การปิดบังข้อมูลที่ระบุตัวบุคคล

ณัฏฐณิชา ชัยศิริพานิช[1] และ ประวิตรานันท์ บุตรโพธิ์[2]

[1]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

[2]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Emails: 60070135@it.kmitl.ac.th, 60070148@it.kmitl.ac.th

บทคัดย่อ

ในปัจจุบันเทคโนโลยีส่งผลให้การดำเนินชีวิตสะดวกขึ้น ซึ่งก็ส่งผลต่อพฤติกรรมการทำธุรกรรมกับทางธนาคารเช่นกัน กล่าวคือ ลูกค้ามักดำเนินการทำธุรกรรมออนไลน์ หรือดำเนินการทำธุรกรรมผ่านทางศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ (Call Center) ซึ่งรายละเอียดต่าง ๆ ที่ลูกค้าดำเนินการทำธุรกรรมผ่านศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์นั้นมีรายละเอียดข้อมูลส่วนบุคคลค่อนข้างมาก และทางธนาคารได้มีการบันทึกเสียงเพื่อใช้เป็นหลักฐานการระบุตัวตนลูกค้า และนำข้อมูลไปวิเคราะห์เพื่อพัฒนาประสิทธิภาพขององค์กร แต่ในกระบวนการวิเคราะห์นั้น หากยังมีข้อมูลส่วนบุคคลของลูกค้าอยู่ อาจส่งผลให้ผู้ที่ประสงค์ร้ายสามารถลักลอบนำข้อมูลส่วนบุคคลของลูกค้าไปเผยแพร่โดยไม่ได้รับอนุญาตได้ ดังนั้น การรักษาความลับและข้อมูลส่วนตัวของลูกค้าเป็นเรื่องที่ทางธุรกิจต้องพึงตระหนักเป็นอย่างมาก

ทางผู้จัดทำจึงได้สร้างโครงงานฉบับนี้ขึ้นโดยมีวัตถุประสงค์เพื่อปิดบังการสนทนาที่ประกอบด้วยข้อมูลส่วนบุคคลทั้งของลูกค้าและพนักงานผู้ให้บริการ โดยมีการสร้างแบบจำลองที่สามารถแปลงเสียงพูดให้อยู่ในรูปแบบของข้อความ และทำการตรวจจับรูปแบบของข้อมูลที่เป็นส่วนบุคคล จากนั้นทำการจับคู่เวลาที่มีข้อมูลส่วนบุคคล และปกปิดเสียงเหล่านั้นออกไป เพื่อที่องค์กรสามารถนำผลลัพธ์ที่ได้ไปวิเคราะห์และพัฒนาประสิทธิภาพทางธุรกิจ

คำสำคัญ – ข้อมูลส่วนบุคคล; ศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ (*Call Center*); การประมวลผลภาษาธรรมชาติ *(Natural Language Processing);* นิพจน์ระบุนาม *(Named Entities);* การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ (*Speech-to-Text*)

1. บทนำ

ปัจจุบันการละเมิดข้อมูลส่วนบุคคลนั้นเกิดขึ้นได้หลายรูปแบบ ซึ่งการละเมิดข้อมูลส่วนบุคคลจากการบันทึกบทสนทนาการทำธุรกรรมกับทางธนาคารก็ถือเป็นหนึ่งในปัญหาการละเมิดสิทธิส่วนบุคคลเช่นกัน ทางผู้จัดทำได้เล็งเห็นถึงความสำคัญของการรักษาข้อมูลส่วนบุคคลของลูกค้าในการทำธุรกรรมกับทางธนาคารผ่านศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ (Call Center) โดยจะมีการทำการตรวจจับการสนทนาบางส่วนกับทางศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ โดยเฉพาะส่วนที่เป็นข้อมูลส่วนบุคคลของลูกค้า เช่น ชื่อ - นามสกุล วันเกิด เบอร์โทรศัพท์ เลขที่บัญชี และเลขหน้าบัตรเครดิต หรือเดบิต ก่อนจะนำข้อมูลการสนทนาเหล่านั้นส่งต่อไปสู่กระบวนการวิเคราะห์เพื่อใช้ในกระบวนการทางธุรกิจ โดยทางผู้จัดทำจะดำเนินการแปลงการสนทนานั้นให้อยู่ในรูปแบบข้อความ ตรวจจับเนื้อหาของข้อความว่าคำใดมีรูปแบบที่เป็นข้อมูลที่สำคัญหรือข้อมูลส่วนบุคคล จากนั้นดำเนินการจับคู่คำกับเวลาในไฟล์บันทึกเสียง และดำเนินการปกปิดข้อความในส่วนนั้นออกไป

2. แนวคิด และเทคโนโลยีที่เกี่ยวข้อง

2.1 แนวคิดที่เกี่ยวข้อง

2.1.1 สิทธิความเป็นอยู่ส่วนบุคคล

มีการบัญญัติรับรองสิทธิดังกล่าวในรัฐธรรมนูญแห่งราชอาณาจักรไทย พ.ศ. 2560 มาตรา 32 ว่า “สิทธิของบุคคลในครอบครัว เกียรติยศ ชื่อเสียง ตลอดจนความเป็นอยู่ส่วนบุคคล ย่อมได้รับความคุ้มครอง การกล่าวหรือไขข่าวแพร่หลายซึ่งข้อความหรือภาพไม่ว่าด้วยวิธีใดไปยังสาธารณชนอันเป็นการละเมิดหรือกระทบถึงสิทธิของบุคคลในครอบครัว เกียรติยศ ชื่อเสียง หรือความเป็นอยู่

ส่วนบุคคล จะกระทำมิได้ เว้นแต่กรณีที่เป็น ประโยชน์ต่อสาธารณะ บุคคลย่อมมีสิทธิได้รับความคุ้มครองจากการแสวงประโยชน์โดยมิชอบจากข้อมูลส่วนบุคคลที่เกี่ยวกับตน ทั้งนี้ ตามที่กฎหมายบัญญัติ" [1]

2.2 เทคโนโลยีที่เกี่ยวข้อง

2.2.1 การทำเหมืองข้อมูล (Data Mining)

เป็นกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลชุดนั้น โดยอาศัยหลักการทางสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ [2]

2.2.2 การรู้จำเสียงพูด (Speech Recognition)

เป็นสิ่งที่ช่วยให้โปรแกรมสามารถประมวลผลคำพูดของมนุษย์ให้อยู่ในรูปแบบลายลักษณ์อักษร โดยเน้นที่การแปลงเสียงพูดจากรูปแบบคำพูดเป็นข้อความ [3]

2.2.3 Google Speech Recognition

พิมพ์อธิบายรายละเอียดสั้น ๆ พร้อม ref

2.2.4 การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP)

การประมวลผลภาษาธรรมชาติ คือ หนึ่งในสาขาของวิทยาศาสตร์คอมพิวเตอร์ที่เกี่ยวข้องกับปัญญาประดิษฐ์ (Artificial Intelligence) และภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) เป็นศาสตร์ที่ศึกษาเกี่ยวกับการทำให้คอมพิวเตอร์สามารถสื่อสารโต้ตอบด้วยภาษาของมนุษย์ และทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์มากขึ้น เช่น Siri, Google Assistant และ Alexa [4]

2.2.5 Stanford Named Entity Recognizer (Stanford NER)

เป็นการประยุกต์ใช้จากภาษาจาวา (Java) สำหรับการรู้จำนิพจน์ระบุนาม (Named Entity Recognizer: NER) ซึ่งเป็นการจัดประเภทของคำในข้อความ เช่น ชื่อสิ่งของ ชื่อบุคคล และบริษัท เป็นการกำหนดโครงสร้างการสกัดคุณสมบัติที่เหมาะสมสำหรับการรู้จำนิพจน์ระบุนาม (Named Entity Recognition: NER) [5]

2.2.6 Natural Language Toolkit (NLTK)

เป็นแพลตฟอร์มที่นิยมในโปรแกรมภาษาไพทอน (Python) เพื่อทำงานกับข้อมูลภาษาของมนุษย์ พร้อมกับชุดของไลบรารีที่ช่วยในการประมวลข้อความ แบ่งประเภทของคำ (Classification) การแบ่งโทเค็นของคำ (Tokenization) การตัดคำ (Stemming) การติดแท็กคำ (Tagging) และการแยกวิเคราะห์คำ (Parsing) [6]

2.2.7 spaCy

เป็นไลบรารีสำหรับการทำการระมวลผลภาษาธรรมชาติขั้นสูงในภาษาไพทอน (Python) โดยที่ spaCy ถูกออกแบบมาสำหรับการประยุกต์ใช้งานจริง และช่วยสร้างแอปพลิเคชันที่สามารถประมวลผล และทำความเข้าใจข้อความจำนวนมาก สามารถใช้ในการดำเนินการสกัดข้อมูล (Information Extraction) หรือระบบการทำความเข้าใจภาษาธรรมชาติเพื่อดำเนินการประมวลผลข้อความล่วงหน้าสำหรับการเรียนรู้เชิงลึก (Deep Learning) [7]

2.2.8 Regular Expressions

เป็นสัญลักษณ์ที่ใช้ระบุชุดของอักขระตัวอักษร เมื่อชุดของอักขระตัวอักษรที่เฉพาะเจาะจงนั้นอยู่ในชุดอักขระตัวอักษรที่มีการกำหนดให้เป็น Regular Expressions โดยทั่วไปแล้วจะใช้สัญลักษณ์ "*", "+", "?", "()" และ "|" ในการกำหนดเงื่อนไขของชุดตัวอักษร [8]

2.2.9 ...

.....

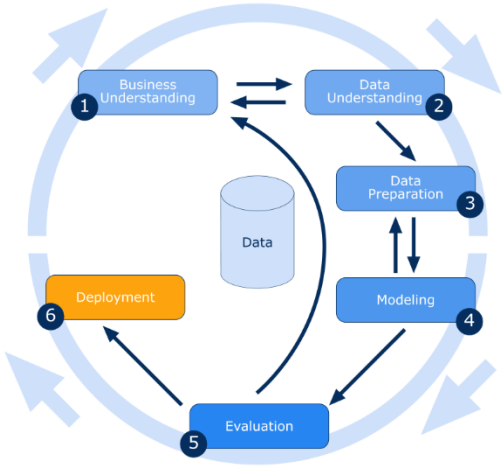
2.2.10 Jaccard's Coefficient Similarity

เป็นสถิติประยุกต์แนวคิดในทฤษฎีเซตเพื่อนำมาใช้เปรียบเทียบความคล้ายคลึงและความหลากหลายของกลุ่มตัวอย่าง แนวคิดของค่าสัมประสิทธิ์ Jaccard's Coefficient Similarity คือ การวัดค่าความคล้ายคลึงระหว่างกลุ่มประชากร 2 กลุ่ม โดยคำนวณจากขนาดของประชากรที่ทั้งสองกลุ่มมีตัวอย่างร่วมกัน (อินเตอร์เซกชันในทฤษฎีเซต) หารด้วยขนาดของประชากรทั้งหมดจากทั้งสองกลุ่มตัวอย่าง (ยูเนียนในทฤษฎีเซต) [10] ดังสมการที่ 1

$$Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \qquad (1)$$

3. ขั้นตอน และวิธีการดำเนินงานวิจัย

3.1 กระบวนการทำเหมืองข้อมูล (Data Mining Process)



รูปที่ 1. กระบวนการทำเหมืองข้อมูล

3.1.1 การทำความเข้าใจธุรกิจ (Business Understanding)

เมื่อเข้าสู่ยุคที่มีการแข่งขันสูง หลาย ๆ ธนาคารเริ่มนำเทคโนโลยีต่าง ๆ เข้ามาประยุกต์ใช้ในการให้บริการเพื่อเพิ่มความสะดวกสบายต่อลูกค้า รวมถึงต้องนำความพึงพอใจจากลูกค้า หรือปัญหาต่าง ๆ ทั้งทางออนไลน์ และการสนทนาผ่านโทรศัพท์ มาดำเนินการวิเคราะห์เพื่อเพิ่มประสิทธิภาพขององค์กรให้ดีที่สุด ข้อมูลส่วนบุคคลของลูกค้าจึงจำเป็นต้องมีการปกปิดก่อนจะเข้าสู่กระบวนการวิเคราะห์นั้น เพื่อป้องกันการละเมิดสิทธิส่วนบุคคลของลูกค้า และเพิ่มความน่าเชื่อถือขององค์กร

3.1.2 การทำความเข้าใจข้อมูล (Data Understanding)

ชุดข้อมูลประกอบไปด้วยชุดข้อมูลบทสนทนาระหว่างลูกค้าและศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ในรูปแบบข้อความ และชุดข้อมูลบทสนทนาระหว่างลูกค้าและศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ในรูปแบบเสียง ซึ่งรายละเอียดของข้อมูลในแต่ละบทสนทนาจะประกอบไปด้วยข้อมูลส่วนบุคคลของลูกค้า เช่น ชื่อ - นามสกุล ที่อยู่ เบอร์โทรศัพท์ วันเกิด เลขบัตรประชาชน เลขที่บัญชี และเลขหน้าบัตรเดบิต หรือบัตรเครดิต ต่าง ๆ ประเภทของการสนทนาประกอบไปด้วยการสนทนาประเภทสอบถามอัตราแลกเปลี่ยนของค่าเงินต่าง ๆ หรือรายงานปัญหาต่าง ๆ ของลูกค้า หรือการสอบถามรายละเอียดการทำธุรกรรมต่าง ๆ กับทางธนาคาร

3.1.3 การเตรียมข้อมูล (Data Preparation)

ทางผู้จัดทำได้ดำเนินการสร้างชุดข้อมูลบทสนทนาระหว่างลูกค้าและศูนย์บริการข้อมูลลูกค้าทางโทรศัพท์ในรูปแบบข้อความขึ้นเองเป็นจำนวนทั้งหมด 23 บทสนทนา จากการวิเคราะห์ประโยคในบทสนทนาคิดเป็น 566 ประโยค ค่าเฉลี่ยใน 1 บทสนทนาจะมีประโยคโดยเฉลี่ยจำนวน 24.61 ประโยค หากแบ่งย่อยลงไปเป็นการวิเคราะห์คำที่ยังไม่ผ่านการทำความสะอาดข้อมูลมีทั้งหมด 4,095 คำ ค่าเฉลี่ยใน 1 บทสนทนาจะมีคำโดยเฉลี่ยจำนวน 178.04 คำ และหากวิเคราะห์คำผ่านการทำความสะอาดข้อมูลแล้ว กล่าวคือ ดำเนินการตัดเครื่องหมายวรรคตอนและ Stop words บางส่วนออก มีทั้งหมด 1732 คำ ค่าเฉลี่ยใน 1 บทสนทนาจะมีคำโดยเฉลี่ยจำนวน 75.3 คำ

จากนั้นนำข้อมูลบทสนทนาที่ได้สร้างขึ้นมาดำเนินการบันทึกเสียง เนื่องจากบทสนทนาเป็นบทสนทนาภาษาอังกฤษ ทางผู้จัดทำได้มีการนำบทสนทนาไปบันทึกเสียงโดยใช้ระบบสั่งการด้วยเสียงของระบบปฏิบัติการ iOS หรือที่เป็นที่รู้จักกันในนามของ “สิริ” (Siri) ในการช่วยอ่านบทสนทนาเหล่านั้น ใน 1 บทสนทนาจะประกอบไปด้วยเสียงของพนักงานและลูกค้า โดยที่เสียงของพนักงานจะมีเพียงเพศเดียว คือ เพศหญิง โดยใช้เสียงของ “Siri Female” และในส่วนของเสียงลูกค้าจะแบ่งออกเป็น 2 เพศ ได้แก่ เพศหญิง ใช้เสียงของ “Samantha” และเพศชาย ใช้เสียงของ “Siri Male”

ประเภทไฟล์ของการบันทึกเสียงคือ “.m4a” ซึ่งทางผู้จัดทำจะต้องดำเนินการแปลงประเภทของไฟล์เสียงให้เป็น “.wav” เพื่อให้แบบจำลองการแปลงเสียงให้อยู่ในรูปแบบข้อความสามารถประมวลผลข้อมูลได้

3.1.4 กระบวนการพัฒนาแบบจำลอง (Modeling Process)

ขั้นตอนนี้แบ่งเป็น 3 กระบวนการหลัก ๆ ได้แก่ การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ การตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ และการจับคู่คำที่เป็นข้อมูลส่วนบุคคลกับระยะเวลาที่พูดในไฟล์เสียงเพื่อปกปิด มีรายละเอียดการดำเนินงาน ดังนี้

การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ

การตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ คือ หลังจากได้ดำเนินการแปลง

เสียงพูดให้อยู่ในรูปแบบข้อความโดยใช้ Google Speech Recognition แล้ว ข้อมูลที่ได้จะอยู่ในรูปแบบไฟล์ JSON จากนั้นจึงนำข้อมูลมาดำเนินการวิเคราะห์ต่อ เริ่มจากกระบวนการตรวจจับนิพจน์ระบุนาม (Named Entities Tagger Process) ขั้นตอนนี้มีการใช้แบบจำลองทั้งหมด 3 แบบจำลอง เพื่อเพิ่มความแม่นยำในการตรวจจับนิพจน์ระบุนาม ได้แก่ Stanford NER, NLTK และ spaCy มีกระบวนการดำเนินงาน ดังนี้

- พัฒนาแบบจำลองของ Stanford NER โดยเลือกประเภทของนิพจน์ระบุนามในการติดแท็กบทสนทนาทั้งหมดเป็นจำนวน 5 ประเภท ได้แก่ PERSON, ORGANIZATION, LOCATION, DATE และ MONEY ซึ่งในฟังก์ชันมีการทำ Word Tokenization เพื่อแยกโทเค็นของคำในข้อความ ต่อมามีการติดแท็กนิพจน์ระบุนามจากอัลกอริทึมของ Stanford NER จากนั้นสร้างเงื่อนไขเก็บเฉพาะโทเค็นที่เป็นนิพจน์ระบุนามเท่านั้น จากนั้นจึงแก้ไขประเภทของนิพจน์ระบุนามที่ถูกติดแท็กเพื่อให้ประเภทของนิพจน์ระบุนามตรงกับแบบจำลองอื่น ๆ เช่น คำว่า “ORG” ที่ทางแบบจำลองติดแท็กไว้ จะดำเนินการเปลี่ยนเป็นคำว่า “ORGANIZATION” เพื่อให้ตรงกับแบบจำลองทั้ง 2 แบบ และสะดวกต่อการนำไปประเมินผล จากนั้นทำการจับคู่โทเค็นที่แบบจำลองแบ่งออกมาเทียบกับโทเค็นที่ Google Speech Recognition แบ่งไว้ เพื่อให้แน่ใจว่าโทเค็นที่ถูกติดแท็กนั้นตรงกับระยะเวลาที่ Google Speech Recognition ทำนายออกมา และเก็บค่าของคำที่ติดแท็กได้ พร้อมกับประเภทของนิพจน์ระบุนาม
- พัฒนาแบบจำลองของ NLTK โดยทางผู้จัดทำได้ดำเนินการเลือกการติดแท็กในบทสนทนาเป็นจำนวนทั้งหมด 6 ประเภท ได้แก่ ORGANIZATION, PERSON, LOCATION, GPE, DATE และ MONEY โดยเริ่มจากการทำ Word Tokenization จากนั้นทำการติดแท็กนิพจน์ระบุนามจากอัลกอริทึม NLTK ซึ่งต้องมีการติดแท็กส่วนของประโยค (Part-of-Speech) ก่อนจึงจะติดแท็กได้ และกระบวนการหลังจากนั้นมีวิธีการทำเช่นเดียวกันกับ Stanford NER คือ เลือกโทเค็นที่เป็นนิพจน์ระบุนาม และทำการเปลี่ยนประเภทนิพจน์ระบุนามให้เหมือนกันทุกแบบจำลอง จากนั้นจับคู่โทเค็นที่แบบจำลองแบ่งเทียบกับโทเค็นของ Google Speech Recognition และเก็บค่าของโทเค็น
- พัฒนาแบบจำลองของ spaCy โดยทางผู้จัดทำได้ดำเนินการเลือกการติดแท็กในบทสนทนาเป็นจำนวนทั้งหมด 6 ประเภท ได้แก่ ORGANIZATION, PERSON, LOCATION, GPE, DATE และ MONEY ในฟังก์ชันมีการใช้อัลกอริทึมของ spaCy ซึ่งในอัลกอริทึมจะดำเนินการวิเคราะห์ข้อความต่าง ๆ อัตโนมัติ ซึ่งสามารถเรียกดูค่าได้จากอัลกอริทึมได้ทันที และกระบวนการหลังจากนั้นมีวิธีการทำเช่นเดียวกันกับ Stanford NER และ NLTK คือ เลือกโทเค็นที่เป็นนิพจน์ระบุนาม และทำการเปลี่ยนประเภทนิพจน์ระบุนามให้เหมือนกันทุกแบบจำลอง จากนั้นจับคู่โทเค็นที่แบบจำลองแบ่งเทียบกับโทเค็นของ Google Speech Recognition และเก็บค่าของโทเค็น

ต่อมาดำเนินการเลือกการทำนายประเภทของนิพจน์ระบุนามที่เหมือนกันตั้งแต่ 2 จาก 3 โมเดลขึ้นไป จากการสร้างฟังก์ชันจับคู่โทเค็นที่มีการทำนายนิพจน์ระบุนามค่าเดียวกัน และเก็บค่าของโทเค็นนั้นใหม่ เพื่อนำไปใช้วิเคราะห์กระบวนการถัดไป ในที่นี้ ทางผู้จัดทำขอแทนผลลัพธ์ของกระบวนการนี้ว่าค่าทำนายจริง

ขั้นตอนสุดท้ายคือการสร้างนิพจน์ระบุนามเพิ่ม เพื่อติดแท็กเลขที่เป็นข้อมูลส่วนบุคคลโดยใช้ Regular Expressions ขั้นตอนนี้จะมีการดึงโทเค็นคำของ Google Speech Recognition เฉพาะที่เป็นเลขมาตรวจสอบเงื่อนไขเพื่อติดแท็กเลขที่เป็นข้อมูลส่วนบุคคลเท่านั้น โดยแบ่งประเภทของเลขที่เป็นข้อมูลส่วนบุคคลไว้ 5 ประเภท คือ IDCARD (เลขบัตรประชาชน 13 หลัก) PHONENUM (เบอร์โทรศัพท์ 10 หลัก) ACCNUM (เลขบัญชี 9 หลัก) CARDNUM (เลขบัตรเดบิต หรือบัตรเครดิต 16 หลัก) และ PIINUM (เลขอื่น ๆ ที่ไม่เข้าเงื่อนไขประเภทก่อนหน้านี้ แต่มีตั้งแต่ 9 หลักขึ้นไป มีไว้ในกรณีที่ Google Speech Recognition แปลงเป็นข้อความออกมาได้ไม่แม่นยำ) จากนั้นนำค่าที่ได้ไปรวมกับค่าทำนายจริง และเก็บค่านั้นไว้ในรูปแบบไฟล์ CSV เพื่อนำไปดำเนินการต่อในขั้นถัดไป

การจับคู่คำที่เป็นข้อมูลส่วนบุคคลกับระยะเวลาที่พูดในไฟล์เสียงเพื่อปกปิด

3.1.5 การประเมินผล (Evaluation)

มีการประเมินผลกระบวนการทั้งหมด 2 กระบวนการหลัก ๆ คือ ประเมินผลความแม่นยำของการแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ และกระบวนการ

ประเมินผลความแม่นยำของการตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ จากการสร้างผลเฉลยของการทำนายข้อความและโทเค็นต่าง ๆ เพื่อใช้ตรวจสอบความแม่นยำในการทำนายของแบบจำลองทั้งหมด และในส่วนของการประเมินผลความแม่นยำของการแปลงเสียงพูดให้อยู่ในรูปแบบข้อความนั้น ได้มีการนำแนวคิดของ Jaccard's Coefficient Similarity มาประยุกต์ใช้ในการประเมินผล

3.1.6 การนำไปใช้จริง (Deployment)

หลังจากที่ทำการประเมินผลการทำนายแล้ว จึงนำมาประยุกต์ใช้กับองค์กรต่าง ๆ ที่ต้องการรักษาความเป็นส่วนบุคคลของลูกค้า โดยการนำชุดข้อมูลเสียงที่บันทึกไว้ทั้งหมดเข้าสู่แบบจำลองการปกปิดข้อมูลที่ระบุตัวบุคคล จากนั้นระบบจะดำเนินการปกปิดคำที่เป็นข้อมูลส่วนบุคคลจากไฟล์เสียงนั้น เพื่อให้สามารถนำข้อมูลส่วนอื่นไปวิเคราะห์ทางธุรกิจในด้านต่าง ๆ ได้

4. ผลการดำเนินงานเบื้องต้น

4.1 การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ


```
{'transcript': "Hello, you have called virtual bank. This is Linda speaking. How may I help you? Hi Linda. I was just at your bill branch and I think I left my debit card in the ATM machine. Okay. Do you have your debit card number? I don't know. Okay. Well, do you have the checking account number associated with the debit card, but I do have are you ready? I will give you what I have got 760-545-6789. Okay. That's +765-450-600-7089. Correct? What is your identification number? 774-589-6589 665 okay. I have +774-580-960-5896 65 and what is your name sir? It is Robert. Appel board. Okay.I have Robert Applebaum yet. And what is your date of birth Mr. Appelbaum, July 7th, 1974. Okay, July 7th, 1974. Yes, and your phone number. It is 610-265-1715. Okay, I have 610-265-1715. Yes. Okay, Mr. Appelbaum. I have just this pended your card. If it is in the machine, we will contact you as lift the suspension 00. Thank you, sir. Thank you.", 'values': {'start': [0.0, 0.4, 1.2, 1.3, 1.8, 2.2, 2.4, 3.2, 3.4, 3.8, 4.3, 5.3, 5.3, 5.5, 5.7, 6.2, 6.8, 7.2, 8.0, 8.2, 8.3, 8.7, 8.8, 9.0, 9.5, 9.8, 9.8, 10.0, 10.2, 10.4, 10.7, 11.1, 11.2, 11.6, 11.7, 11.8, 12.3, 13.1, 14.2, 14.2, 14.4, 14.6, 15.0, 15.1, 15.4, 16.4, 16.5, 16.7, 18.2, 18.9, 19.2, 19.3, 19.4, 19.6, 19.9, 20.5, 20.8, 21.1, 21.8, 21.9, 22.3, 22.4, 23.1, 23.3, 23.4, 23.6, 24.6, 24.8, 25.1, 25.9, 26.1, 26.2, 26.5, 26.6, 26.7, 26.8, 27.2, 30.6, 31.8, 32.7, 36.0, 37.1, 37.2, 37.3, 37.5, 38.1, 38.9,
```


รูปที่ 2. ตัวอย่างการแปลงข้อมูลเสียงให้อยู่ในรูปแบบข้อความโดยใช้ Google Speech Recognition

จากรูปที่ 2 ทางผู้จัดทำได้ดำเนินการแปลงให้อยู่ในรูปแบบของ Dictionary และสร้างคีย์ที่ชื่อว่า transcript ไว้เก็บข้อความในบทสนทนาทั้งหมด ในส่วนของโทเค็นคำ ได้มีการสร้างคีย์ที่ชื่อว่า values ไว้เก็บค่าของเวลาที่เริ่มพูดโทเค็นนั้น ๆ (start) เวลาที่พูดจบ (end) และโทเค็นนั้น ๆ (word)

นอกจากนี้ ยังได้มีการประเมินผลความแม่นยำในการทำนายของแบบจำลอง โดยการนำข้อมูลบทสนทนาจริงเทียบกับข้อมูลที่แบบจำลองทำนายโดยใช้ Jaccard's Coefficient Similarity ดังนี้


```
'Hello, you have called virtual bank, this is Linda speaking. How may I help you? Hi Linda. I was just at your Ville branch and I think I left my Debit card in the ATM machine. Okay. Do you have your Debit card number? I don't have. Okay, well do you have the checking account number associated with the Debit card? That I do have. Are you ready? I will give you what I have got. 765-456-789. Okay. That's 765-456-789. Correct. What is your identification number? 774-589-658-9665. Okay, I have 774-589-658-9665 and what is your name sir? It is Robert Applebaum. Okay. I have Robert Applebaum. Yes. And what is your date of birth Mr. Applebaum? July 7th, 1974. Okay. July 7th, 1974. Yes. And your phone number? It is 610-265-1715. Okay. I have 610-2651715. Yes. Okay Mr. Applebaum. I have just suspended your card. If it is in the machine, we will contact you and lift the suspension. Oh, thank you, Sure. Thank you.'
```


รูปที่ 3. ข้อมูลบทสนทนาจริง


```
"Hello, you have called virtual bank. This is Linda speaking. How may I help you? Hi Linda. I was just at your bill branch and I think I left my debit card in the ATM machine. Okay. Do you have your debit card number? I don't know. Okay. Well, do you have the checking account number associated with the debit card, but I do have are you ready? I will give you what I have got 760-545-6789. Okay. That's +765-450-600-7089. Correct? What is your identification number? 774-589-6589 665 okay. I have +774-580-960-5896 65 and what is your name sir? It is Robert. Appel board. Okay.I have Robert Applebaum yet. And what is your date of birth Mr. Appelbaum, July 7th, 1974. Okay, July 7th, 1974. Yes, and your phone number. It is 610-265-1715. Okay, I have 610-265-1715. Yes. Okay, Mr. Appelbaum. I have just this pended your card. If it is in the machine, we will contact you as lift the suspension 00. Thank you, sir. Thank you."
```


รูปที่ 4. บทสนทนาที่แบบจำลองทำนาย

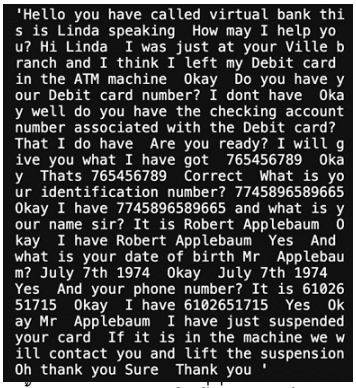
```
acc = Jaccard_Similarity(dict_, ori_text)
acc = acc*100

print('Accuracy of the conversation:', '%.2f' %acc, '%')
```

```
Accuracy of the conversation: 57.02 %
```

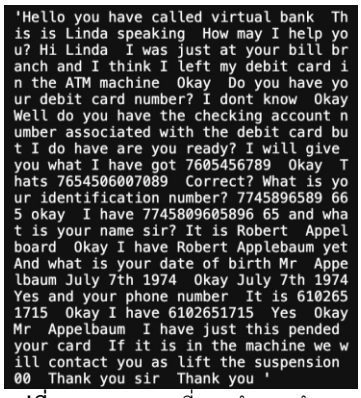
รูปที่ 5. ค่าของความแม่นยำในการทำนาย

จากรูปที่ 5 ความแม่นยำในการทำนายคิดเป็นร้อยละ 57.02 ซึ่งเมื่อเทียบบทสนทนารูปที่ 3 และ 4 จะสังเกตได้ว่าสิ่งที่ส่งผลให้ค่าความแม่นยำของแบบจำลองไม่สูงนั้นส่วนใหญ่แล้วขึ้นอยู่กับเครื่องหมายวรรคตอนของข้อมูลบทสนทนาทั้งสอง ดังนั้น จึงดำเนินการสร้างฟังก์ชันตัดเครื่องหมายวรรคตอนของบทสนทนาทั้งสองออก เพื่อประเมินผลค่าความแม่นยำใหม่ ดังรูปที่ 6, 7 และ 8



```
'Hello you have called virtual bank thi
s is Linda speaking  How may I help yo
u? Hi Linda  I was just at your Ville b
ranch and I think I left my Debit card
in the ATM machine  Okay  Do you have y
our Debit card number? I dont have  Oka
y well do you have the checking account
number associated with the Debit card?
That I do have  Are you ready? I will g
ive you what I have got  765456789  Oka
y  Thats 765456789  Correct  What is yo
ur identification number? 7745896589665
Okay I have 7745896589665 and what is y
our name sir? It is Robert Applebaum  O
kay  I have Robert Applebaum  Yes  And
what is your date of birth Mr  Applebau
m? July 7th 1974  Okay  July 7th 1974
Yes  And your phone number? It is 61026
51715  Okay  I have 6102651715  Yes  Ok
ay Mr  Applebaum  I have just suspended
your card  If it is in the machine we w
ill contact you and lift the suspension
Oh thank you Sure  Thank you '
```

รูปที่ 6. ข้อมูลบทสนทนาจริงที่ผ่านการทำความสะอาด



```
'Hello you have called virtual bank  Th
is is Linda speaking  How may I help yo
u? Hi Linda  I was just at your bill br
anch and I think I left my debit card i
n the ATM machine  Okay  Do you have yo
ur debit card number? I dont know  Okay
Well do you have the checking account n
umber associated with the debit card bu
t I do have are you ready? I will give
you what I have got 7605456789  Okay  T
hats 7654506007089  Correct? What is yo
ur identification number? 7745896589 66
5 okay  I have 7745809605896 65 and wha
t is your name sir? It is Robert  Appel
board  Okay I have Robert Applebaum yet
And what is your date of birth Mr  Appe
lbaum July 7th 1974  Okay July 7th 1974
Yes and your phone number  It is 610265
1715  Okay I have 6102651715  Yes  Okay
Mr  Appelbaum  I have just this pended
your card  If it is in the machine we w
ill contact you as lift the suspension
00  Thank you sir  Thank you '
```

รูปที่ 7. บทสนทนาที่แบบจำลองทำนายที่ผ่านการทำความสะอาด

```
acc = Jaccard_Similarity(clean_text(dict_)/
    , clean_text1(ori_text))
acc = acc*100

print('Accuracy of the conversation:', '%.2
Accuracy of the conversation: 71.43 %
```

รูปที่ 8. ค่าของความแม่นยำในการทำนาย (ใหม่)

จากรูปที่ 8 ความแม่นยำในการทำนายคำพูดของแบบจำลองคิดเป็นร้อยละ 71.43 สามารถเห็นได้ชัดว่าค่าความแม่นยำสูงขึ้นอย่างชัดเจน เมื่อตัดเครื่องหมายวรรคตอนออกเบื้องต้น

4.2 การตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ

เมื่อดำเนินการนำข้อมูลในรูปแบบข้อความที่ได้จาก Google Speech Recognition มาเข้าฟังก์ชันต่าง ๆ ของแบบจำลอง Stanford NER, NLTK และ spaCy พร้อมกับนำเข้าฟังก์ชันของการเลือกค่าทำนายจริง และสร้างนิพจน์ระบุนามเพิ่มสำหรับเลขที่เป็นข้อมูลส่วนบุคคลโดยใช้ Regular Expressions ทางผู้จัดทำก็ได้ดำเนินการเก็บค่าของการทำนายของทุก ๆ แบบจำลองไว้ในรูปแบบตาราง ดังรูปที่ 9

indx	word	start_time	end_time	stanford_pred	nltk_pred	spacy_pred	real_ents
0	Hello,	0.0	0.4	DATE	LOCATION	O	O
1	you	0.4	1.2	O	O	O	O
2	have	1.2	1.3	O	O	O	O
3	called	1.3	1.8	O	O	O	O
4	virtual	1.8	2.2	O	O	O	O
5	bank.	2.2	2.4	O	O	O	O
6	This	2.4	3.2	O	O	O	O
7	is	3.2	3.4	O	O	O	O
8	Linda	3.4	3.8	PERSON	PERSON	PERSON	PERSON
9	speaking.	3.8	4.3	O	O	O	O

รูปที่ 9. ตารางการทำนายประเภทของนิพจน์ระบุนาม

จากรูปที่ 9 ทางผู้จัดทำได้เก็บค่าการทำนายของโทเค็นทุก ๆ คำ ไว้ในตารางเดียวกันตามประเภทของนิพจน์ระบุนาม แถวใดมีการทำนายเป็นคำว่า “O” หมายความว่าโทเค็นนั้นไม่ได้เป็นนิพจน์ระบุนาม และมีการเก็บค่าการทำนายทั้งหมด 4 คอลัมน์ ได้แก่ stanford_pred (ค่าที่แบบจำลอง Stanford NER ทำนาย) nltk_pred (ค่าที่ NLTK ทำนาย) spacy_pred (ค่าที่ spaCy ทำนาย) และคอลัมน์สุดท้าย real_ents (ค่าทำนายที่แท้จริง จากการเลือกค่าทำนายที่เหมือนกันตั้งแต่ 2 ใน 3 ของแบบจำลอง และการติดแท็กค่าของเลขที่เป็นข้อมูลส่วนบุคคลจากการใช้ Regular Expressions) นอกจากนี้ ยังได้ดำเนินการเก็บบันทึกค่าการทำนายจริง เฉพาะโทเค็นที่มีการติดแท็กนิพจน์ระบุนามขึ้นมาอีก 1 ตาราง เพื่อดำเนินการบันทึกให้อยู่ในรูปแบบไฟล์ CSV และนำไปปกปิดเสียงในขั้นตอนถัดไป ดังรูปที่ 10

indx	word	start_time	end_time	real_ents
8	Linda	3.4	3.8	PERSON
16	Linda.	6.8	7.2	PERSON
34	ATM	11.7	11.8	ORGANIZATION
76	760-545-6789.	27.2	30.6	PHONENUM
79	+765-450-600-7089.	32.7	35.7	IDCARD
86	774-589-6589	38.9	42.7	PHONENUM
91	+774-580-960-5896	45.4	49.0	IDCARD
101	Robert.	51.9	52.3	PERSON
107	Robert	55.0	55.4	PERSON
108	Applebaum	55.4	56.0	PERSON
118	Appelbaum,	59.8	60.3	PERSON
119	July	60.3	61.6	DATE
120	7th,	61.6	62.1	DATE
121	1974.	62.1	63.5	DATE
123	July	64.9	66.0	DATE
124	7th,	66.0	66.6	DATE
125	1974.	66.6	68.3	DATE
133	610-265-1715.	71.9	75.4	PHONENUM
137	610-265-1715.	77.4	80.7	PHONENUM
141	Appelbaum.	83.1	83.6	PERSON

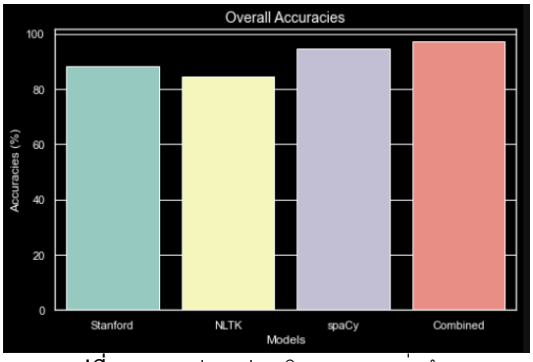
รูปที่ 10. ตารางค่าทำนายจริงเฉพาะที่มีการติดแท็กนิพจน์ระบุนาม

ทางผู้จัดทำมีการประเมินผลความแม่นยำในการทำนายนิพจน์ระบุนามของแต่ละแบบจำลอง โดยการนำโทเค็นที่ Google Speech Recognition แบ่งออกมาไปทำการเฉลยนิพจน์ระบุนามจริง เพื่อที่จะนำไปประเมินผลความแม่นยำของการทำนายนิพจน์ระบุนามในทุก ๆ แบบจำลอง

```
Stanford Accuracy: 88.17%
NLTK Accuracy: 84.62%
spaCy Accuracy: 94.67%

** Combined Models and using Regular Expressions Accuracy:
97.04% **
```

รูปที่ 11. การประเมินผลความแม่นยำของแต่ละแบบจำลอง



รูปที่ 12. กราฟการประเมินผลความแม่นยำของแต่ละแบบจำลอง

จากรูปที่ 11 สามารถสรุปได้ ดังนี้

- ความแม่นยำของการทำนายนิพจน์ระบุนามของ Stanford NER คิดเป็นร้อยละ 88.17
- ความแม่นยำของการทำนายนิพจน์ระบุนามของ NLTK คิดเป็นร้อยละ 84.62
- ความแม่นยำของการทำนายนิพจน์ระบุนามของ spaCy คิดเป็นร้อยละ 94.67
- ความแม่นยำของการทำนายนิพจน์ระบุนามของการรวมแบบจำลองและการทำ Regular Expressions คิดเป็นร้อยละ 97.04

สังเกตได้ว่า เมื่อดำเนินการรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกัน และสร้างเงื่อนไขจาก Regular Expressions นั้น ส่งผลให้ค่าความแม่นยำในการทำนายนิพจน์ระบุนามสูงที่สุด

นอกจากนี้ ทางผู้จัดทำได้ประเมินผลความแม่นยำของนิพจน์ระบุนามในแต่ละประเภท เพื่อวิเคราะห์ว่าประเภทใดมีค่าความแม่นยำแตกต่างกันอย่างไร สามารถสรุปได้ ดังนี้

```
Stanford Accuracy: 98.82%
NLTK Accuracy: 93.49%
spaCy Accuracy: 100.00%

** Combined Models and using Regular Expressions Accuracy:
100.00% **
```

รูปที่ 13. การประเมินผลความแม่นยำในการติดแท็กคำว่า “PERSON”

จากรูปที่ 13 ความแม่นยำในการติดแท็กคำว่า “PERSON” ของแบบจำลอง Stanford NER คิดเป็นร้อยละ 98.82 แบบจำลอง NLTK คิดเป็นร้อยละ 93.49 แบบจำลอง spaCy คิดเป็นร้อยละ 100 และเมื่อรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 100

```
Stanford Accuracy: 99.41%
NLTK Accuracy: 100.00%
spaCy Accuracy: 99.41%

** Combined Models and using Regular Expressions Accuracy:
99.41% **
```

รูปที่ 14. การประเมินผลความแม่นยำในการติดแท็กคำว่า “ORGANIZATION”

จากรูปที่ 14 ความแม่นยำในการติดแท็กคำว่า “ORGANIZATION” ของแบบจำลอง Stanford NER คิดเป็นร้อยละ 99.41 แบบจำลอง NLTK คิดเป็นร้อยละ 100 แบบจำลอง spaCy คิดเป็นร้อยละ 99.41 และเมื่อรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 99.41

```
Stanford Accuracy: 98.82%
NLTK Accuracy: 99.41%
spaCy Accuracy: 100.00%

** Combined Models and using Regular Expressions Accuracy:
100.00% **
```

รูปที่ 15. การประเมินผลความแม่นยำในการติดแท็กคำว่า “LOCATION”

จากรูปที่ 15 ความแม่นยำในการติดแท็กคำว่า “LOCATION” ของแบบจำลอง Stanford NER คิดเป็นร้อยละ 98.82 แบบจำลอง NLTK คิดเป็นร้อยละ 99.41 แบบจำลอง spaCy คิดเป็นร้อยละ 100 และเมื่อรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 100

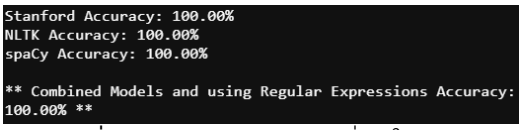
```
Stanford Accuracy: 94.67%
NLTK Accuracy: 96.45%
spaCy Accuracy: 100.00%

** Combined Models and using Regular Expressions Accuracy:
100.00% **
```

รูปที่ 16. การประเมินผลความแม่นยำในการติดแท็กคำว่า “DATE”

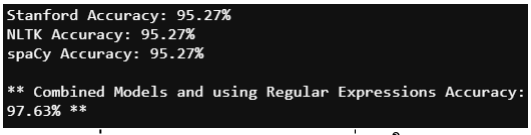
จากรูปที่ 16 ความแม่นยำในการติดแท็กคำว่า “DATE” ของแบบจำลอง Stanford NER คิดเป็นร้อยละ 94.67 แบบจำลอง NLTK คิดเป็นร้อยละ 96.45 แบบจำลอง spaCy คิดเป็นร้อยละ 100 และเมื่อรวมการทำนายของ

แต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 100



รูปที่ 17. การประเมินผลความแม่นยำในการติดแท็กคำว่า "MONEY"

จากรูปที่ 17 ความแม่นยำในการติดแท็กคำว่า "MONEY" ของแบบจำลอง Stanford NER คิดเป็นร้อยละ 100 แบบจำลอง NLTK คิดเป็นร้อยละ 100 แบบจำลอง spaCy คิดเป็นร้อยละ 100 และเมื่อรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 100 ในบางครั้งอาจสรุปได้ว่าบทสนทนานี้ไม่มีการกล่าวถึงค่าเงิน จึงส่งผลให้แบบจำลองทุกแบบมีค่าความแม่นยำสูงสุด



รูปที่ 18. การประเมินผลความแม่นยำในการติดแท็กติดแท็ก PII Number ทุกประเภท

จากรูปที่ 18 ทางผู้จัดทำได้ประเมินผลความแม่นยำของเลขที่เป็นข้อมูลส่วนบุคคลทุก ๆ ประเภทเข้าด้วยกัน สามารถสรุปได้ว่า ความแม่นยำในการติดแท็กประเภทของ PII Number ทุกประเภทของแบบจำลอง Stanford NER คิดเป็นร้อยละ 95.27 แบบจำลอง NLTK คิดเป็นร้อยละ 95.27 แบบจำลอง spaCy คิดเป็นร้อยละ 95.27 และเมื่อรวมการทำนายของแต่ละแบบจำลองเข้าด้วยกันพร้อมกับสร้างเงื่อนไขจาก Regular Expressions มีความแม่นยำคิดเป็นร้อยละ 97.63 สาเหตุที่แบบจำลองทั้ง 3 แบบมีค่าความแม่นยำเท่ากันเพราะไม่ได้มีการติดแท็กเลขในแบบจำลองทั้ง 3 แบบ แต่มีการติดแท็กในการรวมแบบจำลองเท่านั้น

4.3 การจับคู่คำที่เป็นข้อมูลส่วนบุคคลกับระยะเวลาที่พูดในไฟล์เสียงเพื่อปกปิด

อธิบายยยยยยย

5. บทสรุป

5.1 สรุปผลโครงงาน

5.1.1 การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความ

การแปลงเสียงพูดให้อยู่ในรูปแบบข้อความนั้น หากเป็นการประเมินผลโดยไม่คำนึงถึงความถูกต้องของเครื่องหมายวรรคตอน ถือว่าค่าของความแม่นยำอยู่ในระดับที่ดี อาจจะมีการแปลงชื่อบุคคลที่ไม่ตรงกับข้อมูลบทสนทนาจริงเล็กน้อย อาจเป็นสาเหตุมาจากเสียงที่ใช้ในการดำเนินการบันทึกเสียงที่แต่ละบุคคลมีสำเนียงการพูดที่ไม่เหมือนกัน เช่น นามสกุล Applebaum เมื่อเป็นเสียงของ Siri Male ทางแบบจำลองแปลงได้เป็น 2 โทเค็น คือ "Appel" และ "board." แต่เมื่อเป็นเสียงของ "Siri Female" ทางแบบจำลองกลับแปลงคำได้ถูกต้อง จึงสรุปได้ว่าบางครั้งสำเนียงการพูดของแต่ละตัวบุคคลอาจส่งผลต่อความแม่นยำของการแปลงข้อมูลเสียงให้อยู่ในรูปข้อความ นอกจากนี้ ยังมีการแปลงเลขที่ผิดพลาดไปบ้าง เช่น เมื่อสิริพูดว่า "oh" ในบางครั้งแบบจำลองจะแปลงเป็นเลข "0" ซึ่งส่งผลให้ความแม่นยำของแบบจำลองลดลง

5.1.2 การตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ

ในขั้นตอนนี้ ผู้จัดทำจะอธิบายรายละเอียดของแต่ละแบบจำลอง ดังนี้

- Stanford NER สามารถติดแท็กบุคคล และค่าเงินได้ค่อนข้างแม่นยำ ส่วนนิพจน์ระบุนามประเภทอื่น ๆ มีความแม่นยำเฉลี่ยเท่า ๆ กันกับแบบจำลองอื่น ๆ แต่ในการติดแท็กวันที่ ด้วยข้อจำกัดของแบบจำลองที่ไม่มีการติดแท็กตัวเลขที่เป็นประเภท Cardinal เหมือนแบบจำลองอื่น จึงส่งผลให้มีการติดแท็กตัวเลขธรรมดาเป็นประเภทของวันที่ (Date) ทำให้ความแม่นยำของแบบจำลองลดลง
- NLTK สามารถติดแท็กองค์กรได้แม่นยำมากที่สุด ส่วนนิพจน์ระบุนามประเภทอื่น ๆ มีความแม่นยำเฉลี่ยเท่า ๆ กันกับแบบจำลองอื่น ๆ แต่แบบจำลองนี้มักมีการติดแท็กที่ผิดพลาดตรงส่วนของสถานที่ กล่าวคือ หากโทเค็นนั้นขึ้นต้นด้วยตัวอักษรพิมพ์ใหญ่ เช่น คำว่า "Hello" แบบจำลองจะติดแท็กเป็นสถานที่ทันที
- spaCy จากผลลัพธ์การประเมินผลความแม่นยำ จะสังเกตได้ว่าส่วนใหญ่แล้ว spaCy มีค่าความแม่นยำสูงในการติดแท็กโทเค็น แต่หากให้สรุปเป็นรายประเภท จะสามารถสรุปได้ว่า แบบจำลองนี้สามารถติดแท็กบุคคล

สถานที่ วันที่ และค่าเงินได้ดีที่สุด ส่วนนิพจน์ระบุนามประเภทอื่น ๆ มีความแม่นยำเฉลี่ยเท่า ๆ กันกับแบบจำลองอื่น ๆ แต่เนื่องจากการติดแท็กของแบบจำลองนี้ยังมีความไม่แม่นยำบ้าง ทางผู้จัดทำจึงมีความเห็นว่าควรรวมแบบจำลองเข้าด้วยกันเพื่อเพิ่มประสิทธิภาพในการติดแท็ก

ในส่วนของการรวมแบบจำลองเข้าด้วยกัน มีความแม่นยำค่อนข้างสูง ซึ่งเฉลี่ยแล้วคิดเป็นร้อยละ 90 ถือเป็นค่าความแม่นยำที่น่าพึงพอใจ

และการตรวจจับเลขที่เป็นข้อมูลส่วนบุคคลโดยใช้ Regular Expressions ก็มีความแม่นยำค่อนข้างสูง แต่ในบางครั้งอาจไม่แม่นยำอย่างสมบูรณ์เนื่องจากรูปแบบการแปลงตัวเลขของ Google Speech Recognition อาจแบ่งโทเค็นได้ไม่ตรงกับตัวเลขที่ควรจะเป็น เช่น เลขบัตรเดบิต หรือบัตรเครดิต 16 หลัก ทางแบบจำลองอาจมีรูปแบบการแปลงตัวเลขได้เพียงแค่ 13 หลัก แล้วจึงแบ่งเลขอีก 3 หลักหลังเป็นอีกโทเค็น ซึ่งในเงื่อนไขมักจะติดแท็กเลขที่มากกว่า 9 หลักขึ้นไปโดยไม่สนใจเครื่องหมายต่าง ๆ เช่น +111-111-111-1111 หรือ 111-111-1111 เป็นต้น แต่หากพิจารณาถึงภาพรวมของค่าความแม่นยำแล้ว ถือเป็นที่น่าพึงพอใจ

5.1.3 การจับคู่คำที่เป็นข้อมูลส่วนบุคคลกับระยะเวลาที่พูดในไฟล์เสียงเพื่อปกปิด

อธิบาย

5.2 ปัญหาในการทำโครงงานและสรุปผล

โดยส่วนใหญ่แล้ว ปัญหาในการทำโครงงานนี้ คือ ความแม่นยำของการแปลงข้อมูลเสียงให้อยู่ในรูปแบบข้อความนั้น มีความแม่นยำในระดับปานกลางจนถึงค่อนข้างสูง แต่เมื่อดำเนินการเข้าสู่กระบวนการตรวจจับคำที่เป็นข้อมูลส่วนบุคคลจากข้อมูลรูปแบบข้อความ ส่งผลให้แบบจำลองไม่สามารถติดแท็กประเภทของโทเค็นที่ควรจะมีนิพจน์ระบุนามได้ เช่น ชื่อบุคคล หรือส่วนเล็ก ๆ ของเลขที่เป็นข้อมูลสำคัญ จึงอาจส่งผลให้เป็นปัญหาต่อการปิดบังคำที่เป็นข้อมูลส่วนบุคคลในขั้นตอนสุดท้ายได้

5.3 แนวทางในการพัฒนาต่อ

ทางผู้จัดทำจะดำเนินการหาวิธีการเพิ่มค่าความแม่นยำของการแปลงข้อมูลเสียงให้อยู่ในรูปแบบข้อความให้มีความแม่นยำมากขึ้น เพื่อให้การติดแท็กโทเค็นตรงเงื่อนไขมากที่สุด และอาจมีการดำเนินการพัฒนาต่อเพิ่มในด้านของการตรวจจับข้อมูลส่วนบุคคล เช่น หลังจากที่ติดแท็กโทเค็นนั้นแล้ว อาจมีการฝึกฝนแบบจำลองอื่น ๆ เพิ่มเติม เพื่อตรวจจับว่าโทเค็นนั้น ๆ เป็นข้อมูลส่วนบุคคลที่จำเป็นต้องปกปิดจริงหรือไม่ แต่ด้วยวิธีการนั้นอาจจะต้องดำเนินการสร้างชุดข้อมูลพร้อมกับการเฉลยผลการตรวจจับว่าเป็นข้อมูลส่วนบุคคลหรือไม่ เป็นจำนวนมาก เพื่อให้แบบจำลองสามารถทำนายได้อย่างแม่นยำ

เอกสารอ้างอิง

[1] A. B. Green, C. D. Black, and E. F. White, "Article Title," *Journal*, vol. 100, no. 1, pp. 1-10, Dec. 2000.

[2] C. D. Black, A. B. Green, and E. F. White, *Book Title*, 3rd ed. New York: McGraw-Hill, 2001.

[3] สมชาย สกุลดี. "ชื่อบทความ". **ชื่อวารสาร** ปีที่ 10, ฉบับที่ 2 (10 กุมภาพันธ์ 2553). หน้า 10-15.

[4] สมหญิง เจริญดี. **ชื่อหนังสือ.** พิมพ์ครั้งที่ 2. กรุงเทพฯ: สำนักพิมพ์เจริญทัศน์, 2553.

[5] J. K. Pink, "Article Title," in *Proc. International Conference on Green Computing*, Paris, France, Jan. 2012, pp. 50-55.

[6] สมศักดิ์ มงคล. "**ชื่อวิทยานิพนธ์**". (วิทยานิพนธ์ปริญญามหาบัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒประสานมิตร, 2543).

[7] สมศรี บุญมาก. "ชื่อบทความ". ชื่อการประชุมวิชาการ. 2549. หน้า 45-48.

[8] R. Good. (2011, Feb 10). Computers (2nd ed.) [Online]. Available: http://www.computers.com

[9] J. Better, "How to Write," Ph.D. dissertation, Dept. Elect. Eng., Amazing University, Cambridge, MA, 2003.