่ว่าในประโยค อาจจะมีหัวข้อที่ถูกกล่าวถึงมากกว่า 1 หัวข้อ โดยในแต่ละหัวข้อจะเกิดจากการรวมกันของคำหลาย ๆ คำ ซึ่งแต่ละคำในหัวข้อก็จะมีความน่าจะเป็นที่แตกต่างกัน

D. Sentiment Analysis

Sentiment Analysis หรือ Opinion mining เป็นการศึกษาเกี่ยวกับความรู้สึก อารมณ์ ทัศนคติ จากการสังเกตเนื้อหาของการสนทนาเหล่านั้น [13] โดยในการทำ sentiment analysis นั้นมีวิธีการวิเคราะห์พื้นฐาน 2 แบบคือ

1) lexicon-based: คือการนำคำที่ต้องการหาค่าทัศนคติไปเปรียบเทียบกับ lexicon ที่มีค่าทัศนคติกำกับอยู่ เพื่อใช้เป็นตัวแทนของค่าทัศนคติของคำนั้น ๆ

2) machine learning-based: คือ การนำข้อมูลที่มีค่าทัศนคติมาฝึกฝนการวิเคราะห์ของคอมพิวเตอร์ เพื่อนำไปเป็นฐานสำหรับการวิเคราะห์ข้อมูลอื่น ๆ ต่อไป

III. Related Work

เราได้แบ่งงานวิจัยที่เกี่ยวข้องออกเป็น 2 กลุ่ม คือ 1. งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ทัศนคติของผู้ใช้งานโปรแกรมบนโทรศัพท์เคลื่อนที่ และ 2. งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ขอมูลภาษาไทย

1) งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ทัศนคติของผู้ใช้งานโปรแกรมบนโทรศัพท์: จากการศึกษาพบว่างานวิจัยทางด้านการวิเคราะห์ทัศนคติของผู้ใช้งานโปรแกรมบนโทรศัพท์เคลื่อนที่นั้นมีจำนวนมาก แต่ส่วนใหญ่จะมีแต่งานวิจัยที่วิเคราะห์ข้อมูลภาษาอังกฤษเพียงอย่างเดียว [14]–[16] เป็นเหตุให้งานวิจัยเหล่านี้อาจไม่สามารถนำมาใช้ในการวิเคราะห์ข้อมูลภาษาอื่น ๆ ได้

Ning Chen และคณะ [14] ได้นำเสนอ AR-Miner เป็นงานวิจัยที่จะสกัดและจัดอันดับประโยคที่มีสารประโยชน์ (informative reviews) จากข้อความของผู้ใช้งาน โดยพวกเขาใช้วิธีการแบ่งกลุ่มคำที่มีประโยชน์และไม่มีประโยชน์ (classifier) ด้วยวิธี Expectation Maximization for Naive Bays (EMNB) [17] ซึ่งเป็นการจัดกลุ่มโดยการนำข้อมูลที่ทราบคำตอบมาเป็นแบบในการหาข้อมูลที่ไม่ทราบคำตอบ จากนั้นจึงจะนำข้อมูลที่เป็นประโยชน์มาแบ่งกลุ่มตามหัวข้อที่ผู้ใช้ได้กล่าวถึง โดยการเปรียบเทียบระหว่างวิธีการ LDA และ ASUM

Emitza Guzmann และ Wiem Maalej [16] ได้เสนอวิธีการสกัดหา feature และ sentiment ของโปรแกมจาก review ของผู้ใชงาน โดยพวกเขาจะใชเฉพาะ noun, verb, and adjective ที่

อยู่ในประโยคมาแทน feature และใช้ SentiStrengh [18] ในการหา sentiment ของแต่ละคำ แล้วนำคะแนนที่มากที่สุดของคำในประโยคที่มาเป็นคะแนนของประโยค และใช้ LDA ในการจับกลุ่ม feature ต่าง ๆ ซึ่งในการวัดความถูกต้องนั้นพวกเขาได้ให้นักวิจัยอีกคนที่ไม่มีส่วนเกี่ยวกับการทำงานวิจัยนี้มาเปรียบเทียบ

Phong Minh Vu และคณะ [15] ได้เสนอวิธีการหาประโยคที่ผู้ใช้งานได้กล่าวต่อว่าหรือกล่าวถึงปัญหาของโปรแกรม จากการหา keyword ของคำในประโยค โดย keyword ที่พวกเขาใช้จะเป็นคำ noun และ verb และจับกลุ่ม keyword เหล่านี้ด้วยการหาค่าความใกล้เคียงของคำด้วยวิธี cosine similarity [19] ซึ่งพวกเขาเปรียบเทียบความถูกต้องของงานวิจัยโดยการให้ผู้เชี่ยวชาญ 8 คนตรวจสอบความเหมาะสมของกลุ่มคำสำคัญที่หาได้

โดยงานวิจัยเหล่านี้มีความคล้ายกับงานวิจัยนี้ตรงที่ต่างก็ต้องการหาวิธีที่จะแสดงถึงหัวขอที่ผู้ใช้งานต้องการสื่อสาร โดยเฉพาะอย่างยิ่งงานวิจัยของ Emitza ที่มีการวิเคราะห์หา sentiment แต่ทั้งนี้ทั้งนั้นงานวิจัยเหล่านี้อยู่เป็นการวิจัยที่ทำอยู่บนฐานของภาษาอังกฤษ ซึ่งสำหรับการวิเคราะห์ภาษาไทยนั้นจะมีความแตกต่างกับการวิเคราะห์ภาษาอังกฤษอยู่บางส่วน

2) งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลภาษาไทย: จากการศึกษาพบว่ามีงานวิจัยที่เกี่ยวกับการวิเคราะห์ข้อมูลภาษาไทยอยู่จำนวนหนึ่ง ซึ่งมีทั้ง การสรุปบทความของเอกสารจากย่อหน้า [20], การจัดกลุ่มคำตามอารมณ์พื้นฐานของมนุษย์ [21], การวิเคราะห์ทัศนคติบนสื่อสังคมออนไลน์ [22], และการวิเคราะห์ทัศนคติของผู้ใช้งานโรงแรม [23] แต่ผู้วิจัยยังไม่พบงานวิจัยที่เกี่ยวกับการวิเคราะห์ทัศนคติของโปรแกรมบนโทรศัพท์

Chuleerat Jaruskulchai และ Canasai Kruengkrai [20] ได้เสนอวิธีการสรุปบทความของเอกสารภาษาไทยด้วยการสกัดหาย่อหน้าที่สำคัญ โดยพวกเขาได้แบ่งข้อความออกเป็นย่อหน้าและตัดคำของแต่ละย่อหน้าด้วยวิธีการ Longest matching จากนั้นจึงนำคำที่ตัดได้ไปหาความสัมพันธ์ระหว่างแต่ละย่อหน้า เพื่อเชื่อมโยงยอหน้าที่มีความเกี่ยวของกัน จากคำที่ต้องการค้นหาโดยใช้นักศึกษาจากคณะศิลปศาสตร์ สาขาภาษาไทย สรุปเอกสารเหล่านี้เพื่อใช้วัดความถูกต้องของงานวิจัย

Piyatida และ Sukree Sinthupinyo [21] ได้เสนอวิธีการจัดกลุ่มคำที่แสดงถึงอารมณ์ จากคำนามและคำกริยา โดยใช้ SWATH ตัดคำซึ่งเป็นโปรแกรมที่ใช้ในการตัดคำที่พัฒนาโดย NECTEC และใช้ ORCHID corpus เพื่อหา POS ของคำเหล่านั้น แล้วนำคำที่เป็นคำนามและกริยามาจัดกลุ่มตามอารมณ์พื้นฐานของมนุษย์ 6 อย่างคือ โกรธ, ขยะแขยง, กลัว, ดีใจ เสียใจ, และตกใจ [24] ด้วยวิธีการจัดกลุ่มแบบ Naive Bays

Choochart Haruechaiyasak และคณะ [22] ได้เสนอ S-Sense เป็นเครื่องมือสำหรับการวิเคราะห์ทัศนคติบนสื่อสังคมออนไลน์ เช่น twitter และ pantip (webboard ที่แพร่หลายของคนไทย) เป็นต้น โดยใช้ LEXiTRON ซึ่งเป็นพจนานุกรมไทย-อังกฤษ และอังกฤษ-ไทย ที่พัฒนาโดย NECTEC เพื่อแปลภาษาและจับกลุ่มของคำที่มีความหมายใกล้เคียงกัน และใช้ Utilization on REsource for Knowledge Acquisition (UREKA) ซึ่งเป็นส่วนประกอบส่วนหนึ่งของงานวิจัยของเขา ในการสกัดหัวข้อ/คำสำคัญ ออกมาจากขอความ โดยงานวิจัยนี้จะแบ่งกลุ่มของข้อความออกเป็น การประกาศ การแสดงความต้องการ การแสดงคำถาม และความคิดเห็น โดยพวกเขาได้แยกความคิดเห็นออกเป็น เชิงบวก และ เชิงลบ

Choochart Haruechaiyasak และคณะ [23] ได้เสนอวิธีการวิเคราะห์ทัศนคติของผู้ที่มาใช้โรงแรม โดยใช้ lexicon และ corpus สำหรับค้นหาคำและทัศนคติของคำ โดยงานวิจัยนี้มีหัวข้อสำหรับการวิเคราะห์ทัศนคติที่ชัดเจนคือ 1. การบริการ, และ 2. อาหารเช้า โดยการศึกษาหา pattern ของประโยค เพื่อจับกลุ่มประโยคให้อยู่ในหัวข้อที่ต้องการ และนำกลุ่มหัวข้อเหล่านั้นไปวิเคราะห์ทัศนคติต่อไป

โดยงานวิจัยเหล่านี้ต่างก็ต้องมีการหา POS เพื่อที่จะกรองชนิดของคำที่ต้องการนำมาวิจัย (ซึ่งส่วนมากจะเป็น noun, verb, adjective) โดยงานวิจัย S-Sense และ การหาทัศนคติของผู้ใช้งานโรงแรม ต่างมีความต้องการที่จะหาทัศนคติของผู้ใช้งานเหมือนงานวิจัยฉบับนี้ แต่ก็มีความแตกต่างกันอยู่บ้างตรงที่ การหาทัศนคติของผู้ใช้งานโรงแรม นั้นจะมีหัวข้อที่ต้องการจะหาอยู่แล้วอย่างแน่นอน และของ S-Sense จะสามารถใช้กับสื่อสังคมออนไลน์ได้ แต่ไม่สามารถนำมาใช้กับการวิเคราะห์โปรแกรมที่อยู่ในโทรศัพท์ ได้ ซึ่งสำหรับงานวิจัยฉบับนี้ ต้องการที่จะวิเคราะห์หาหัวขอและทัศนคติของโปรแกรมที่อยู่ในโทรศัพท์ และในแต่ละโปรแกรมก็จะมีหัวข้อที่ถูกกล่าวถึงต่างกัน ทำให้ไม่สามารถที่จะกำหนดหัวข้อที่ต้องการจะวิเคราะห์ได้อย่างชัดเจน

IV. Approach

การวิจัยนี้มีเป้าหมายในการหาหัวข้อและทัศนคติของความคิดเห็นของผู้ใช้งานโปรแกรมที่อยู่ตาม app store โดยใช้วิธีการประมวลผลภาษาธรรมชาติ และการทำเหมืองข้อมูล Figure 1

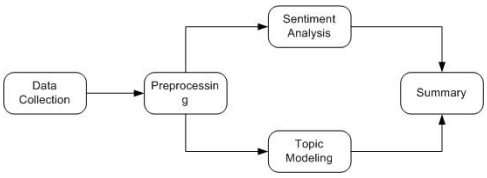

Fig. 1. Overview of approach

จะแสดงขั้นตอนทั้งหมดในการวิจัย โดยจะเริ่มตั้งแต่ 1. Data Collection 2. Prepossessing 3. Sentiment Analysis 4. Topic Extraction 5. Summary

A. Data Collection

TABLE I
no. of review in each app.

Application	no. of review
Man Man	1279
H-Tv	691
K-mobile	1055

เราได้รวบรวมข้อมูลความคิดเห็นของผู้ใช้งานจากโปรแกรมประเภทต่าง ๆ บน Google Play store โดยการรวบรวมจากบนหน้าเว็บไซต์สำหรับดาวน์โหลดโปรแกรมนั้น ๆ โดยรวบรวมจากโปรแกรม "แม่น แม่น" (virtual keyboard), "H-Tv" (TV Online), "K-mobile" (Internet Mobile Banking) โดยเราได้รวบรวมข้อมูลในช่วง กุมภาพันธ์ 2015 - พฤษภาคม 2016 และช่วง มิถุนายน 2016 - สิงหาคม 2016 (นับจากวันที่ผู้ใช้งานแสดงความคิดเห็น) ซึ่งมีปริมาณข้อมูลตาม Table I โดยข้อมูลที่ผู้วิจัยได้รวบรวมมา ได้แก่ author, title, detail, rate, review-date ซึ่งแสดงตัวอย่างตาม Table II

B. Prepossessing

หลังจากที่เราได้ข้อมูลที่ต้องการแล้ว เราจะนำข้อมูลเหล่านั้นมาหา POS ก่อนเพื่อใช้ในการทำงานขั้นต่อไป แต่ก่อนที่เราจะหา POS ได เราจะตองตัดประโยค และตัดคำกอน

1) sentence extraction: เนื่องจากขอมูลความคิดเห็น 1 ความคิดเห็นอาจจะไม่ได้มีเพียงประโยคเดียว ดังนั้นเราจึงจำเป็นต้องแบงประโยคออกมาเสียกอน เนื่องจากประโยคภาษาไทยนั้นเราไมมี pattern ที่ตายตัวในการแบงประโยคเหมือนอยางภาษาอังกฤษ และในปัจจุบันมีคำสมัยใหม่เพิ่มขึ้นมาอีกมากมาย ทำให้ลำบากในการใช้เครื่องมือในการแบ่งประโยค อีกทั้งยังจำเป็นต้องใช corpus ที่มีขอมูลของรูปประโยคที่คอนขางมากเพื่อใชในการจำแนกประโยคต่าง ๆ ดังนั้นเราจึงใช้วิธี manual ในการแบ่งประโยค

โดยเราใช้ pattern ในการแบ่งประโยคคือ

1. ถ้าเจอคำวา "ครับ"/"คะ" เราจะถือวาเป็นการจบประโยค
2. ถ้าเจอคำวา "แต" เราจะถือวาเป็นการขึ้นประโยคใหม่

2) word segmentation: เมื่อเราแบงประโยคเรียบร้อยแล้ว เราจะนำประโยคที่ได้แต่ละประโยคมาตัดแยกคำเพื่อนำไปหา POS ตอไป โดยในการตัดคำนั้นเราได้ใชเครื่องมือที่ชื่อ LexTo ซึ่งพัฒนาโดย NECTEC เป็นซึ่งใช้วิธีการตัดคำแบบ longest matching ในการตัดคำ

3) pos tagger: เราหา pos ของคำโดยใช้ RDRPOStagger ซึ่งมี ORCHID เป็น corpus สำหรับการคำนวน

C. Sentiment Analysis

ส่วนนี้เป็นการนำประโยคที่มีการกำหนด pos ของคำแล้วมา คำนวนหาทัศนคติของประโยค โดยสำหรับการหา sentiment ของคำในภาษาไทยนั้นยังไมมี corpus ที่เผยแพร ดังนั้นเราจึงเลือกใช SentiWordNet [25] ซึ่งเป็น corpus สำหรับหา sentiment ของคำในภาษาอังกฤษ

ดังนั้นขั้นตอนแรกของการหา sentiment ของงานวิจัยนี้จึงเป็นการแปลคำศัพทจากไทย-อังกฤษ โดยเราเลือกใช LEXiTRON [26] เป็นพจนานุกรมในการแปลคำศัพท์ โดยการหาคำที่มี POS ตรงกัน ทำให้เราได้ synonym ภาษาอังกฤษ

ขั้นต่อมาเราจะนำ synonym ที่ได้มาหา sentiment ใน SentiWordNet โดยค่าของ sentiment ที่ได้จะอยู่ในช่วง [-1,1] ดังตัวอย่างใน Table III แต่เนื่องจากค่าที่ได้จาก SentiWordNet เราพบวา คำบางคำ ที่ให้ความรูสึกในเชิงลบของรูปประโยค มีคาที่ได้เป็นบวก ดังนั้นเราจึงจำเป็นต้องสร้างลิสต์คำที่คาดว่าให้ความรู้สึกเชิงลบ แล้วนำมาเทียบกับ sentiment ที่ได้ ถ้า sentiment ที่ได้เป็นบวก เราจะกลับค่า sentiment นั้นให้เป็นลบแทน และในกรณีที่มีคำวา "ไม" นำหนาคำ ๆ นั้น เราก็จะกลับคา sentiment ของคำนั้นแทน

จากนั้นเราจะหาค่า sentiment ของประโยคโดยการนำค่า sentiment ทั้งหมดของประโยคนั้น ๆ มาเฉลี่ยเป็นคะแนนของประโยค

D. Topic Extraction

ส่วนนี้เป็นส่วนที่อธิบายถึงวิธีการหาหัวข้อของประโยค โดยเราใช้วิธี LDA ในการคนหาหัวขอ ซึ่งจะทำหลังจากการหา pos ของคำ โดยเรากำหนดให้ number of topic ที่ต้องการเป็น 20 เนื่องจากเราไมทราบหัวข้อที่แน่นอน จากนั้นเราจะเลือกหัวข้อที่เหมาะสมออกมาจากหัวขอทั้งหมดที่ได

E. Summary

หลังจากที่เราได้กลุ่มคำของ topic ต่าง ๆ และ sentiment ของประโยคแล้ว เราจะรวมรวบประโยคที่มีคำตรงกับในกลุ่มคำของ topic เพื่อนำมาแสดงถึงคะแนน sentiment ของ topic นั้น ๆ และนำคะแนนที่ได้มาหาค่าเฉลี่ย เพื่อแสดงถึง sentiment รวมของหัวขอนั้น ๆ วาคะแนนเป็นบวก หรือเป็นลบ

นอกจากนี้เรายังสามารถแจกแจงได้วาแต่ละหัวข้อมีประโยคที่มีทัศนคติเป็นบวก หรือเป็นลบ อยูอยางละกี่ประโยคได้อีกดวย

V. Result

เราได้ตรวจสอบความถูกต้องของงานวิจัยโดยให้ผู้เชี่ยวชาญประเมินทัศนคติของประโยค ซึ่งจะได้ค่าความถูกต้องตาม Table IV

Limitation

เนื่องจากเรายังหาวิธีที่จะใช้ในการแบ่งประโยคที่ชัดเจนยังไม่ได้จึงทำให้การแบงประโยคนั้นอาจยังไมถูกตอง รวมถึงคำบางคำอาจจะเป็นคำสมัยใหม่ หรือภาษาวัยรุ่น ทำให้คำเหล่านี้ไม่มีอยู่ในcorpus ที่เราใช้งาน จึงเป็นเหตุให้เราอาจไม่สามารถหาทัศนคติของคำเหล่านี้ได้

และในการหา sentiment โดยการแปลภาษาไทย-อังกฤษ คำบางที่แปลได้ อาจแปลได้ไม่ตรงตามความต้องการของประโยค ทั้งนี้เนื่องมาจาก การหา POS และ การพองรูปในภาษาไทย

อีกทั้งในเรื่องของการหาหัวข้อที่ไม่มีความแน่นอนของโปรแกรมตาง ๆ จึงทำให้เรากำหนดจำนวนหัวขอที่ต้องการไม่ได้

VI. Conclusion

งานวิจัยนี้ ได้เสนอแนวคิดในการหาหัวข้อและทัศนคติของโปรแกรมในโทรศัพท์เคลื่อนที่ ดวยวิธีการ NLP และ Topic modeling ซึ่งผลลัพธ์ที่ได้อาจจะยังไม่น่าพอใจมากนัก แตยังสามารถนำมาเป็นแนวทางในการวิจัยต่อ ๆ ไปได้

TABLE II
example of review

author	title	review	rate	date
โชคชัย มหาวงนันท์	โชคชัย มหาวงศ์นันท์	ใช้ได้ดีครับ	5	10/04/2015
bie slow life		พักหลังนี่อัพบ่อยนะครับ	4	09/19/2015
ornanohg Hongrrimon		ชอบคะใช้ง่าย มีตัวการ์ตูนให้ด้วย	5	09/20/2015
Terdsak chompusri		เรียบง่ายแต่ใช้ได้ดีจริง ๆครับชอบมาก	5	09/22/2015
Worapote Panomauppatum	วรพจน์ พนมอุปถัมภ์	ใช้ได้เยี่ยมมาก	5	09/25/2015
Nate Makboon	เนตร มากบุญ	ดีมากครับ สะดวกดีแม่นสุดยอด	5	09/24/2015

TABLE III
Top 10 sentiment of each word in Man Man app

negative		positive	
word	sentiment	word	sentiment
ลบ	-0.33621	น่ารัก	0.21843
เสียดาย	-0.33621	รัก	0.21843
เกลียด	-0.33621	เพลิน	0.21843
ดุ	-0.33621	ดี	0.21843
สายตายาว	-0.33621	สวย	0.21843
ขยายตัว	-0.33621	สุดยอด	0.21843
ห่วย	-0.33621	มันส์	0.21843
ปวด	-0.33621	ไว	0.21843
เสียใจ	-0.33621	ชอบ	0.21843
ไม่ได้	-0.33621	สนุก	0.21843

TABLE IV
F-measure and Accuracy for sentiment analysis

Application	F-measure	Accuracy
Man Man	0.352147	0.608059
H-Tv	0.336605	0.483669

References

[1] A. Begel and T. Zimmermann, "Analyze this! 145 questions for data scientists in software engineering," in Proceedings of the 36th International Conference on Software Engineering. ACM, 2014, pp. 12–23.

[2] P. Charoenpornsawat, "Feature-based thai word segmentation," Master's thesis, Computer Engineering, Master. Chulalongkorn University, Bangkok, 1999.

[3] Y. Poovarawan and W. Imarrom, "Thai syllable separator by dictionary," in Proceedings of the 9th Annual Meeting on Electrical Engineering of the Thai Universities, Khonkaen, Thailand, December 1986, p. 14.

[4] V. Sornlertlamvanich, "Word segmentation for thai in machine translation system," Machine Translation, National Electronics and Computer Technology Center, Bangkok, pp. 50–56, 1993.

[5] K. Asanee, T. Chalathip, and S. Sapon, "A statistical approach to thai word filtering," 1995.

[6] National Electronics and Computer Technology Center. LexTo : Text lexeme tokenizer. [Online]. Available: http://www.sansarn.com/lexto

[7] P. Varasai, C. Pechsiri, T. Sukvaree, V. Satayamas, and A. Kawtrakul, "Building an annotated corpus for text summarization and question answering." in LREC, 2008.

[8] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, "Orchid: Thai part-of-speech tagged corpus," National Electronics and Computer Technology Center Technical Report, pp. 5–19, 1997.

[9] NLTK Project. (2015) Natural language toolkit. [Online]. Available: http://www.nltk.org

[10] D. Q. Nguyen, D. Q. Nguyen, D. D. Pham, and S. B. Pham, "Rdrpostagger: A ripple down rules-based part-of-speech tagger," in Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 17–20. [Online]. Available: http://www.aclweb.org/anthology/E14-2005

[11] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 815–824.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[13] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data. Springer, 2012, pp. 415–463.

[14] N. Chen, J. Lin, S. C. Hoi, X. Xiao, and B. Zhang, "Ar-miner: mining informative reviews for developers from mobile app marketplace," in Proceedings of the 36th International Conference on Software Engineering. ACM, 2014, pp. 767–778.

[15] P. M. Vu, T. T. Nguyen, H. V. Pham, and T. T. Nguyen, "Mining user opinions in mobile app reviews: A keyword-based approach (t)," in Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on. IEEE, 2015, pp. 749–759.

[16] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in Requirements Engineering Conference (RE), 2014 IEEE 22nd International. IEEE, 2014, pp. 153–162.

[17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," Machine learning, vol. 39, no. 2-3, pp. 103–134, 2000.

[18] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544–2558, 2010.

[19] C. D. Manning, P. Raghaven, and H. Schutze. (2009) An introduction to information retrieval. [Online]. Available: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[20] C. Jaruskulchai and C. Kruengkrai, "A practical text summarizer by paragraph extraction for thai," in Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11. Association for Computational Linguistics, 2003, pp. 9–16.

[21] P. Inrak and S. Sinthupinyo, "Applying latent semantic analysis to classify emotions in thai text," in Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, vol. 6. IEEE, 2010, pp. V6–450.

[22] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, "S-sense: a sentiment analysis framework for social media sensing," in Sixth International Joint Conference on Natural Language Processing, 2013, p. 6.

[23] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and C. Sangkeettrakarn, "Constructing thai opinion mining resource: A case study on hotel reviews," in 8th Workshop on Asian Language Resources, 2010, pp. 64–71.

[24] P. Ekman, "An argument for basic emotions," Cognition & emotion, vol. 6, no. 3-4, pp. 169–200, 1992.

[25] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An

enhanced lexical resource for sentiment analysis and opinion mining." in LREC, vol. 10, 2010, pp. 2200–2204.
[26] National Electronics and Computer Technology Center. LEXiTRON. [Online]. Available: http://lexitron.nectec.or.th/2009_1