

การเรียนรู้เชิงลึกสำหรับเข้ารหัสบัญชีการจำแนกโรคระหว่างประเทศ (ไอซีดี) จากบันทึกเวชระเบียน
โดยใช้คลังข้อมูลภาษาไทยและภาษาอังกฤษ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2566

Deep Learning for Coding International Classification of Diseases (ICD) from Medical
Records Using Thai and English Corpus



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty Of Engineering
Chulalongkorn University
Academic Year 2023

หัวข้อวิทยานิพนธ์	การเรียนรู้เชิงลึกสำหรับเข้ารหัสสัญญาณการจำแนกโรค ระหว่างประเทศ (ไอซีดี) จากบันทึกเวชระเบียนโดยใช้ คลังข้อมูลภาษาไทยและภาษาอังกฤษ
โดย	น.ส.ณัฐชา ชาชูม
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เกริก ภิรมย์โสภา
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ นายแพทย์กฤษณ์ เจริญลาภ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัย เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ หนูไพโรจน์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.เกริก ภิรมย์โสภา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ นายแพทย์กฤษณ์ เจริญลาภ)

..... กรรมการ
(รองศาสตราจารย์ ดร.พีรพล เวทีกุล)

..... กรรมการภายนอกมหาวิทยาลัย
(นายแพทย์วิหวัศ เจนบุญไทย)

ณัฐชา ซาซุม : การเรียนรู้เชิงลึกสำหรับเข้ารหัสบัญชีการจำแนกโรคระหว่างประเทศ (ไอซีดี) จากบันทึกเวชระเบียนโดยใช้คลังข้อมูลภาษาไทยและภาษาอังกฤษ. (Deep Learning for Coding International Classification of Diseases (ICD) from Medical Records Using Thai and English Corpus) อ.ที่ปรึกษาหลัก : รศ. ดร.เกริก ภิรมย์โสภิต, อ.ที่ปรึกษาร่วม : ผศ. นพ.กฤษณ์ เจริญลาภ

งานวิจัยนี้เสนอแบบจำลองสำหรับจำแนกรหัสไอซีดีแบบหลายผลากเพื่อเป็นตัวช่วยจำแนกรหัสไอซีดี การให้รหัสไอซีดีไม่ครบส่งผลให้โรงพยาบาลไม่ได้รับเงินค่าชดเชยอย่างเหมาะสม เราจึงมุ่งเน้นที่จะช่วยโรงพยาบาลให้รหัสไอซีดีอย่างครบถ้วนในขั้นตอนการเบิกจ่ายเงินค่ารักษาพยาบาลซึ่งผลลัพธ์ที่ได้คือการสนับสนุนทางการเงินแก่โรงพยาบาล โดยปกติแล้วปัญหาการให้รหัสไอซีดีแบบหลายผลากเป็นชุดข้อมูลบันทึกเวชระเบียนที่มีการกระจายตัวแบบหางยาวซึ่งไม่ควรละทิ้งข้อมูลใด ๆ ดังนั้น การจำแนกแบบหลายผลากของเราจึงเป็นการรวมกันของวิธีการรวมแบบจำลองการเรียนรู้เชิงลึกทั้งสาม (Bidirectional Long-short term memory, Convolutional neural network and Transformer encoder) ด้วยการเลือกค่าทำนายสูงสุดและแบบจำลองทางสถิติคือ Multinomial Naïve Bayes ร่วมกับวิธี Binary Relevance ในส่วนของแบบจำลองการเรียนรู้เชิงลึกจะรับผิดชอบกลุ่มรหัสไอซีดีทั่วไป (โรคที่พบบ่อย) ในขณะที่แบบจำลองทางสถิติจะจัดการกับกลุ่มรหัสไอซีดีพบบยาก (โรคที่ผู้ป่วยไม่ค่อยเป็น) ผลการทำนายรหัสไอซีดี-10 จากบันทึกการรักษาด้วยแบบจำลองนี้ให้ค่า Jaccard index ที่ 0.792 สำหรับกลุ่มรหัสไอซีดีทั่วไปและที่ 0.205 สำหรับกลุ่มรหัสไอซีดีพบบยาก ส่วนของผลการทำนายรหัสไอซีดี-9 จากบันทึกการให้ยา (หัตถการรักษา) ให้ค่า Jaccard index ที่ 0.963 สำหรับกลุ่มรหัสไอซีดีทั่วไปและที่ 0.201 สำหรับกลุ่มรหัสไอซีดีพบบยาก ซึ่งแบบจำลองนี้มีความสามารถในการแนะนำรหัสไอซีดีที่ขั้นตอนการเบิกจ่ายเงินค่ารักษาพยาบาล

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์

ปีการศึกษา 2566

ลายมือชื่อนิสิต

ลายมือชื่อ อ.ที่ปรึกษาหลัก

ลายมือชื่อ อ.ที่ปรึกษาร่วม

6270077221 : MAJOR COMPUTER SCIENCE

KEYWORD: ICD codes, multi-label classification, Thai language, Deep learning,
medical record

Nutcha Zazum : Deep Learning for Coding International Classification of
Diseases (ICD) from Medical Records Using Thai and English Corpus.
Advisor: Assoc. Prof. KRERK PIROMSOPA, Ph.D. Co-advisor: Asst. Prof. CHRIS
CHAROENLAP, Ph.D.

We propose an ensemble model for multi-label ICD classification to assist in ICD coding. An incomplete ICD code prevents hospitals from receiving a full compensation. We aim at helping hospitals to complete the ICD code in the reimbursement process. The eventual result is to financially support the hospital. Naturally, the multi-label ICD is a long-tailed distribution dataset of medical records. No data should be dropped out. Therefore, our multi-label classification is a combination of a maximizing ensemble model which has 3 deep learning models (bidirectional Long-short term memory, convolutional neural network and 4 multi-head attention Transformers) and a binary relevance with multinomial Naïve Bayes (statistical) model. The deep learning models are responsible for frequent ICD codes while the statistical model handles infrequent ones. Our model can predict ICD-10-TM from course notes up to 0.792 of Jaccard index for frequent ICD group and 0.205 for infrequent group. The prediction of ICD-9-CM from medication note with procedures achieves 0.963 of Jaccard index for frequent ICD group and 0.201 for infrequent group. The model is capable of suggesting ICD code in the reimbursement process.

Field of Study: Computer Science

Academic Year: 2023

Student's Signature

Advisor's Signature

Co-advisor's Signature

กิตติกรรมประกาศ

ขอมอบความสำเร็จนี้ให้กับคุณพ่อที่รักยิ่ง

ขอกล่าวขอบคุณ รศ. ดร. เกริก ภิรมย์โสภา อาจารย์ที่ปรึกษาหลักที่คอยให้คำแนะนำแนวทางการวิจัย แก่ไขข้อบกพร่อง ช่วยเหลือและให้คำปรึกษาทั้งด้านการเรียนและการใช้ชีวิตจนการวิจัยนี้ประสบความสำเร็จในที่สุด ผศ. นพ. กฤษณ์ เจริญลาภ อาจารย์ที่ปรึกษาร่วมที่คอยช่วยเหลือในการขอข้อมูลและให้คำแนะนำการทำวิจัยนี้ อาจารย์วิหวัธ เจนบุญไทย ที่ให้ความรู้เกี่ยวกับรหัสไอซีดีและพาไปศึกษาขั้นตอนการเบิกจ่ายเงินของโรงพยาบาล

ขอขอบคุณโรงพยาบาลจุฬาลงกรณ์ที่มอบข้อมูลเพื่อการวิจัย ศูนย์วิจัยและพัฒนาเทคโนโลยีเพื่ออุตสาหกรรมแห่งจุฬาลงกรณ์มหาวิทยาลัยที่อำนวยความสะดวกเกี่ยวกับทรัพยากรคอมพิวเตอร์

ขอขอบคุณพ่อกับแม่ที่มอบโอกาสให้ได้ต่อเรียนปริญญาโท น้ำดื่มที่คอยเดินจับมืออยู่เคียงข้างในทุกช่วงเวลาทั้งสุขและเศร้า แม่คอยที่คอยสนับสนุน ปกป้องและให้กำลังใจ น้ำทัยที่คอยให้กำลังใจและคำแนะนำ

ขอบคุณญาติพี่น้อง จิงจรัสทรัพย์ ที่คอยช่วยเหลือ รับฟังปัญหาและให้พลังบวกเสมอมา วิชยะ ศิริภพพร ที่คอยช่วยเหลือแก้ปัญหาต่าง ๆ เจ๊กี้กที่คอยเทศนาเรื่องสนุกคลายเครียด โบทที่รับฟังความคิดที่แสนรุ่มรวย ขอขอบคุณตัวข้าพเจ้าที่ไม่ล้มเลิกความตั้งใจและก้าวเดินไปด้วยพลังจากทุกท่านที่กล่าวมา

สุดท้ายนี้ข้าพเจ้าหวังเป็นอย่างยิ่งว่างานวิจัยนี้จะสร้างประโยชน์ให้กับผู้ที่สนใจนำไปศึกษาค้นคว้าต่อไป

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ณัฐชา ชาชุม

สารบัญ

	หน้า
.....	ค
บทคัดย่อภาษาไทย.....	ค
.....	ง
บทคัดย่อภาษาอังกฤษ.....	ง
กิตติกรรมประกาศ.....	จ
สารบัญ.....	ฉ
สารบัญตาราง.....	ช
สารบัญรูปภาพ.....	ญ
1. บทนำ.....	1
1.1. สาเหตุและปัจจัยที่ส่งผลต่ออัตราการเบิกจ่าย.....	1
1.2. บัญชีจำแนกทางสถิติระหว่างประเทศของโรคและปัญหาสุขภาพที่เกี่ยวข้องหรือรหัสไอซีดี (International Classification of Diseases: ICD).....	3
1.3. กระบวนการเบิกจ่ายค่ารักษาพยาบาล (Reimbursement process).....	5
1.4. ระบบกลุ่มวินิจฉัยโรคร่วม (Diagnostic Related Groups system: DRGs).....	7
2. แนวคิดและทฤษฎี.....	9
2.1. ประเภทของการจำแนก (Classification).....	9
2.2. การวัดประเมิณผล (Metrics).....	10
2.3. การตัดคำในภาษาไทย (Tokenization in Thai).....	13
3. บททบทวนวรรณกรรม.....	17
3.1. การจำแนกข้อความ (Text Classification).....	17
3.2. การจำแนกข้อความจากบันทึกเวชระเบียนหลายภาษา (Medical record classification with multilingual).....	18

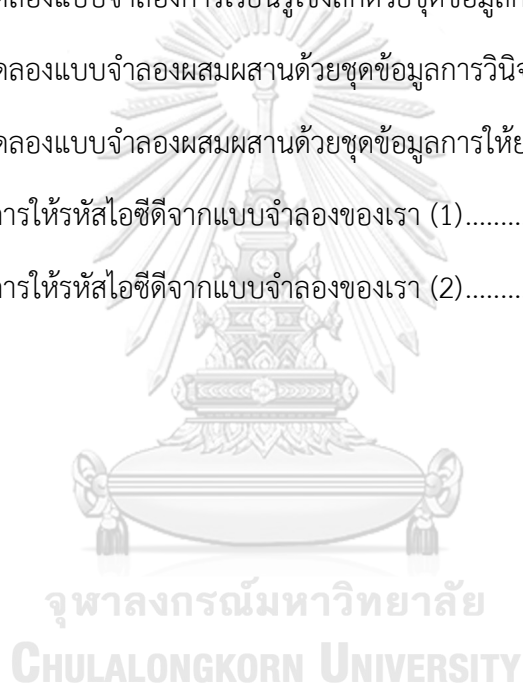
3.3. การจำแนกรหัสไอซีดี (ICD classification)	18
4. ระเบียบวิธี	21
4.1. ชุดข้อมูล (Dataset)	21
4.1.1. ชุดข้อมูลความรู้จำเพาะ (Domain Knowledge dataset).....	21
4.1.2. ชุดข้อมูลอักษรย่อทางการแพทย์ (Medical acronym).....	22
4.1.3. ชุดข้อมูลเวชระเบียน (Medical record dataset)	22
4.2. กระบวนการทดลอง (Methodology).....	27
4.2.1. การเตรียมข้อมูล (Data preparation)	28
4.2.2. กระบวนการประมวลผลล่วงหน้า (Preprocessing).....	30
4.2.3. การฝังคำ (Word Embedding).....	32
4.2.4. การสร้างแบบจำลอง (Modeling).....	34
4.3. การปรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter tuning).....	43
4.4. การวัดผลแบบจำลอง (Evaluation).....	45
5. ผลการทดลอง.....	47
5.1. ผลการทดลองแบบจำลองเชิงสถิติ (Experiment 1)	47
5.2. ผลการทดลองแบบจำลองเชิงลึก (Experiment 2).....	48
5.3. ผลการทดลองแบบจำลองผสมผสาน (Experiment 3).....	50
6. อภิปรายและสรุปผล	52
6.1. การอภิปรายผล (Discussion)	52
6.2. การสรุปผล (Conclusion)	59
บรรณานุกรม.....	60
ประวัติผู้เขียน	65

สารบัญตาราง

หน้า

ตารางที่ 1 ตัวอย่างเปรียบเทียบค่าสัมพัทธ์ระหว่างการให้รหัสไอซีดีครบถ้วนและไม่ครบถ้วน	2
ตารางที่ 2 ตัวอย่างการให้รหัสไอซีดีจากบันทึกเวชระเบียน	6
ตารางที่ 3 ตัวอย่างข้อมูลการจำแนกแบบไบนารี.....	9
ตารางที่ 4 ตัวอย่างข้อมูลการจำแนกแบบหลายคลาส	10
ตารางที่ 5 ตัวอย่างข้อมูลการจำแนกแบบหลายฉลาก	10
ตารางที่ 6 confusion matrix.....	11
ตารางที่ 7 สรุปความแตกต่างของตัวตัดคำภาษาไทย	15
ตารางที่ 8 ตารางเปรียบเทียบของแบบจำลองที่เกี่ยวข้องระหว่างปัญหาที่แตกต่างกัน	19
ตารางที่ 9 แหล่งที่มาของข้อมูลไอซีดี.....	21
ตารางที่ 10 ตัวอย่างอักขระย่อและคำเต็ม	22
ตารางที่ 11 ตารางแสดงรหัสที่มีจำนวนตัวอย่างมากที่สุด 5 อันดับแรกจากรูปที่ 4	23
ตารางที่ 12 จำนวนรหัสไอซีดี-10 ในแต่ละช่วงจำนวนตัวอย่าง	24
ตารางที่ 13 ตัวอย่างบันทึกการให้ยาและรหัสไอซีดีที่เกี่ยวข้อง.....	25
ตารางที่ 14 ตารางแสดงรหัสที่มีจำนวนตัวอย่างมากที่สุด 5 อันดับแรก	25
ตารางที่ 15 จำนวนรหัสไอซีดี-9 ในแต่ละช่วงจำนวนตัวอย่าง	27
ตารางที่ 16 ข้อมูลทางสถิติเชิงปริมาณของชุดบันทึกการรักษา.....	29
ตารางที่ 17 ข้อมูลทางสถิติเชิงปริมาณของชุดบันทึกการให้ยา	30
ตารางที่ 18 ตัวอย่างข้อความก่อนและหลังผ่านการประมวลผลตามขั้นตอน	31
ตารางที่ 19 ผลการทดลองเบื้องต้นเปรียบเทียบประสิทธิภาพแบบจำลองที่ใช้การฝังคำที่แตกต่างกัน	32
ตารางที่ 20 เปรียบเทียบประสิทธิภาพของแบบจำลองที่ความยาวของข้อความต่างกัน	36

ตารางที่ 21 ชุดค่าไฮเปอร์พารามิเตอร์ของ CNN หลังการปรับ.....	44
ตารางที่ 22 ชุดค่าไฮเปอร์พารามิเตอร์ของ Bi-LSTM หลังการปรับ.....	44
ตารางที่ 23 ชุดค่าไฮเปอร์พารามิเตอร์ของ Transformer encoder หลังการปรับ.....	44
ตารางที่ 24 ผลการทดลองแบบจำลองทางสถิติด้วยชุดข้อมูลการวินิจฉัยโรค	47
ตารางที่ 25 ผลการทดลองแบบจำลองทางสถิติด้วยชุดข้อมูลการให้ยาและหัตถการ	48
ตารางที่ 26 ผลการทดลองแบบจำลองการเรียนรู้เชิงลึกด้วยชุดข้อมูลการวินิจฉัยโรค.....	49
ตารางที่ 27 ผลการทดลองแบบจำลองการเรียนรู้เชิงลึกด้วยชุดข้อมูลการให้ยาและหัตถการ	50
ตารางที่ 28 ผลการทดลองแบบจำลองผสมผสานด้วยชุดข้อมูลการวินิจฉัยโรค.....	51
ตารางที่ 29 ผลการทดลองแบบจำลองผสมผสานด้วยชุดข้อมูลการให้ยาและหัตถการ	51
ตารางที่ 30 ตัวอย่างการให้รหัสไอซีดีจากแบบจำลองของเรา (1).....	53
ตารางที่ 31 ตัวอย่างการให้รหัสไอซีดีจากแบบจำลองของเรา (2).....	54



สารบัญรูปภาพ

	หน้า
รูปที่ 1 (a) โครงสร้างของรหัส ICD-9-CM และ (b) โครงสร้างรหัส ICD-10-TM.....	5
รูปที่ 2 กระบวนการเบิกจ่ายค่ารักษาพยาบาล	5
รูปที่ 3 การกระจายตัวของจำนวนตัวอย่างต่อรหัสไอซีดี-10	23
รูปที่ 4 แสดงภาพรวมของจำนวนตัวอย่างในกลุ่มรหัสไอซีดี-10 ส่วนที่พบบ่อย	23
รูปที่ 5 การกระจายตัวของจำนวนตัวอย่างต่อรหัสไอซีดี-9	26
รูปที่ 6 แสดงภาพรวมของจำนวนตัวอย่างในกลุ่มรหัสไอซีดี-9 ส่วนที่พบบ่อย	26
รูปที่ 7 ภาพรวมขั้นตอนและโครงสร้างแบบจำลอง.....	27
รูปที่ 8 ตัวอย่างคำใกล้เคียงของคำว่า ‘cholera’	33
รูปที่ 9 ตัวอย่างเวกเตอร์ของคำจากการผ่าน TextVectorization	34
รูปที่ 10 การทดลองทั้งหมดในงานวิจัย	35
รูปที่ 11 แผนภาพการแบ่งชุดข้อมูลสำหรับแต่ละการทดลอง	35
รูปที่ 12 แสดงของเขตของช่วงค่าคำตอบจากการคำนวณด้วยฟังก์ชัน sigmoid (อ้างอิงรูปจาก https://guopai.github.io/ml-blog16.html)	37
รูปที่ 13 แสดงภาพโครงสร้างของแบบจำลอง Multinomial Naïve Bayes	38
รูปที่ 14 โครงสร้างของแบบจำลอง Bidirectional LSTM.....	39
รูปที่ 15 โครงสร้างของแบบจำลอง Convolutional neural network	40
รูปที่ 16 โครงสร้างของแบบจำลอง Transformer encoder.....	42
รูปที่ 17 โครงสร้างของการรวมแบบจำลองด้วยวิธีการ Ensemble.....	43
รูปที่ 18 ตัวอย่างการคำนวณของค่า Precision@k	45
รูปที่ 19 ภาพประกอบการอธิบายการซ้อนทับของคำตอบ	46

1. บทนำ

กระบวนการเบิกจ่ายค่ารักษาพยาบาลในหลายประเทศกำหนดให้โรงพยาบาลส่งรายงานระบุรหัสไอซีดี (International Classification of Diseases: ICD) การระบุรหัสไอซีดีโรคหลักและโรคร่วมแตกต่างกันจะได้รับอัตราค่าเบิกจ่ายค่ารักษาแตกต่างกันไป สาเหตุและปัจจัยร่วมที่ส่งผลให้การอัตราการเบิกจ่ายที่โรงพยาบาลพึงได้รับไม่เหมาะสมนั้น มีดังต่อไปนี้

1.1. สาเหตุและปัจจัยที่ส่งผลต่ออัตราการเบิกจ่าย

1.1.1. การส่งเบิกจ่ายล่าช้า การส่งเอกสารเบิกจ่ายต้องเสร็จสิ้นภายในระยะเวลาที่กำหนด สำหรับการส่งเบิกจ่ายล่าช้า เงินชดเชยที่ได้รับจะต้องถูกหักเงินค่าปรับ

1.1.2. ขาดแคลนบุคลากรผู้มีความรู้เกี่ยวกับรหัสไอซีดีหรือมีผู้เชี่ยวชาญการให้รหัสไอซีดีจำนวนน้อย ซึ่งส่งผลต่อความรวดเร็วในการระบุและตรวจสอบความถูกต้องของรหัสไอซีดีก่อนส่งเอกสารไปเบิกจ่าย

1.1.3. ความเปลี่ยนแปลงของรหัสไอซีดี เนื่องจากรหัสไอซีดีมีการปรับปรุงเพิ่มเติมเป็นระยะเพื่อตอบสนองความเปลี่ยนแปลงของโรคที่เกิดขึ้นใหม่ การติดตามการเปลี่ยนแปลงนี้เป็นงานที่ยาก

1.1.4. ความไม่ครบถ้วนของรหัส ปัญหานี้เป็นผลจากการบันทึกรายละเอียดไม่ครบถ้วนและการระบุรหัสไอซีดีไม่ถูกต้อง ซึ่งปัญหานี้คิดเป็น 10-40 % ของปัญหาทั้งหมด [1] ดังนั้นเราจึงตั้งใจแก้ไขปัญหานี้

ตัวอย่างแสดงความแตกต่างของค่าอัตราสัมพัทธ์ระหว่างรายงานที่ให้รหัสไอซีดีครบและไม่ครบตามตารางที่ 1 เป็นการเปรียบเทียบค่า RW ระหว่าง 2 กรณีที่ผู้ป่วยจัดอยู่ในกลุ่มโรคร่วมเดียวกัน ซึ่งผลที่ได้ให้ค่า RW แตกต่างกันอยู่ 5.334 อ้างอิงตามอัตราการจ่ายเงินชดเชยคิดจากผลคูณของค่า RW กับอัตราจ่ายมาตรฐานของกลุ่มผู้ป่วยตามระบบกลุ่มวินิจฉัยโรคร่วม (DRG) ส่งผลให้โรงพยาบาลสูญเสียเงินค่าชดเชยที่สมควรได้รับในกรณีให้รหัสไอซีดีไม่ครบถึง 5 เท่าของอัตราจ่ายค่าชดเชยตามจริง ดังนั้นงานวิจัยนี้จึงมุ่งเน้นลดปัญหาการให้รหัสไอซีดีไม่ครบถ้วนโดยสร้างแบบจำลองที่จะมาเป็นตัวช่วยแนะนำรหัสไอซีดีให้กับผู้ให้รหัส (Expert coder) ลดระยะเวลาการให้รหัสและทดแทนช่องว่างทางประสบการณ์ของแพทย์และผู้ให้รหัสใหม่ สำหรับรายละเอียดเพิ่มเติมการคำนวณค่าสัมพัทธ์และระบบกลุ่มวินิจฉัยโรคร่วมจะอธิบายในตอนถัดไป

ตารางที่ 1 ตัวอย่างเปรียบเทียบค่าสัมพัทธ์ระหว่างการให้รหัสไอซีดีครบถ้วนและไม่ครบถ้วน

ตัวอย่าง	<p>ผู้ป่วยชายอายุ 45 ปี มีประวัติเป็นโรคไขมันสูงและทาลาสซีเมีย เข้ามาด้วยอาการแผลผ่าตัดมีหนองไหลมา 3 วัน ไม่มีไข้ เดินได้ปกติ case spondylolisthesis L3-5, scoliosis ทำการผ่าตัด TPS L3-5, Laminectomy, PLIF L4-5, PLF</p> <p>รักษาโดยผ่าตัด Re-explored with Debridement ส่งตรวจ MRI 2 ครั้ง ระหว่างการรักษามีการแพ้ยา (Fosfomycin): ซีดจนต้องให้เลือด 4 ยูนิท ทำแผล ทำกายภาพบำบัด รวมทั้งสิ้นอยู่โรงพยาบาล 27 วัน</p>	
	ให้รหัสไอซีดีไม่ครบ	ให้รหัสไอซีดีครบ
โรคหลัก (Principal diagnosis)	<p>T81.4 Infection following a procedure</p>	
โรคร่วม (Comorbidity)	<p>G06.2 Extradural and subdural abscess, unspecified</p> <p>T36.9 Poisoning by adverse effect of and underdosing of unspecified systemic antibiotic.</p> <p>D62 Acute posthemorrhagic anemia</p> <p>E83.5 Disorders of calcium metabolism</p>	
โรคแทรกหรืออื่น ๆ	<p>Y83.8 Surgical operation with formation of external stoma as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure</p> <p>Y83.8 Surgical operation with formation of external stoma as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure</p> <p>Y40.9 Systemic antibiotic, unspecified</p>	

		D56.3 Thalassemia minor E78.9 Disorder of lipoprotein metabolism, unspecified
หัตถการ (Procedure)	86.22 Injury to other specified intrathoracic organs without mention of open wound into cavity 88.93 Magnetic resonance imaging of spinal canal	86.22 Injury to other specified intrathoracic organs without mention of open wound into cavity 88.93 Magnetic resonance imaging of spinal canal 99.04 Transfusion of packed cells 93.39 Other physical therapy therapeutic procedures
DRG	18040	18044
RW	1.6429	6.9869

การสร้างแบบจำลองที่สามารถแนะนำรหัสไอซีดีที่ถูกต้องเพื่อลดจำนวนเงินชดเชยที่โรงพยาบาลสูญเสียไปจากการเบิกจ่ายด้วยรหัสไอซีดีไม่ครบถ้วน มีความจำเป็นต้องทำความเข้าใจเกี่ยวกับรหัสไอซีดี ความสัมพันธ์ของรหัสไอซีดีต่อกระบวนการเบิกจ่ายและระบบที่เกี่ยวข้องกับการคำนวณอัตราการเบิกจ่าย ซึ่งจะอธิบายต่อไปนี้

1.2. บัญชีจำแนกทางสถิติระหว่างประเทศของโรคและปัญหาสุขภาพที่เกี่ยวข้องหรือรหัสไอซีดี (International Classification of Diseases: ICD)

บัญชีจำแนกโรคระหว่างประเทศ (International Classification of Diseases) ฉบับแรกถือกำเนิดขึ้นจากกลุ่มประเทศทางยุโรป โดยเริ่มจากการจำแนกโรคที่เป็นสาเหตุการตาย จนกระทั่งมีการก่อตั้งองค์การอนามัยโลก (World Health Organization หรือ WHO) ในช่วงสงครามโลกครั้งที่ 2 องค์การอนามัยโลกได้เห็นความสำคัญของ ICD และเลือกใช้ ICD เป็นเครื่องมือหลักในการจำแนกและให้รหัสโรค ซึ่งรวมไปถึงรหัสโรคที่อาจไม่รุนแรงถึงแก่ชีวิตและสาเหตุการบาดเจ็บต่าง ๆ เข้าไว้ในบัญชีจำแนกโรคระหว่างประเทศ

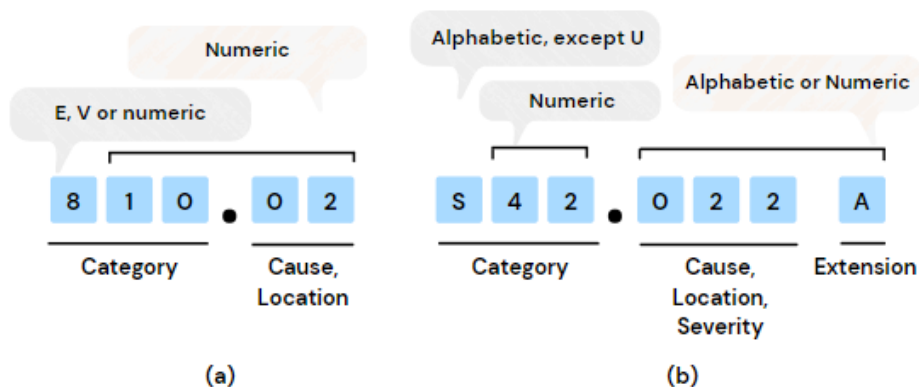
และต่อมาได้เปลี่ยนชื่อเป็น International Statistical Classification of Diseases and Related Health Problems หลังจากนั้นองค์การอนามัยโลกได้มีการปรับปรุงแก้ไขบัญชีจำแนกโรคนี้อย่างต่อเนื่อง จนกระทั่งได้จัดทำบัญชีจำแนกโรคระหว่างประเทศฉบับที่ 10 (ICD-10) เสร็จใน พ.ศ. 2535 และเริ่มใช้ใน พ.ศ. 2537 โดยมีประเทศไทย เดนมาร์ก และ เช็กโกสโลวะเกียเป็นสามประเทศแรกที่เริ่มใช้ ICD-10

โรคและปัญหาการเจ็บป่วยในแต่ละประเทศย่อมมีความแตกต่างกัน หลายประเทศจึงมีแนวคิดในการดัดแปลงรหัส ICD เพื่อให้เหมาะสมในการทำงานยิ่งขึ้น ประเทศอเมริกา ดัดแปลง ICD-9 เป็น ICD-9-CM ประเทศออสเตรเลียดัดแปลง ICD-10 เป็น ICD-10-AM สำหรับประเทศไทย หลังจากเริ่มใช้ ICD-10 มาระยะหนึ่ง พบว่า ICD-10 ไม่ครอบคลุมบางโรคที่พบบ่อยในประเทศไทย จึงเกิดแนวคิดในการจัดทำ ICD-10 ฉบับประเทศไทย หรือ ICD-10-TM (TM ย่อมาจาก Thai modification) ขึ้นใน พ.ศ. 2544 โดยได้รับความเห็นชอบและสนับสนุนจากองค์การอนามัยโลก การจัดทำ ICD-10-TM โดยมีจุดมุ่งหมายในการพัฒนาให้รหัสที่เหมาะสมกับการใช้งานในประเทศไทย และเป็นพื้นฐานในการจัดทำระบบกลุ่มวินิจฉัยโรคร่วม (DRG) ฉบับประเทศไทย

สรุปรหัสไอซีดีคือมาตรฐานการจำแนกการวินิจฉัยโรคซึ่งแปลงโรค อาการ ความผิดปกติ สาเหตุของการบาดเจ็บเพิ่มเติม และหัตถการต่าง ๆ สำหรับการรักษา ให้อยู่ในรูปแบบของรหัสมาตรฐานทางสถิติซึ่งถูกจัดทำขึ้นโดยองค์การอนามัยโลก

ปัจจุบันองค์กรประกันในประเทศไทยกำหนดข้อบังคับและกฎเกณฑ์ในการเข้ารหัสส่งเบิกโดยให้ใช้รหัส ICD-10-TM สำหรับระบุโรคและอาการและใช้รหัส ICD-9-CM สำหรับระบุหัตถการการรักษา โครงสร้างของรหัสไอซีดีทั้งสองมีส่วนที่แตกต่างกันดังแสดงในรูปที่ 1 สำหรับรหัส ICD-9-CM โดยทั่วไปประกอบขึ้นจากตัวเลขทั้งหมดไม่เกิน 5 หลักตามรูปที่ 1(a) แบ่งด้วยเครื่องหมายจุดออกเป็น 3 หลักแรกจากทางซ้ายคือหมวดหมู่และ 2 หลักหลังเครื่องหมายจุดเป็นส่วนขยายใช้บ่งบอกตำแหน่งและสาเหตุ ยกเว้นรหัสหมวดอาการบาดเจ็บภายนอกจะขึ้นต้นหลักแรกทางซ้ายสุดด้วยตัวอักษร E และ V ในส่วนของโครงสร้าง ICD-10-TM จะมีรายละเอียดและจำนวนหลักของรหัสมากกว่า รหัสไอซีดีนี้ประกอบขึ้นด้วยอย่างน้อย 7 หลักตามรูปที่ 1(b) แบ่งกลุ่มด้วยเครื่องหมายจุดและวรรค ส่วนของ 3 หลักแรกจากทางซ้ายคือหมวดหมู่ประกอบด้วยตัวอักษรภาษาอังกฤษตัวแรกและตัวเลขใน 2 หลักที่เหลือ ต่อมาที่ส่วนของ 3 หลักหลังเครื่องหมายจุด (หลักที่ 4-6) ประกอบด้วยตัวเลข

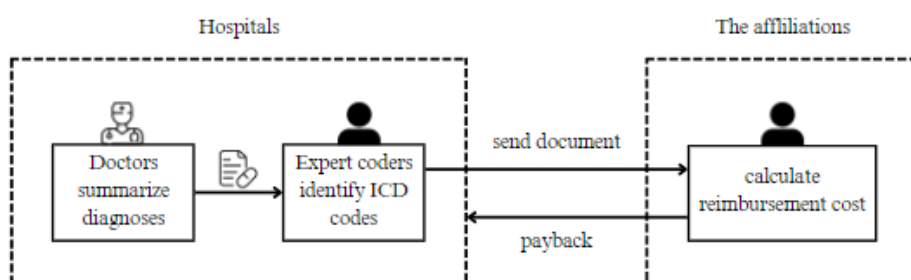
เท่านั้น ใช้บ่งบอกสาเหตุ ตำแหน่งและความรุนแรง หลักสุดท้ายถูกแบ่งด้วยวรรคอาจแสดงด้วยตัวอักษรภาษาอังกฤษหรือตัวเลขเพียงตัวเดียวใช้ระบุส่วนขยายเพิ่มเติม



รูปที่ 1 (a) โครงสร้างของรหัส ICD-9-CM และ (b) โครงสร้างรหัส ICD-10-TM

1.3. กระบวนการเบิกจ่ายค่ารักษาพยาบาล (Reimbursement process)

กระบวนการเบิกจ่ายค่ารักษาพยาบาลคือขั้นตอนการยื่นขอรับเงินชดเชยค่ารักษา และการบริการของโรงพยาบาลผู้ให้บริการต่อองค์กรประกัน โดยอัตราค่าชดเชยจะขึ้นกับผลการจัดกลุ่มผู้ป่วยด้วยระบบกลุ่มวินิจฉัยโรคร่วมตามรายงานการรักษาที่สรุปการวินิจฉัยโรคและเหตุการณ์ด้วยรหัสไอซีดี ขั้นตอนการปฏิบัติงานแสดงในรูปที่ 2



รูปที่ 2 กระบวนการเบิกจ่ายค่ารักษาพยาบาล

กระบวนการเบิกจ่ายแบ่งออกเป็น 2 ฝั่งคือ ฝั่งของโรงพยาบาลที่ให้บริการ และฝั่งขององค์กรประกัน สำหรับฝั่งโรงพยาบาลผู้ให้บริการจะเริ่มจากแพทย์เขียนใบสรุปการวินิจฉัย (Discharge summaries) ที่มีรายละเอียดของโรคและเหตุการณ์ที่เกี่ยวข้องทั้งหมด หลังจากจำหน่ายผู้ป่วยออกจากโรงพยาบาล (Hospital discharge) และส่งให้กับผู้ให้รหัส

ไอซีดีหรือแพทย์ผู้มีความรู้เกี่ยวกับรหัสไอซีดีตรวจสอบความครบถ้วน หากข้อมูลไม่ครบถ้วน เอกสารจะถูกส่งกลับให้แพทย์ผู้รับผิดชอบแก้ไขอีกครั้ง เนื่องจากผู้ให้รหัสไม่สามารถการเปลี่ยนแปลงบันทึกการรักษาและเนื้อหาภายในใบสรุปการวินิจฉัยโรคได้ หากไม่แก้ไขบันทึกจะไม่สามารถให้รหัสได้อย่างครบถ้วน ส่งผลให้โรงพยาบาลผู้ให้บริการสูญเสียเงินชดเชยตามสมควร (บทนำ-ปัญหาที่ 4) ดังนั้นผู้ให้รหัสจะระบุไอซีดีให้ตรงกับโรคและหัตถการตามเอกสารเมื่อข้อมูลครบถ้วน เหตุการณ์ที่ต้องส่งเอกสารให้แพทย์ทำการบันทึกเพิ่มเติมคำวินิจฉัยใหม่นั้นส่งผลให้เกิดความล่าช้าในกระบวนการและการส่งเบิกถูกหักเงินตามอัตราค่าปรับที่กำหนด และหากไม่แก้ไขใบสรุปการวินิจฉัยโรคผู้ให้รหัสไอซีดีพบว่าควรเพิ่มรหัสไอซีดีจะไม่สามารถให้รหัสเพิ่ม ตารางที่ 2 แสดงตัวอย่างของความแตกต่างระหว่างกระบวนการสรุปการวินิจฉัยของแพทย์และการให้รหัสไอซีดีตามการบันทึกของผู้เชี่ยวชาญการให้รหัส ทางฝั่งองค์กรประกันรับเอกสารที่ให้รหัสไอซีดีแล้วมาจัดกลุ่มผู้ป่วยด้วยระบบจัดกลุ่มโรคร่วม (Diagnostic Related Groups system: DRG) ซึ่งเป็นการจัดกลุ่มโรคที่มีต้นทุกการรักษาใกล้เคียงกันให้อยู่กลุ่มเดียวกันเพื่อประเมินอัตราจ่ายเงินค่าชดเชยให้กับโรงพยาบาลผู้ให้บริการ

ตารางที่ 2 ตัวอย่างการให้รหัสไอซีดีจากบันทึกเวชระเบียน

ตัวอย่าง	ผู้ป่วยอายุ 30 ปี ได้รับการวินิจฉัยว่าเป็น ACL insufficiency เข่าขวา แพทย์เลือกผ่าตัด Arthroscopic ACL reconstruction with semitendinosus graft ระหว่างกรดผ่าตัดพบว่าหมอนรองกระดูกหัวเข่าฉีกขาดด้วย จึงตัดหมอนรองกระดูกออกไป	
	แพทย์สรุปในใบสรุปการวินิจฉัย	ผู้เข้ารหัสให้รหัสไอซีดี
โรคหลัก (Principal diagnosis)	ACL insufficiency right knee	M23.51 Chronic instability of knee, anterior cruciate ligament
โรคร่วม (Comorbidity)	Tear of meniscus right knee	M23.21 Derangement of meniscus due to old tear of injury
หัตถการ (Procedure)	1. Arthroscopic ACL reconstruction with semitendinosus graft 2. Excision of meniscus	81.45 Other repair of cruciate ligament (833-48-01) 80.26 Arthroscopy (833-99-02) 80.6 Excision of semilunar cartilage of knee (833-26-03) 83.43 Excision of muscle or fascia for graft (903-26-01)

1.4. ระบบกลุ่มวินิจฉัยโรคร่วม (Diagnostic Related Groups system: DRGs)

ระบบกลุ่มวินิจฉัยโรคร่วม (DRGs) คือระบบการจัดกลุ่มผู้ป่วยเพื่อทำนายต้นทุน ทรัพยากรที่ใช้ในการรักษาและการให้บริการ โดยมีแนวคิดที่ว่าผู้ป่วยกลุ่มเดียวกันจะใช้ ทรัพยากรในการรักษาและใช้เวลาพักรักษาตัวในโรงพยาบาลใกล้เคียงกัน [2] หลักการของ DRGs คือการจ่ายเงินล่วงหน้าให้แก่โรงพยาบาลผู้ให้บริการแบบเฉลี่ยความเสี่ยงระหว่าง องค์การประกันกับโรงพยาบาลผู้ให้บริการ โดยการจัดกลุ่มผู้ป่วย DRG จะใช้ข้อมูลดังต่อไปนี้

- อายุปี (Age) 0 – 124 ปี
- อายุวัน (Age Day) 0 – 364/365 วัน
- เพศ (Gender) 1 = ชาย, 2=หญิง
- น้ำหนัก (Admission weight) หน่วยกิโลกรัม (kg)
- จำนวนวันนอน (Length of stay: LOS) หน่วยเป็นวันและชั่วโมง (day, hour)
- ประเภทการจำหน่าย (Discharge type) 1 - 9

ร่วมกับข้อมูลในใบสรุปการวินิจฉัยโรค ซึ่งประกอบไปด้วย โรคหลัก (Principal diagnosis) โรคร่วม (Comorbidity) โรคแทรก (Complication) และหัตถการ (Procedure) ซึ่งต้อง ระบุด้วยรหัสไอซีดี 10 (ICD-10-TM) และ รหัสไอซีดี 9 (ICD-9-CM) ตามลำดับอย่างถูกต้อง

หลังจากจัดกลุ่มผู้ป่วย DRG เราสามารถคำนวณอัตราการจ่ายเงินของแต่ละกลุ่ม ผู้ป่วยจากค่าน้ำหนักสัมพัทธ์ (Relative weight: RW) ดังนี้

$$\text{ค่าน้ำหนักสัมพัทธ์} = \frac{\text{ค่าเฉลี่ยในการใช้ทรัพยากรในการรักษาผู้ป่วย DRG นั้น}}{\text{ต้นทุนเฉลี่ยของการรักษาผู้ป่วยทั้งหมด}}$$

ค่า RW ใช้เป็นค่ามาตรฐานสำหรับการประเมินต้นทุนการใช้ทรัพยากรในการรักษาและ สามารถสะท้อนความรุนแรงของผู้ป่วยได้ ในกรณีที่ผู้ป่วยมีจำนวนวันนอนสูงหรือต่ำกว่า เกณฑ์ อัตราการจ่ายเงินจะคำนวณเพิ่มเติมตามจำนวนวันนอนแล้วเรียกว่าค่า adjust RW สุดท้ายจำนวนเงินที่ทางองค์กรประกันต้องจ่ายจะเท่ากับผลคูณของค่า RW หรือ adjust RW กับอัตราค่ารักษาพื้นฐานของผู้ป่วยแต่ละกลุ่ม DRG

$$\text{จำนวนเงินเบิกจ่าย} = \text{ค่า RW} * \text{อัตราค่ารักษาพื้นฐาน (base rate)}$$

เพื่อให้ได้แบบจำลองที่เหมาะสม เราจำเป็นต้องทำความเข้าใจรูปแบบการจำแนกรหัสไอซีดีที่เหมาะสมกับข้อมูลของเรา กระบวนการที่เกี่ยวข้องตลอดการสร้างแบบจำลอง รวมไปถึงประเภทการวัดผลการเรียนรู้ของแบบจำลอง ซึ่งเราจะอธิบายตามลำดับในบทที่ 2 แนวคิดและทฤษฎี



2. แนวคิดและทฤษฎี

2.1. ประเภทของการจำแนก (Classification)

เนื่องจากแบบจำลองของเราเป็นแบบจำลองการจำแนกประเภทเรียนรู้แบบมีผู้สอน (Supervised learning) คือแบบจำลองสามารถหาคำตอบได้ด้วยตนเองจากการศึกษาชุดข้อมูลตัวอย่าง การจำแนกประเภทเรียนรู้แบบมีผู้สอนสามารถแบ่งประเภทการจำแนกตามรูปแบบการให้คำตอบของแบบจำลอง ซึ่งแบ่งออกเป็น 3 ประเภท

2.1.1. การจำแนกประเภทแบบไบนารี (Binary classification)

การจำแนกประเภทแบบไบนารี คือการจำแนกเพียง 2 คลาส ดังแสดงในตารางที่ 3 การจำแนกนี้จำแนกข้อมูลให้ได้คำตอบเพียง 1 คลาสจาก 2 คลาส ตัวอย่างงานที่ใช้การจำแนกประเภทแบบไบนารี ได้แก่ การจำแนกอีเมลที่ไม่พึงประสงค์ (Spam mail classification: Spam, No-spam) การจำแนกเพศ (gender classification: male, female)

ตารางที่ 3 ตัวอย่างข้อมูลการจำแนกแบบไบนารี

ทรงผม (Hair style)	เพศ (Gender)
ผมยาวและหยิก	หญิง
ผมสั้นสีน้ำตาล	ชาย
ผมตรงยาวสีน้ำตาล	ชาย

2.1.2. การจำแนกประเภทแบบหลายคลาส (Multi-class classification)

การจำแนกประเภทแบบหลายคลาส คือการจำแนกที่มีจำนวนคลาสมากกว่า 2 คลาส ดังแสดงในตารางที่ 4 การจำแนกนี้จำแนกข้อมูลให้ได้คำตอบเพียง 1 คลาส จากจำนวนหลาย ๆ คลาส ตัวอย่างงานที่ใช้การจำแนกประเภทแบบหลายคลาส ได้แก่ การจำแนกหมู่เลือด (A, B, O, AB) การจำแนกคะแนนประเมินภาพยนตร์ (1-5 ดาว)

ตารางที่ 4 ตัวอย่างข้อมูลการจำแนกแบบหลายคลาส

ภาพยนตร์	คะแนนประเมิน (Rate)
Monster Hunter (2020)	3
Ready Player One (2018)	5
John Wick (2017)	4

2.1.3. การจำแนกประเภทแบบหลายฉลาก (Multi-label classification)

การจำแนกแบบหลายฉลาก คือการจำแนกที่มีจำนวนคลาสอย่างน้อย 2 คลาส ดังแสดงในตารางที่ 5 การจำแนกนี้จำแนกข้อมูลให้ได้คำตอบมากกว่า 1 คลาส จากจำนวนทั้งหมดของคลาส ตัวอย่างงานที่ใช้การจำแนกประเภทแบบหลายฉลาก ได้แก่ การจำแนกประเภทของภาพยนตร์ (แอคชั่น ไซไฟ ผจญภัย เป็นต้น) การจำแนกรหัสไอซีดีก็เป็นการจำแนกแบบหลายฉลากเพราะบันทึกเวชระเบียนสามารถระบุรหัสไอซีดีได้มากกว่า 1 รหัสอย่างที่แสดงตัวอย่างในตารางที่ 2

ตารางที่ 5 ตัวอย่างข้อมูลการจำแนกแบบหลายฉลาก

ภาพยนตร์	แอคชั่น	ไซไฟ	ผจญภัย
John Wick (2017)	1	0	0
Men in Black (2019)	1	1	0
Monster Hunter (2020)	1	0	1

2.2. การวัดประเมิณผล (Metrics)

เราใช้ confusion matrix ในการวัดประสิทธิภาพของแบบจำลอง Confusion matrix คือตารางขนาด 2x2 ซึ่งประกอบไปด้วยคอลัมน์ของผลเฉลยและแถวของผลทำนายดังแสดงในตารางที่ 6

ตารางที่ 6 confusion matrix

Confusion matrix		ผลทำนาย (Predicted class)	
		Positive (+)	Negative (-)
ผลเฉลย (Actual class)	Positive (+)	True Positive (TP)	False Negative (FN)
	Negative (-)	False Positive (FP)	True Negative (TN)

จากตารางที่ 6 confusion matrix ประกอบด้วย 4 ค่า ดังต่อไปนี้

- ผลบวกจริง (True positive) คือการทำนายค่าผลบวกถูกต้อง นั่นหมายความว่าผู้ป่วยตรวจพบว่าเป็นโรคได้อย่างถูกต้อง
- ผลบวกเท็จ (False positive) คือการทำนายค่าลบเป็นค่าบวก นั่นหมายความว่าคนปกติตรวจพบว่าเป็นโรคอย่างไม่ถูกต้อง
- ผลลบจริง (True negative) คือการทำนายค่าผลลบถูกต้อง นั่นหมายความว่าคนปกติตรวจพบว่าเป็นโรคอย่างถูกต้อง
- ผลลบเท็จ (False negative) คือการทำนายค่าบวกเป็นค่าลบ นั่นหมายความว่าผู้ป่วยตรวจพบว่าเป็นโรคอย่างไม่ถูกต้อง

ดังนั้นแบบจำลองที่มีประสิทธิภาพควรมีค่าผลบวกจริงและผลลบจริงสูงและมีค่าผลบวกเท็จและผลลบเท็จต่ำ สำหรับแบบจำลองทางการแพทย์ควรมีค่าผลบวกเท็จสูงกว่าผลลบเท็จ เพราะการทำนายว่าเป็นโรคไว้ก่อนถือเป็นการป้องกันและเตรียมมือหากพบว่าป่วยในภายหลัง ค่าต่าง ๆ จาก confusion matrix สามารถคำนวณค่า precision recall accuracy F1-score และ Jaccard index ได้ตามสมการต่อไปนี้

2.2.1. Precision

Precision คือสัดส่วนของค่าทำนายผลบวกที่ถูกต้องต่อค่าทำนายผลบวกทั้งหมดตามสมการ (1) ซึ่งแบบจำลองที่มีประสิทธิภาพควรมีค่านี้สูง อีกนัยหนึ่งคือการที่ค่า precision สูงมาจากค่าผลบวกเท็จต่ำ ซึ่งหมายความว่าแบบจำลองทำนายเป็นโรคที่ไม่เกี่ยวข้องน้อยที่สุด

$$\text{precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (1)$$

2.2.2. Recall

Recall คือสัดส่วนของค่าทำนายผลบวกที่ถูกต้องต่อผลรวมทั้งของค่าทำนายที่มีผลเฉลยเป็นผลบวกตามสมการ (2) ซึ่งแบบจำลองที่มีประสิทธิภาพควรมีค่านี้สูง อีกนัยหนึ่งคือการทำ recall สูงมาจากค่าผลลบเท็จต่ำ นั่นหมายความว่าแบบจำลองทำนายว่าผู้ป่วยตรวจพบว่าเป็นโรคในอัตราที่ต่ำ

$$\text{recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (2)$$

2.2.3. Accuracy

Accuracy คือสัดส่วนของค่าทำนายที่ถูกต้องต่อค่าทำนายทั้งหมดตามสมการ (3)

$$\text{accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{False positive} + \text{True negative} + \text{False negative}} \quad (3)$$

2.2.4. F1-Score

F1-score คือค่าเฉลี่ยถ่วงน้ำหนักของ precision และ recall ตามสมการ (4) ซึ่งเหมาะสมกับการกระจายตัวของคลาสไม่เท่ากันหรือเหตุการณ์ที่ผลบวกเท็จและผลลบเท็จมีค่าใกล้เคียงกัน แบบจำลองที่มีประสิทธิภาพควรมีค่า F1-score ใกล้ 1

$$F1 = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \quad (4)$$

2.2.5. Jaccard index

Jaccard index หรือ Jaccard similarity มาจากจำนวนรหัสไอซีดีที่ทำนายถูกต้อง (True positive)หารด้วยผลรวมของรหัสไอซีดีจากการทำนายและรหัสไอซีดีจากผลเฉลยตามสมการ (5) ค่ายิ่งเข้าใกล้ 1 หมายความว่าการทำงานของเราใกล้เคียงกับผลเฉลย

$$\text{Jaccard index} = \frac{\text{True positive}}{\text{True positive} + \text{False positive} + \text{False negative}} \quad (5)$$

2.3. การตัดคำในภาษาไทย (Tokenization in Thai)

วิธีการตัดคำในภาษาไทยแตกต่างจากภาษาอังกฤษอย่างมาก ในภาษาอังกฤษมีรูปแบบการตัดคำที่ชัดเจน คือ ช่องว่าง (white space) ใช้เว้นวรรคระหว่างคำสองคำ และจุด (full-stop) ใช้แสดงจุดสิ้นสุดประโยค แต่สำหรับภาษาไทยเราอาจมีการเว้นวรรคเพื่อขึ้นประโยคใหม่แต่ไม่เป็นในรูปแบบที่ตายตัว และเขียนคำติดไปยาวตลอดทั้งประโยค ไม่มีสัญลักษณ์ขึ้นระหว่างคำสองคำที่ชัดเจน ดังนั้นการตัดคำจึงมีความซับซ้อนมาก การตัดคำที่ถูกต้องมีความสำคัญต่อการนำคำศัพท์ไปวิเคราะห์ต่อ

ตัวตัดคำ (Tokenizer) สำหรับตัดคำภาษาไทยในปัจจุบันมีหลากหลายแบบให้เลือกใช้ผ่านไลบรารี PyThaiNLP [3] ตัวตัดคำภาษาไทยสามารถแบ่งออกเป็น 2 แบบตามหลักการอ้างอิงเพื่อตัดคำ คือตัวตัดคำอ้างอิงตามพจนานุกรม (Dictionary-based) และตัวตัดคำอ้างอิงตามการเรียนรู้ของเครื่อง (Machine learning-based) ซึ่งสามารถดูสรุปความแตกต่างระหว่างตัวตัดคำทั้ง 2 ประเภทนี้ได้ในตารางที่ 7

สำหรับตัวตัดคำอ้างอิงตามพจนานุกรมมีวิธีการตัดคำ 2 วิธี ได้แก่ วิธีการตัดคำแบบยาวที่สุด (Longest matching) และวิธีตัดคำแบบสอดคล้องมากที่สุด (Maximal matching)

- วิธีการตัดคำแบบยาวที่สุด (Longest matching) เป็นวิธีการค้นหาคำโดยเริ่มจากตัวอักษรซ้ายสุดของประโยคไล่ไปทางขวาทีละตัวจนกว่าจะพบคำในพจนานุกรม และหากพบมากกว่า 1 คำที่ตัวอักษรเริ่มต้นเดียวกันให้เลือกคำที่ยาวที่สุด [4] เราสามารถเลือกใช้ตัวตัดคำนี้โดยเรียกฟังก์ชัน “longest” จาก PyThaiNLP ได้

- วิธีตัดคำแบบสอดคล้องมากที่สุด (Maximal matching) เป็นวิธีการหารูปแบบการตัดคำในประโยคที่เป็นไปได้ทั้งหมด โดยจะเลือกรูปแบบที่ให้จำนวนค่าน้อยที่สุด และในกรณีที่จำนวนค่าเท่ากันจะตัดสินด้วยหลักการตัดคำแบบที่ยาวที่สุด เราสามารถเลือกใช้ตัวตัดคำนี้โดยเรียกฟังก์ชัน “newmm” จาก PyThaiNLP ได้

เนื่องจากการตัดคำแบบอ้างอิงตามพจนานุกรมมีข้อจำกัดที่ไม่รองรับคำที่อยู่นอกเหนือพจนานุกรมที่อ้างอิง (Out-of-Vocabulary) ต่อมาจึงพัฒนาตัวตัดคำให้เรียนรู้การตัดคำจากตัวอย่างฝึกผ่านวิธีการเรียนรู้ของเครื่อง ตัวตัดคำอ้างอิงตามการเรียนรู้ของเครื่อง (Machine Learning-based) ใช้หลักการฝึกตัดคำผ่านชุดข้อมูลตัวอย่างให้เรียนรู้การตัดคำอย่างถูกต้อง เป็นลักษณะการเรียนรู้แบบมีผู้สอน (Supervised learning) ตัวตัดคำรูปแบบนี้ที่นิยมใช้คือ Deepcut และ AttaCut ซึ่งสามารถเรียกใช้ได้จาก PyThaiNLP และทั้งสองตัวตัดคำนี้รองรับการตัดคำทั้งภาษาไทยและภาษาอังกฤษ

- Deepcut เป็นตัวตัดคำที่สร้างจากแบบจำลองที่ผสมระหว่าง 1D-Convolutional neural network เข้ากับ deep neural network ซึ่งแบบจำลองถูกฝึกด้วยชุดข้อมูล BEST จาก NECTEC ที่ประกอบด้วยบทความ ข่าว วรรณกรรม และสารานุกรม [5]
- AttaCut เป็นตัวตัดคำที่สร้างจากแบบจำลอง Convolutional neural network แบบ 3 ชั้น ซึ่งจุดสำคัญอยู่ที่เวกเตอร์ฝังคำ (Word embedding) ที่สร้างจากการรวมกันของเวกเตอร์ฝังตัวอักษร (Character embedding) กับเวกเตอร์ฝังพยางค์ (Syllable embedding) และฝึกด้วยชุดข้อมูล Thai National Corpus [6]

จากการเปรียบเทียบประสิทธิภาพด้านการตัดคำทั้งสองตัวตัดคำ Deepcut และ AttaCut มีผลดีมากและค่อนข้างใกล้เคียงกัน แต่เมื่อเปรียบเทียบด้านความเร็ว AttaCut ใช้เวลาประมวลผลเร็วกว่า Deepcut 10 เท่า จึงนำมาสู่ข้อสรุปในการเลือกใช้ตัวตัดคำ AttaCut ในงานวิจัยนี้ด้วยประสิทธิภาพการตัดคำที่ดีทั้งภาษาไทยและภาษาอังกฤษ ความรวดเร็วในการประมวลผล และรองรับคำศัพท์นอกพจนานุกรม

ตารางที่ 7 สรุปความแตกต่างของตัวตัดคำภาษาไทย

	ตัวตัดคำ	วิธีการ	ตัวอย่าง
อิงตามพจนานุกรม (Dictionary-based)	Newmm	วิธีตัดคำแบบสอดคล้องมากที่สุด (Maximal matching)	ประเทศไทย รวม เลือดเนื้อ ชาติ เชื้อ ไทย เป็น พระราชรัฐไพบ ของ ไทย ทุก ส่วน อยู่ ดำรง คง ไว้ ได้ ทั้งหมด ด้วย ไทย ล้วนหมาย รัก สามัคคี ไทย นี้ รัก สงบ แต่ ถึง รบ ไม่ ขาด เอกราช จะ ไม่ ให้ ใคร ข่ม ขี่ สละ เลือด ทุก หยาด เป็น ชาติ พลี เถลิง ประเทศชาติไทยทวีมีชัยชโย
	Longest	วิธีการตัดคำแบบยาวที่สุด (Longest matching)	ประเทศไทย รวม เลือดเนื้อ ชาติ เชื้อ ไทย เป็น พระราชรัฐไพบ ของ ไทย ทุก ส่วน อยู่ ดำรง คง ไว้ ได้ ทั้งหมด ด้วย ไทย ล้วนหมาย รัก สามัคคี ไทย นี้ รัก สงบ แต่ ถึง รบ ไม่ ขาด เอกราช จะ ไม่ ให้ ใคร ข่ม ขี่ สละ เลือด ทุก หยาด เป็น ชาติ พลี เถลิง ประเทศชาติไทยทวีมีชัยชโย
อิงตามการเรียนรู้ของเครื่อง (Machine learning-based)	Deepcut	ผสมระหว่าง 1D-Convolutional neural network (CNN) กับ deep learning	ประเทศไทย รวม เลือดเนื้อ ชาติ เชื้อ ไทย เป็น พระราชรัฐไพบ ของ ไทย ทุก ส่วน อยู่ ดำรง คง ไว้ ได้ ทั้งหมด ด้วย ไทย ล้วนหมาย รัก สามัคคี ไทย นี้ รัก สงบ แต่ ถึง รบ ไม่ ขาด เอกราช จะ ไม่ ให้ ใคร ข่ม ขี่ สละ เลือด ทุก หยาด เป็น ชาติ พลี เถลิง ประเทศชาติไทยทวีมีชัยชโย
	AttaCut	CNN ที่มี Convolutional layer 3 ชั้น	ประเทศไทย รวม เลือดเนื้อ ชาติ เชื้อ ไทย เป็น พระราชรัฐไพบ ของ ไทย ทุก ส่วน อยู่ ดำรง คง ไว้ ได้ ทั้งหมด ด้วย ไทย ล้วนหมาย รัก สามัคคี ไทย นี้ รัก สงบ แต่ ถึง รบ ไม่ ขาด เอกราช จะ ไม่ ให้ ใคร ข่ม ขี่ สละ เลือด ทุก หยาด เป็น ชาติ พลี เถลิง ประเทศชาติไทยทวีมีชัยชโย

เราจึงสร้างแบบจำลองสำหรับแนะนำรหัสไอซีดีที่เกี่ยวข้องกับเนื้อหาในเอกสารบันทึกเวชระเบียน เพื่อเป็นตัวช่วยจัดการกับปัญหาเกี่ยวกับความไม่ครบถ้วนของรหัสไอซีดี เนื่องจากบันทึกการเข้ารับการรักษาต่อครั้งไม่ได้ถูกระบุด้วยรหัสไอซีดีเพียงรหัสเดียวประกอบกับรหัสไอซีดีมีจำนวนมาก แบบจำลองของเราจึงเป็นการจำแนกแบบหลายฉลาก (Multi-label classification) และแต่ละรหัสถูกใช้ด้วยความถี่แตกต่างกันขึ้นอยู่กับโอกาสที่ผู้ป่วยเป็นโรคนั้น ส่งผลให้รหัสไอซีดีมีกระจายตัวแบบแบบหางยาว (Long-tailed distribution) และรองรับข้อความผสมระหว่างภาษาไทยกับภาษาอังกฤษในชุดข้อมูลเวชระเบียน

ในส่วนที่เหลือต่อจากนี้จะกล่าวถึงเนื้อหาดังต่อไปนี้ บทที่ 3 ทบทวนวรรณกรรม บทที่ 4 ระเบียบวิธี บทที่ 5 ผลการทดลอง และสุดท้ายคือการอภิปรายและสรุปงานของเรา



3. บทบาททวนวรรณกรรม

3.1. การจำแนกข้อความ (Text Classification)

มีหลากหลายวิธีสำหรับแก้ปัญหาการจำแนกข้อความ แบบจำลอง Naïve Bayes เป็นแบบจำลองทางสถิติอย่างง่าย ซึ่งจะเลือกคลาสคำตอบที่เกี่ยวข้องที่สุดที่มีค่าความน่าจะเป็นสูงสุดจากความน่าจะเป็นของแต่ละคลาสทั้งหมด ความน่าจะเป็นของแต่ละคลาสมหาจากความน่าจะเป็นของคำในข้อความนำเข้า (input text) ซึ่งความน่าจะเป็นของแต่ละคำอ้างอิงจากที่เคยปรากฏในชุดข้อมูลฝึก (Training dataset) ของแต่ละคลาส เนื่องจากเป็นแบบจำลองที่มีพื้นฐานอ้างอิงจากความน่าจะเป็นจึงมีข้อจำกัดที่ชัดเจนในเรื่องความน่าจะเป็นมีค่าเป็นศูนย์ในคำที่อยู่นอกคลังศัพท์ [7] และความน่าจะเป็นลู่เข้าสู่ศูนย์ในข้อความที่ยาว ต่อมาหลายงานจึงหันไปใช้แบบจำลองการเรียนรู้เชิงลึก (Deep learning) แบบจำลองการเรียนรู้เชิงลึกสำหรับการจำแนกข้อความส่วนใหญ่มีพื้นฐานจาก Convolutional neural network (CNN) [8] และ Bidirectional Long-short term memory (Bi-LSTM) [9] แบบจำลอง CNN นั้นนอกจากจะนิยมใช้ในการจำแนกข้อความแล้วใช้อย่างแพร่หลายใน image processing อีกด้วย สำหรับการจำแนกข้อความจะพิจารณาคำติดกัน N คำด้วยขนาดของ kernel ซึ่งคล้ายคลึงกับจำนวน N ใน N-grams ในส่วนของแบบจำลอง Bi-LSTM นั้นเน้นรวบรวมคุณสมบัติของคำของข้อความที่ยาวจาก 2 ทิศทางคือ จากซ้ายไปขวา (Forward) และขวาไปซ้าย (Backward) และเก็บคุณสมบัติสำคัญไว้ในหน่วยความจำของ LSTM

ปัจจุบันงานแบบจำลองล่าสุดสำหรับการจำแนกข้อความคือ Bidirectional Encoder Representation from Transformers (BERT) ซึ่งถูกพัฒนาแบบจำลองที่ผ่านการฝึกด้วยข้อมูลจำเพาะ (Pre-trained model) เพื่อให้เข้ากับงานที่มีขอบเขตเฉพาะทาง เช่น Clinical BERT [10] ที่ถูกฝึกด้วยชุดข้อมูลบทความทางการแพทย์สำหรับจำแนกข้อความเกี่ยวกับสุขภาพหรือทางการแพทย์ BERT เป็นแบบจำลองที่มีโครงสร้างขนาดใหญ่กว่าแบบจำลองอื่น นั้นส่งผลให้แบบจำลองไม่เหมาะสมกับชุดข้อมูลขนาดเล็ก [11]

3.2. การจำแนกข้อความจากบันทึกเวชระเบียนหลายภาษา (Medical record classification with multilingual)

การจำแนกข้อความจากข้อมูลบันทึกเวชระเบียนนำมาประยุกต์ได้ในหลายงาน เช่น การเข้ารหัสไอซีดี การจำแนกโรคจากข้อความ การทำนายเพื่อวินิจฉัยโรค เป็นต้น โดยส่วนใหญ่จะรองรับเฉพาะเวชระเบียนภาษาอังกฤษ [12] วิธีการแก้ปัญหาสำหรับเวชระเบียนภาษาอื่นได้รับการพัฒนาโดยเน้นที่กระบวนการฝังคำ (Word embedding) [13] และการแปลภาษา [14, 15]

การฝังคำ (Word embedding) เป็นกระบวนการสำคัญในการจำแนกข้อความเพื่อแปลงข้อความเป็นเวกเตอร์ สำหรับข้อความภาษาเดียว (Monolingual) โดยปกติจะฝังคำด้วยการฝังคำที่ผ่านการฝึกล่วงหน้าสาธารณะที่เปิดให้นำไปใช้ได้ (An available public pre-trained word embedding) เช่น Word2Vec [16] GloVe [17] และ FastText [18] ซึ่งเป็นการฝังคำแบบคงที่แต่ละแบบถูกฝึกด้วยชุดข้อมูลฝึกที่แตกต่างกัน เพื่อขยายขอบเขตของคลังศัพท์ให้ครอบคลุมกับหลายงานมากขึ้น วิธีการ Meta-embedding เป็นวิธีรวมการฝังคำแบบคงที่เหล่านั้นโดยใช้วิธีการทางคณิตศาสตร์หลากหลายแบบให้ได้การฝังคำที่ครอบคลุมมากขึ้นและสอดคล้องกับงานที่ไม่ใช่ภาษาอังกฤษ [19-21]

เนื่องจากการฝังคำแบบคงที่ถูกรบกวนด้วยข้อมูลทั่วไป ซึ่งไม่สามารถใช้ได้จริงในงานที่มีขอบเขตเฉพาะ ดังนั้นการฝังคำแบบกำหนดเองจึงกลายมาเป็นทางเลือกสำหรับการทำให้เป็นเวกเตอร์ในขอบเขตเฉพาะ สำหรับข้อความที่มีหลายภาษา การแปลภาษาเป็นอีกทางเลือกหนึ่งที่เป็นที่นิยมในการแปลงให้อยู่ในรูปแบบปัญหาภาษาเดียว อีกทางเลือกคือการฝึกบน embedding layer ก่อนที่จะผ่านเข้าสู่แบบจำลองการจำแนก

3.3. การจำแนกรหัสไอซีดี (ICD classification)

การจำแนกรหัสไอซีดีด้วยบันทึกเวชระเบียนมีองค์ประกอบที่ท้าทายอยู่หลายส่วน เพื่อนำไปจำแนกรหัสไอซีดีได้อย่างมีประสิทธิภาพ เราจึงต้องแก้ปัญหา 2 ประเด็นคือการจำแนกแบบหลายฉลากมาก (Extreme multi-label classification) และการกระจายตัวแบบหางยาว (Long-tailed distribution)

เนื่องจากมีจำนวนรหัสไอซีดีจำนวนมากและโครงสร้างแบบอนุกรมวิธาน การจำแนกจึงกลายเป็นปัญหาแบบหลายฉลากมาก บางงานเพิ่มส่วน regularization ใน loss

function [8, 22] หรือรวม hierarchical dependencies ลงในแบบจำลองเพื่อทำให้การทำนายอนุกรมวิธานรหัสไอซีดีขึ้น [9, 23]

วิธี Binary relevance เป็นวิธีอย่างง่ายในการแปลงจาก 1 ปัญหาหลายผลากเป็นหลาย ๆ ปัญหาไบนารี [24] โดยแต่ละการจำแนกแบบไบนารีจะฝึกแบบอิสระต่อกัน มักใช้ร่วมกับการเรียนรู้ของเครื่องโดยเฉพาะแบบจำลอง Naïve Bayes

การจำแนกประเภทแบบหลายผลากมาก (Extreme multi-label classification) เป็นการจำแนกแบบหลายผลากที่มีจำนวนผลากมากจะมากพร้อมกับการกระจายตัวของชุดข้อมูลแบบหางยาว (Long-tailed distribution dataset) การกระจายตัวแบบหางยาวเป็นปัญหาเกี่ยวกับความถี่ของผลากไม่กระจายตัวแบบเอกรูป (Uniform distribution) ในขณะที่บางผลากมีความถี่สูง ผลากอื่นอาจมีน้อยหรือแทบไม่มีเลย ดังนั้นเพื่อรับมือกับปัญหานี้ การตั้งค่าสัดส่วนขั้นต่ำเป็นจำนวนคงที่หรือเปอร์เซ็นต์เพื่อพยายามลดผลกระทบจากข้อมูลที่ไม่สมดุล [25] อย่างไรก็ตามวิธีการนี้ผลากที่พบยากส่วนใหญ่จะถูกละทิ้งซึ่งส่งผลให้ขาดความทั่วไป (Lacking of general) ของแบบจำลอง อีกวิธีหนึ่งเกี่ยวกับความละเอียดของผลากเป็นการลดระดับชั้นของผลากซึ่งสามารถลดจำนวนของผลากที่ต้องพิจารณาได้ [13, 26]

แบบจำลองทั้ง 4 ประกอบไปด้วย Naïve Bayes CNN Bi-LSTM และ BERT มีส่วนที่เหมือนและแตกต่างในการแก้ปัญหาที่ตั้งไว้ เราสรุปลงในตารางที่ 8 ดังนี้

ตารางที่ 8 ตารางเปรียบเทียบของแบบจำลองที่เกี่ยวข้องระหว่างปัญหาที่แตกต่างกัน

ปัญหา	Naïve Bayes	CNN	Bi-LSTM	BERT
การจำแนกข้อความจากข้อความที่ยาว	✗	✓	✓	✓
ชุดข้อมูลขนาดเล็ก	✓	✓	✓	✗
ชุดข้อมูลหลายภาษา	✓	✓	✓	✓
การจำแนกประเภทแบบหลายผลาก	Binary relevance	✓	✓	✓
การกระจายตัวแบบหางยาว	✗	✗	✗	✗

ในด้านของปัญหาจากข้อความที่มีความยาวมาก ชุดข้อมูลหลายภาษา และการจำแนกประเภทแบบหลายคลาส ตัวจำแนกการเรียนรู้เชิงลึก (Deep learning classifiers) ได้แก่ CNN Bi-LSTM และ BERT รองรับข้อความนำเข้าที่ยาวและสามารถจำแนกงานแบบหลายคลาสได้ แต่แบบจำลอง Naïve Bayes ไม่สามารถหลีกเลี่ยงการลู่เข้าความน่าจะเป็นเป็นศูนย์ในข้อความที่มีความยาวมาก ๆ ได้ และตัวจำแนก Naïve Bayes สามารถแปลงให้อยู่ในรูปปัญหาการจำแนกแบบหลายคลาสโดยใช้วิธี Binary relevance

ตัวจำแนกการเรียนรู้เชิงลึกรองรับชุดข้อมูลหลายภาษาโดยใช้การฝังคำ (Word embedding) แต่ตัวจำแนก Naïve Bayes จะรองรับเมื่อฝึกด้วยชุดข้อมูลหลายภาษา ดังนั้นทั้ง 4 ตัวจำแนกสามารถรับมือกับทั้ง 3 ปัญหานี้ได้ อย่างไรก็ตามตัวจำแนกเหล่านั้นก็แตกต่างกันในส่วนของปัญหาชุดข้อมูลขนาดเล็ก ซึ่ง BERT ไม่เหมาะกับชุดข้อมูลขนาดเล็กในขณะที่ตัวจำแนกอื่นกลับทำงานได้ดี สำหรับปัญหาการกระจายตัวแบบหางยาวไม่มีตัวจำแนกใดสามารถรับมือกับปัญหานี้ได้เพียงลำพัง

ประเด็นทั้งหมดที่กล่าวมาในข้างต้นนำไปสู่การออกแบบแบบจำลองของเรา แบบจำลองของเราประกอบด้วย 2 ส่วนเพื่อรับมือกับปัญหาการกระจายตัวแบบหางยาวและลดจำนวนคลาสที่ต้องนำมาพิจารณาในแต่ละแบบจำลอง ส่วนแรกรองรับกลุ่มตัวอย่างของชุดข้อมูลที่สโตนที่พบยากซึ่งจำแนกด้วยแบบจำลอง Naïve Bayes อีกส่วนหนึ่งสำหรับชุดข้อมูลที่สโตนที่พบได้ทั่วไปซึ่งถูกจำแนกด้วยวิธีการรวม (Ensemble method) ของตัวจำแนกที่เหลือและต่อมาก็นำคำตอบจากทั้ง 2 ส่วนมาต่อกันเป็นคำตอบสุดท้าย

4. ระเบียบวิธี

4.1. ชุดข้อมูล (Dataset)

4.1.1. ชุดข้อมูลความรู้จำเพาะ (Domain Knowledge dataset)

ชุดข้อมูลความรู้เฉพาะใช้เป็นคำสำคัญ (Keywords) เพื่อประกอบการตัดสินใจของแบบจำลอง ข้อมูลประกอบด้วยคำคล้าย คลังศัพท์ที่เกี่ยวข้อง และคำบรรยายรหัส ICD-9-CM ICD-10 และ ICD-10-TM แหล่งข้อมูลและที่มาสรุปอยู่ในตารางที่ 9 ส่วนของ ICD-9-CM เป็นคำบรรยายเกี่ยวกับหัตถการรักษารวบรวมข้อมูลมาจากศูนย์มาตรฐานรหัสและข้อมูลแห่งชาติ (Thai health coding center: THCC) [27] สำหรับทั้ง ICD-10 และ ICD-10-TM เป็นคำบรรยายเกี่ยวกับการวินิจฉัยโรคซึ่งรวบรวมมาจากศูนย์มาตรฐานรหัสและข้อมูลแห่งชาติและฐานข้อมูลออนไลน์ขององค์การอนามัยโลก [28] และชุดข้อมูลคลังศัพท์เพิ่มเติม เรารวบรวมจากระบบ (Unified Medical Language System: UMLS) และ SNOMED-CT ซึ่งรวบรวมคำคล้ายต่าง ๆ ที่เกี่ยวข้องทั้งโรคและหัตถการ ชุดข้อมูลเหล่านี้จะแบ่งไปสร้างการฝังคำและนำไปฝึกแบบจำลองร่วมกับชุดข้อมูลเวชระเบียนเพิ่มเติม ให้แบบจำลองเรียนรู้คำศัพท์และความสัมพันธ์ได้มากขึ้น

ตารางที่ 9 แหล่งที่มาของข้อมูลไอซีดี

ข้อมูล	เนื้อหา	แหล่งที่มา
ICD-9-CM	หัตถการ	ศูนย์มาตรฐานรหัสและข้อมูลแห่งชาติ
ICD-10	การวินิจฉัยโรค	ฐานข้อมูลออนไลน์ขององค์การอนามัยโลก
ICD-10-TM	การวินิจฉัยโรค	ศูนย์มาตรฐานรหัสและข้อมูลแห่งชาติ
คลังศัพท์เพิ่มเติม	การวินิจฉัยโรคและหัตถการ	UMLS และ SNOMED-CT

4.1.2. ชุดข้อมูลอักขรย่อทางการแพทย์ (Medical acronym)

ในบันทึกเวชระเบียนเราอาจพบคำย่อได้ ซึ่งอาจมีทั้งคำย่อแบบมาตรฐานและไม่เป็นมาตรฐาน ทั้งนี้แบบจำลองของเราใช้เพียงชุดข้อมูลความรู้เฉพาะนั้นไม่เพียงพอให้ตัดสินใจได้ดีขึ้น ดังนั้นเราจึงรวบรวมชุดข้อมูลคำย่อทางการแพทย์โดยอ้างอิงจากหนังสือพจนานุกรมคำย่อชื่อ “อักขรย่อที่หมอใช้ 2017” [29] ซึ่งเราจะนำคำเต็มมาแทนที่คำย่อในข้อความในขั้นตอนประมวลผลก่อนฝึก ตัวอย่างชุดข้อมูลคำย่อของเราแสดงในตารางที่ 10

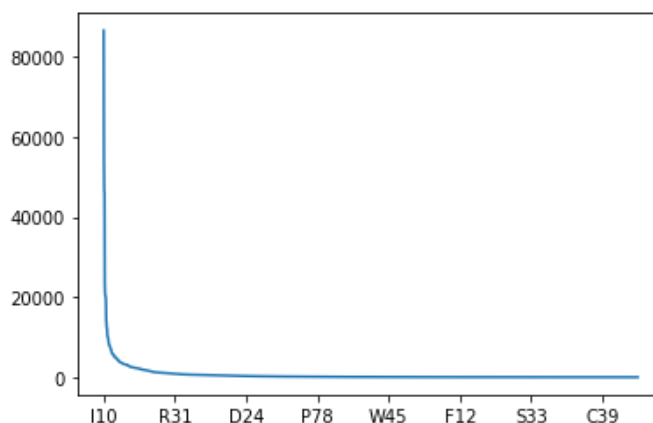
ตารางที่ 10 ตัวอย่างอักขรย่อและคำเต็ม

คำย่อ	คำเต็ม
rt	radiologic technologist
rt	radiotherapy
rt	reaction time
rt	right

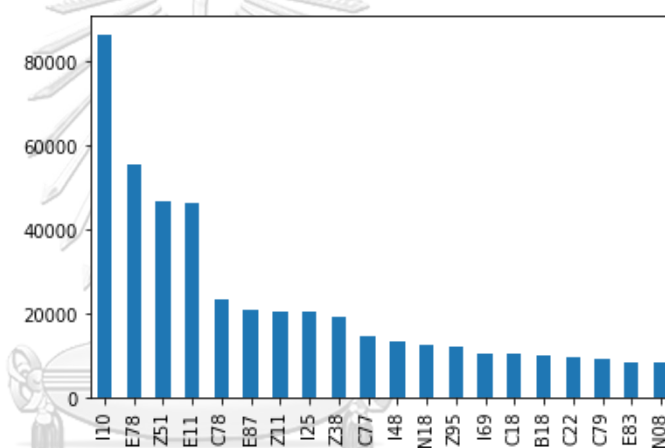
4.1.3. ชุดข้อมูลเวชระเบียน (Medical record dataset)

4.1.3.1. บันทึกการรักษา (Course notes)

บันทึกการรักษา (Course note) บรรยายเกี่ยวกับโรคที่วินิจฉัย อาการของผู้ป่วย ผลตรวจแลปและค่าสัญญาณชีพซึ่งบันทึกด้วยแพทย์และพยาบาล ชุดข้อมูลบันทึกการรักษาของเราระบุด้วยรหัส ICD-10-TM ซึ่งจะพิจารณาเพียงแค่ 3 หลักแรกที่เป็นส่วนของหมวดหมู่โรค เรามีทั้งหมด 309,176 บันทึกกับ 1,875 รหัสไอซีดี ซึ่งแต่ละรหัสไอซีดีมีจำนวนตัวอย่างไม่เท่ากันอย่างมาก (Extremely imbalanced data) และมีการกระจายตัวแบบหางยาวดังแสดงในรูปที่ 3 ซึ่งจะเห็นว่าจำนวนตัวอย่างของรหัสที่พบบ่อย (โรคที่พบได้ทั่วไป) มีจำนวนต่างกันอย่างกับกลุ่มรหัสที่พบบาก (โรคที่พบผู้ป่วยเป็นน้อย) สำหรับรูปที่ 4 และตารางที่ 11 แสดงจำนวนตัวอย่างของรหัสที่มีตัวอย่างมากที่สุด 5 อันดับแรก เราจึงตัดสินใจแบ่งกลุ่มชุดข้อมูลตามจำนวนตัวอย่างของรหัสไอซีดีโดยสำรวจช่วงของขนาดตัวอย่างที่เหมาะสมจะใช้เป็นเกณฑ์ในการแบ่งกลุ่มตามตารางที่ 12



รูปที่ 3 การกระจายตัวของจำนวนตัวอย่างต่อรหัสไอซีดี-10



รูปที่ 4 แสดงภาพรวมของจำนวนตัวอย่างในกลุ่มรหัสไอซีดี-10 ส่วนที่พบมาก

ตารางที่ 11 ตารางแสดงรหัสที่มีจำนวนตัวอย่างมากที่สุด 5 อันดับแรกจากรูปที่ 4

อันดับ	รหัสไอซีดี-10	จำนวนตัวอย่าง
1	I10: Essential (primary) hypertension	87,579
2	E78: Disorders of lipoprotein metabolism and other lipidaemia	55,952
3	Z51: Other medical care	48,923
4	E11: Non-insulin-dependent diabetes mellitus	47,228
5	C78: Secondary malignant neoplasm of respiratory and digestive organs	32,288

ตารางที่ 12 จำนวนรหัสไอซีดี-10 ในแต่ละช่วงจำนวนตัวอย่าง

จำนวนตัวอย่าง (Samples)	จำนวนรหัสไอซีดี	ปริมาณรหัสไอซีดี (%)
1 ตัวอย่าง	119	6.35
< 10 ตัวอย่าง	433	23.14
< 50 ตัวอย่าง	844	45.11
< 100 ตัวอย่าง	1,047	55.96
< 1,000 ตัวอย่าง	1,650	88.19
< 10,000 ตัวอย่าง	1,854	99.09

จากตารางที่ 12 พบว่ามีรหัสไอซีดี-10 ถึง 119 รหัสที่มีตัวอย่างเพียง 1 บันทึก ส่วนจำนวนรหัสไอซีดีที่มีจำนวนตัวอย่างน้อยกว่า 50 และ 100 บันทึกมี 844 และ 1,047 รหัส คิดเป็นร้อยละ 45.11 และ 55.96 ตามลำดับ

เนื่องจากเราต้องการแบ่งกลุ่มชุดข้อมูลบันทึกการรักษาออกเป็น 2 กลุ่ม เพื่อแยกไปฝึกกับแบบจำลองที่แตกต่างกัน เราจึงตัดสินใจเลือกใช้ขนาดตัวอย่างเท่ากับ 100 เป็นเกณฑ์เพราะหากเราใช้ขนาดตัวอย่างเท่ากับ 50 จำนวนตัวอย่างต่อรหัสจะน้อยเกินไปซึ่งอาจส่งผลกระทบต่อประสิทธิภาพการตัดสินใจในแบบจำลอง Naïve Bayes ได้

CHULALONGKORN UNIVERSITY

4.1.3.2. บันทึกการให้ยา (Medication notes)

บันทึกการให้ยา (Medication note) บรรยายเกี่ยวกับหัตถการรักษาและยาที่ใช้ในการรักษาผู้ป่วยขณะพักรักษาตัวอยู่ที่โรงพยาบาลและชุดข้อมูลส่วนนี้ระบุด้วยรหัส ICD-9-CM เรามีตัวอย่างบันทึกการให้ยาอยู่ 308,965 บันทึกกับ 3,157 รหัสไอซีดีที่ไม่มีการแบ่งส่วนของรหัสชัดเจนแบบ ICD-10-TM เราจึงพิจารณาจำแนกรหัสทั้งหมด ไม่เลือกแค่ส่วนหมวดหมู่แบบ ICD-10-TM ซึ่งแสดงในตารางที่ 13 นอกจากนี้ชุดข้อมูลมีการกระจายตัวของข้อมูลแบบเดียวกับชุดบันทึกการรักษาจะเห็นได้ในรูปที่ 5 สำหรับรูปที่ 6 และตารางที่ 14 แสดงจำนวนตัวอย่างของรหัสที่มีตัวอย่างมากที่สุด 5 อันดับแรก เราจึงแก้ปัญหาด้วยวิธีเดียวกันคือข้อมูลออกเป็น 2

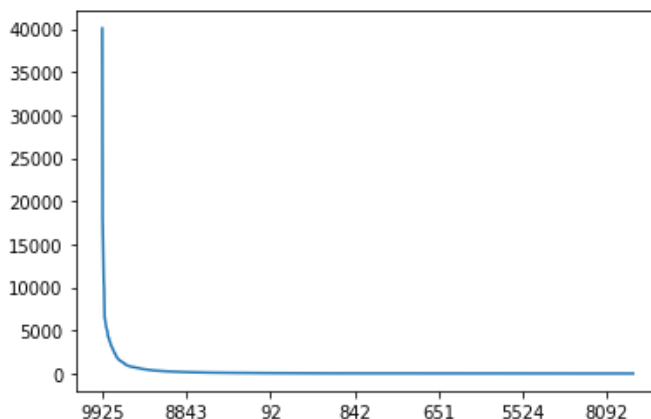
ชุดและเราได้สรุปการสำรวจช่วงขนาดตัวอย่างต่อรหัสเพื่อหาเกณฑ์ขนาดที่เหมาะสมในการแบ่งชุดข้อมูลนี้ลงในตารางที่ 15

ตารางที่ 13 ตัวอย่างบันทึกการให้ยาและรหัสไอซีดีที่เกี่ยวข้อง

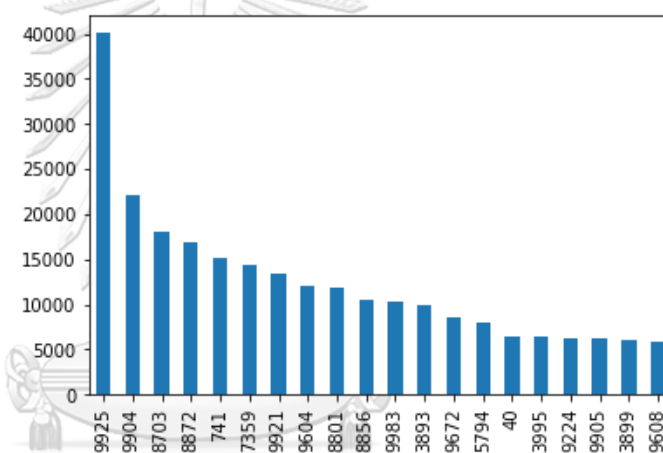
ตัวอย่างบันทึกการให้ยา	icd9cm
' - BD (1:1) 400ml x4 feeds via gastrostomy- Amlodipine(5) 1x1 o pc - Depakin Solution 2.5 ml o q12hr- Lyrica(75) 2xhs- Fluoride(0.25) 3xhs- MOM 30ml o hs- Paracet(500) 2xprn- TWC(15) 1xprn,q8hr- Flunarizine 2xhs- Calcium(CDR) 1x2 opc- Mydocalm 1x3 o pc - Tramol(50 1xprn, q6hr Closed [endoscopic] biopsy of small intestine Biopsy of gum Restoration of tooth by filling Injection of antibiotic Replacement of gastrostomy tube Magnetic resonance imaging of brain and brain stem Magnetic resonance imaging of spinal canal Assisting exercise Ambulation and gait training Insertion of indwelling urinary catheter Other mammography Bone scan	['4514', '2411', '232', '9921', '9702', '8891', '8893', '9311', '9322', '5794', '8737', '9214']

ตารางที่ 14 ตารางแสดงรหัสที่มีจำนวนตัวอย่างมากที่สุด 5 อันดับแรก

อันดับ	รหัสไอซีดี-9	จำนวนตัวอย่าง
1	9925: Injection or infusion of cancer chemotherapeutic substance	40,081
2	9904: Transfusion of packed cells	22,190
3	8703: Computed tomography of paranasal sinuses	18,008
4	8872: Color doppler ultrasound of the heart transesophageal approach	16,846
5	0741: Incision of adrenal gland	15,191



รูปที่ 5 การกระจายตัวของจำนวนตัวอย่างต่อรหัสไอซีดี-9



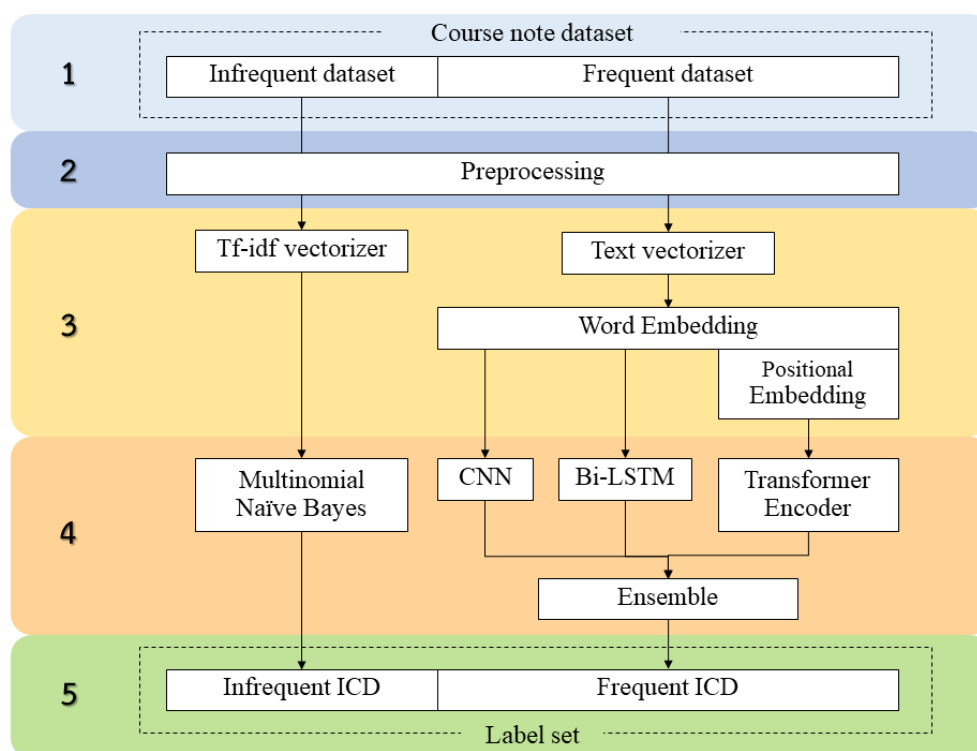
รูปที่ 6 แสดงภาพรวมของจำนวนตัวอย่างในกลุ่มรหัสไอซีดี-9 ส่วนที่พบมาก

จากตารางที่ 15 พบว่ามีรหัสไอซีดี-9 ถึง 337 รหัสที่มีตัวอย่างเพียง 1 บันทึก ส่วนจำนวนรหัสไอซีดีที่มีจำนวนตัวอย่างน้อยกว่า 50 และ 100 บันทึกมี 2,200 และ 2,511 รหัส คิดเป็นร้อยละ 69.69 และ 79.54 ตามลำดับ จะเห็นว่าจำนวนของรหัสไอซีดีทั้ง 2 ช่วงไม่ต่างกันมากแต่ว่าในช่วงเกณฑ์ที่น้อยกว่า 50 ตัวอย่างนั้นมีจำนวนรหัสไอซีดีเกินครึ่งหนึ่งของรหัสไอซีดีทั้งหมด หากยังใช้เกณฑ์ขนาดตัวอย่างเท่ากับ 100 เหมือนชุดข้อมูลบันทึกการรักษอาจเกิด Overfitting ในแบบจำลองสำหรับชุดข้อมูลกลุ่มไอซีดีหายากและขาดสมดุลในแบบจำลอง

ตารางที่ 15 จำนวนรหัสไอซีดี-9 ในแต่ละช่วงจำนวนตัวอย่าง

จำนวนตัวอย่าง (Samples)	จำนวนรหัสไอซีดี	ปริมาณรหัสไอซีดี (%)
1 ตัวอย่าง	337	10.67
< 10 ตัวอย่าง	1,302	41.24
< 50 ตัวอย่าง	2,208	69.69
< 100 ตัวอย่าง	2,511	79.54
< 1,000 ตัวอย่าง	3,014	95.47
< 10,000 ตัวอย่าง	3,146	99.65

4.2. กระบวนการทดลอง (Methodology)



รูปที่ 7 ภาพรวมขั้นตอนและโครงสร้างแบบจำลอง

ในส่วนนี้จะอธิบายโครงสร้างแบบจำลองของพวกเรา มีทั้งหมด 4 ขั้นตอนซึ่งรายละเอียดแสดงในรูปที่ 7 ขั้นแรกคือการเตรียมข้อมูลเป็นขั้นการแบ่งชุดข้อมูลด้วยเกณฑ์ที่กำหนด ขั้นที่สองคือกระบวนการประมวลผลล่วงหน้าซึ่งเกี่ยวกับการทำให้บริสุทธิ์ (Cleansing) และการตัดคำ (Tokenization) ขั้นต่อมาคือการสร้างการฝังคำสำหรับชุดข้อมูลสองภาษา ขั้นสุดท้ายเป็นการรวม 4 ตัวจำแนก เราจะอธิบายแต่ละขั้นตอนต่อจากนี้

4.2.1. การเตรียมข้อมูล (Data preparation)

จากปัญหาการกระจายตัวแบบหางยาว ไม่มีแบบจำลองเดี่ยวที่เข้ากับข้อมูลที่กระจายตัวแบบนั้นและไม่สามารถทำให้ข้อมูลสมดุลได้ด้วยวิธีการเพิ่มตัวอย่าง (Oversampling) เพราะทำให้แบบจำลองเกิด overfitting ซึ่งในชุดข้อมูลมีถึง 119 (ไอซีดี-10) และ 337 (ไอซีดี-9) รหัสที่มีตัวอย่างเพียง 1 ตัวอย่างเท่านั้น อีกทางหนึ่งคือการตัดข้อมูลทิ้งไป วิธีการนี้เป็นการตัดคลาสที่มีจำนวนตัวอย่างน้อยออกแต่การทำเช่นนี้จะทำให้แบบจำลองละทิ้งและไม่ได้ฝึกข้อมูลของคลาสที่พบบยาก (Infrequent classes)

ดังนั้นเราจึงแบ่งชุดข้อมูลบันทึกเวชระเบียนออกเป็น 2 ชุดโดยใช้เกณฑ์ขนาดของตัวอย่าง รหัสไอซีดีที่มีขนาดตัวอย่างมากกว่าเกณฑ์ เราจะเรียกว่ารหัสไอซีดีทั่วไป (Frequent ICD codes) ส่วนที่เหลือเราจะเรียกว่ารหัสไอซีดีพบบยาก (Infrequent ICD codes) ตัวอย่างของรหัสไอซีดีทั่วไปและพบบยากจะเป็นชุดข้อมูลเวชระเบียนทั่วไป (Frequent dataset) และพบบยาก (Infrequent dataset) ตามลำดับ

สำหรับชุดข้อมูลบันทึกการรักษา เราใช้เกณฑ์ขนาดตัวอย่างที่ 100 ได้ชุดข้อมูลทั้งสองที่มีรายละเอียดทางสถิติตามตารางที่ 16 ในชุดข้อมูลบันทึกการรักษาเรามีทั้งหมด 309,176 บันทึกแบ่งเป็นชุดบันทึกการรักษาทั่วไป (รหัสไอซีดีพบบ่อย) 307,284 และชุดบันทึกการรักษาของรหัสที่พบบยาก 26,062 บันทึก จำนวนคำโดยเฉลี่ยในแต่ละบันทึกเท่ากับ 691 และ 878 คำตามลำดับ จำนวนรหัสไอซีดีที่ใช้พิจารณาในแบบจำลองมี 828 และ 1,047 รหัส เฉลี่ย 4.51 และ 1.20 รหัสต่อบันทึกตามลำดับ ในตารางมีค่าความหนาแน่นของรหัส (Label density) เป็นค่าที่ใช้แสดงถึงสถานะของปัญหาการจำแนกแบบหลายผลกว่ามีปริมาณหลากหลายหรือน้อยเท่าใด ซึ่งคำนวณจากการหารค่าเฉลี่ยรหัสต่อบันทึกด้วยจำนวนรหัสทั้งหมด ค่าความหนาแน่นยิ่งน้อย

หมายถึงปัญหานี้มีจำนวนรหัสในการพิจารณาและมีจำนวนตัวอย่างฝึกต่อรหัสน้อย ซึ่งค่าความหนาแน่นที่ได้เท่ากับ 0.0055 และ 0.0011 ตามลำดับและเป็นค่าที่น้อยมาก

ตารางที่ 16 ข้อมูลทางสถิติเชิงปริมาณของชุดบันทึกการรักษา

	ทั้งหมด	ชุดข้อมูลทั่วไป (รหัสที่พบบ่อย)	ชุดข้อมูลของรหัส ที่พบบยาก
จำนวนบันทึก	309,176	307,284	26,062
ปริมาณคำศัพท์	353,423	351,530	88,490
ค่าเฉลี่ยจำนวนคำต่อบันทึก	690	691	878
จำนวนคำสูงสุดต่อบันทึก	18,237	18,237	12,873
จำนวนรหัสไอซีดี	1,875	828	1,047
ค่าเฉลี่ยรหัสต่อบันทึก	4.16	4.51	1.20
ความหนาแน่นของรหัส	0.0022	0.0055	0.0011

สำหรับชุดข้อมูลบันทึกการให้ยา เราเลือกเกณฑ์ขนาดตัวอย่างที่ 50 แทนที่จะเป็น 100 เหมือนชุดข้อมูลที่แล้ว เนื่องจากต้องการรักษาสมดุลจำนวนรหัสไอซีดีไม่ให้แบบจำลองทางสถิติที่พิจารณาชุดข้อมูลรหัสหายากรับภาระมากเกินไปจนเกิด Overfitting ชุดบันทึกการให้ยามีทั้งหมด 308,965 บันทึก แบ่งเป็น 303,231 บันทึกในชุดบันทึกการให้ยาทั่วไปและ 22,084 บันทึกในชุดบันทึกการให้ยาของรหัสหายาก มีจำนวนเฉลี่ยของคำต่อบันทึกอยู่ที่ 299 และ 510 คำ มีรหัสไอซีดีที่พิจารณา 949 และ 2,208 รหัส ค่าความหนาแน่นของรหัสอยู่ที่ 0.00259 และ 0.00054 ตามลำดับ แสดงในตารางที่ 17

ตารางที่ 17 ข้อมูลทางสถิติเชิงปริมาณของชุดบันทึกการให้ยา

	ทั้งหมด	ชุดข้อมูลทั่วไป (รหัสที่พบบ่อย)	ชุดข้อมูลของรหัส ที่หายาก
จำนวนบันทึก	308,965	303,231	22,084
ปริมาณคำศัพท์	29,137	28,507	10,294
ค่าเฉลี่ยจำนวนคำต่อบันทึก	296	297	510
จำนวนคำสูงสุดต่อบันทึก	19,280	19,280	19,280
จำนวนรหัสไอซีดี	3,157	949	2,208
ค่าเฉลี่ยรหัสต่อบันทึก	2.49	2.45	1.19
ความหนาแน่นของรหัส	0.00079	0.00259	0.00054

ขณะที่ฝึกและวัดผลแบบจำลอง เราจะแบ่ง 80% เป็นชุดข้อมูลฝึก (Training set) และ 20% เป็นชุดข้อมูลทดสอบ (Test set) ขณะฝึกแบบจำลองเราแบ่ง 20% จากชุดข้อมูลฝึกมาเป็นชุดข้อมูลตรวจสอบ (Validation set) โดยชุดข้อมูลทั่วไปทั้งบันทึกการรักษาและการให้ยาจะนำไปใช้กับแบบจำลองการเรียนรู้เชิงลึก ส่วนชุดข้อมูลของรหัสหายากจะนำไปใช้กับแบบจำลอง Naïve Bayes

4.2.2. กระบวนการประมวลผลล่วงหน้า (Preprocessing)

กระบวนการประมวลผลล่วงหน้าของบันทึกเป็นกระบวนการกำหนดมาตรฐานข้อความเพื่อลดความกำกวม [30] ประกอบด้วยขั้นตอนดังต่อไปนี้

ขั้นที่ 1

แปลงตัวอักษรภาษาอังกฤษให้เป็นตัวอักษรพิมพ์เล็ก (Lowercase)

ขั้นที่ 2

ลบคำที่ไม่มีตัวอักษรพยัญชนะออก (Non-alphabetical character) เช่น วันที่และเวลา เป็นต้น แต่จะยังเก็บคำ เช่น 20mg L8 ไว้ ดังตัวอย่างในตารางที่ 18

ขั้นที่ 3

ลบเครื่องหมายวรรคตอน (Punctuation) และสัญลักษณ์อื่น ๆ เช่น ดอกจัน ขีด วงเล็บปีกกา เป็นต้น ดังตัวอย่างในตารางที่ 18

ขั้นที่ 4

ตัดคำจากข้อความโดยใช้ตัวตัดคำ AttaCut จากไลบรารี PyThaiNLP ซึ่งรองรับการตัดคำจากทั้งข้อความภาษาไทยและอังกฤษได้อย่างมีประสิทธิภาพและรวดเร็ว

รายละเอียดเพิ่มเติมได้บรรยายไว้ในข้อ 2.3. การตัดคำภาษาไทย

ขั้นที่ 5

ลบคำทั่วไป (Stopwords) ของทั้งคำไทยและภาษาอังกฤษ ซึ่งคำทั่วไปภาษาไทยอ้างอิงจากคลังศัพท์ของ PyThaiNLP และคำทั่วไปภาษาอังกฤษจากคลังศัพท์ของ NLTK [31]

เช่น ฉัน, เธอ คุณ มี I you are is am the เป็นต้น

ขั้นที่ 6

แทนคำย่อด้วยคำเต็มทั้งหมดที่เป็นไปได้โดยอ้างอิงจากหนังสือ “อักษรย่อที่หมอใช้ 2017” ดังตัวอย่างในตารางที่ 18

ตารางที่ 18 ตัวอย่างข้อความก่อนและหลังผ่านการประมวลผลตามขั้นตอน

ขั้นตอนการประมวลผลล่วงหน้า	ตัวอย่างการเปลี่ยนแปลงของข้อความ	
	ก่อน	หลัง
ขั้นที่ 2	12/12/2020: posterior disc bulging L5 / S1	posterior disc bulging L5 S1
ขั้นที่ 3	Tuberculosis of intestine (large)(small)+	Tuberculosis of intestine large small
ขั้นที่ 6	ORIF with dynamic hip blade and antirotating screw	Open reduction with internal fixation with dynamic hip blade and antirotating screw
	OCTR rt	open carpal tunnel release radiologic technologist radiotherapy reaction time Right

4.2.3. การฝังคำ (Word Embedding)

4.2.3.1. การฝังคำที่ผ่านการฝึกเบื้องต้น (Pre-trained Word embedding)

การฝังคำคือกระบวนการแปลงคำจากข้อความนำเข้าเป็นเวกเตอร์ สำหรับแบบจำลองการเรียนรู้เชิงลึก การฝังคำสาธารณะที่ผ่านการฝึกมาก่อนไม่สามารถใช้งานร่วมกับชุดข้อมูลของเราในแง่ของขอบเขตเฉพาะทางและความสัมพันธ์เชิงสองภาษา จากตารางที่ 19 การทดลองเปรียบเทียบเบื้องต้นนี้ทดสอบด้วยแบบจำลอง Bi-LSTM บนชุดข้อมูลบันทึกการรักษาที่นำมาเฉพาะบันทึกของรหัส 50 อันดับแรก ซึ่งจะเห็นว่าแบบจำลองที่ไม่ใช่ Pre-trained word embedding ให้ประสิทธิภาพที่ต่ำกว่าแบบที่ใช้ static pre-trained word embedding แต่เราทราบว่าหน้าที่เวกเตอร์ของคำผ่านการฝึกมาก่อนที่จะนำมาประยุกต์ใช้กับแบบจำลองอื่นนั้นจะช่วยเพิ่มประสิทธิภาพของแบบจำลองและให้ผลการทำนายที่ดีขึ้นได้

ตารางที่ 19 ผลการทดลองเบื้องต้นเปรียบเทียบประสิทธิภาพแบบจำลองที่ใช้การฝังคำที่แตกต่างกัน

การฝังคำ (Word embedding)	ความยาวประโยค	Train	Validation	Test
ไม่ใช่ Pre-trained WE ¹	300	0.7952	0.7949	0.7765
	500	0.7803	0.7808	0.7660
GloVe + Thai2Fit ²	300	0.6881	0.6882	0.6723
	500	0.6931	0.6937	0.6722
FastText ³	300	0.7076	0.7079	0.6960
	500	0.7085	0.7084	0.6961

¹Word Embedding

²การรวมกันของเวกเตอร์ GloVe สำหรับภาษาอังกฤษและ Thai2fit สำหรับคำภาษาไทย [32]

³การรวมกันของเวกเตอร์ FastText ที่ฝึกด้วยคำไทยและคำอังกฤษ

ดังนั้นเราจึงสร้างการฝังคำที่ผ่านการฝึกมาก่อนเอง (Pre-trained word embedding) โดยใช้แบบจำลอง Continuous bag-of-words (CBOW) จากไลบรารี Gensim [33] CBOW จะถูกฝึกด้วยข้อมูลบันทึกเวชระเบียนและคำอธิบายรหัสไอซีดีที่

มีเนื้อความผสมระหว่างภาษาไทยและภาษาอังกฤษ ให้แบบจำลองเรียนรู้ความสัมพันธ์แบบสองภาษาและความสัมพันธ์ระหว่างคำศัพท์กับรหัสไอซีดี การฝังคำของเราจะเป็นเมทริกซ์ที่ตั้งต้นใน embedding layer ของตัวจำแนกการเรียนรู้เชิงลึก รูปที่ 8 คือคำที่มีเวกเตอร์ใกล้เคียงกับคำตัวอย่าง

```
model.wv.most_similar(positive=["cholera"])
[('tetanus', 0.9391313791275024),
 ('paratyphoid', 0.9375733137130737),
 ('pertussis', 0.9347878694534302),
 ('typhoid', 0.9327477812767029),
 ('A21', 0.9316585063934326),
 ('ลิซมาเนีย', 0.9293429851531982),
 ('brucellosis', 0.928986668586731),
 ('whooping', 0.9285757541656494),
 ('A23', 0.9271388053894043),
 ('B55', 0.9263607263565063)]
```



รูปที่ 8 ตัวอย่างคำใกล้เคียงของคำว่า 'cholera'

เราใช้ embedding layer เพื่อสร้างเวกเตอร์ของข้อความนำเข้า $[w_1, w_2, \dots, w_n]$ เมื่อ w_i แทนตำแหน่งของคำและ n แทนความยาวของข้อความ Embedding layer จะเลือกเวกเตอร์ที่ฝังล่วงหน้า (Pre-trained vector) ของคำเหล่านั้นตามตำแหน่งประกอบขึ้นเป็นเมทริกซ์ของการฝังคำจากข้อความนำเข้า ซึ่งมีขนาด $n * d_{emb}$ เมื่อ d_{emb} มีขนาด 300

ใน Transformer encoder เราเพิ่มการฝังตำแหน่งของคำแสดงความสัมพันธ์เชิงลำดับเข้าไป ดังนั้นการฝังคำของแบบจำลอง Transformer encoder เกิดจากผลบวกของเวกเตอร์การฝังคำและเวกเตอร์การฝังตำแหน่งของคำ (Positional embedding) โดยใช้ TokenAndPositionEmbedding layer จาก Keras NLP [34] ซึ่งเป็น API สำหรับกลุ่มงานศึกษาการประมวลผลภาษาธรรมชาติ

4.2.3.2. การแปลงเวกเตอร์ของข้อความ (Text vectorization)

แบบจำลอง Naïve Bayes ใช้ตัวแปลงเป็นเวกเตอร์ (Vectorizer) ต่างจากแบบจำลองอื่น เราแปลงคำให้เป็นเวกเตอร์ของ TF-IDF (Term Frequency–Inverse Document Frequency) ขนาด N มิติ เมื่อ N เป็นขนาดของคลังศัพท์ (Vocabulary size) โดยใช้โมดูล TfidfVectorizer จากไลบรารี Scikit-learn [35] ดังนั้นแบบจำลอง Naïve Bayes จะรับข้อความที่ผ่านการทำให้บริสุทธิ์แล้ว (Cleansing) ซึ่งยังมีลักษณะเป็นข้อความอักขรอยู่เข้าไปยัง TfidfVectorizer โดยตรง แต่แบบจำลองการเรียนรู้เชิงลึกจะนำข้อความอักขรแปลงเป็นเวกเตอร์ที่ประกอบไปด้วยตัวเลขแทนตำแหน่งของคำในพจนานุกรมโดยใช้ TextVectorization จาก Keras API

```
>>> vectorizer.get_vocabulary()

dict_items([('เจ็บ', 148), ('มือ', 388), ('right', 75), ('hand', 3374),...])

sample = "เจ็บมือ right hand"

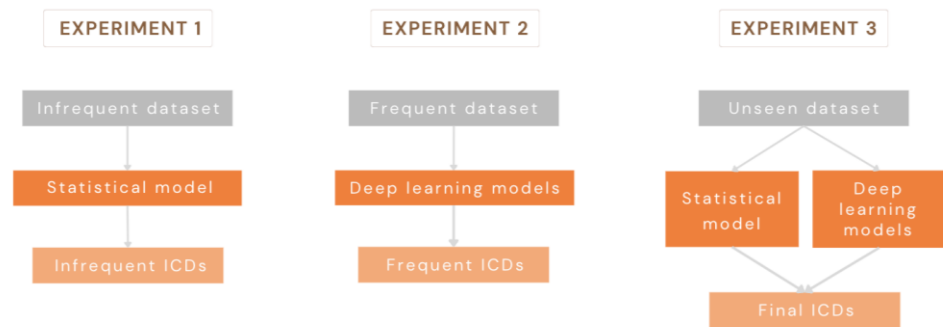
output = vectorizer([[clean_text(sample)]])
output.numpy()[0, :6]

array([ 148,  388,   75, 3374,    0,    0])
```

รูปที่ 9 ตัวอย่างเวกเตอร์ของคำจากการผ่าน TextVectorization

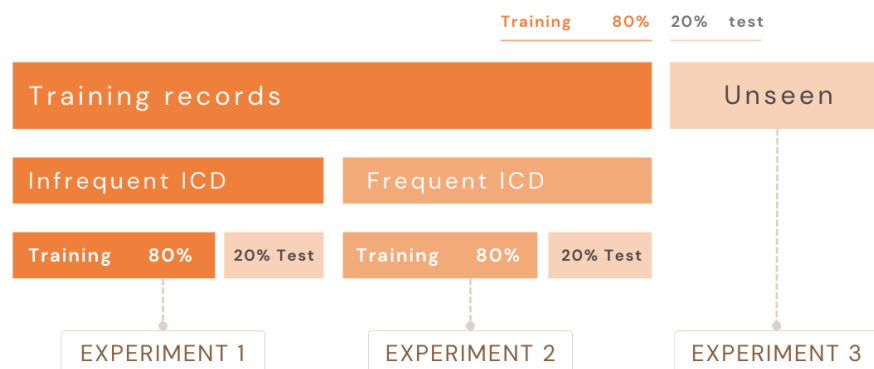
4.2.4. การสร้างแบบจำลอง (Modeling)

การวิจัยแบ่งการทดลองออกเป็น 3 การทดลอง รูปที่ 10 แสดงภาพไดอะแกรมของการทดลองทั้ง 3 แบบซึ่งมีรายละเอียดดังนี้ การทดลองที่ 1 (Experiment 1) วัดประสิทธิภาพการจำแนกรหัสไอซีทีกลุ่มหายากด้วยแบบจำลอง Naïve Bayes การทดลองที่ 2 (Experiment 2) วัดประสิทธิภาพการจำแนกรหัสไอซีทีกลุ่มพบได้ทั่วไปด้วยแบบจำลองการเรียนรู้เชิงลึก (Deep learning) และการทดลองที่ 3 (Experiment 3) วัดประสิทธิภาพการจำแนกรหัสไอซีทีทั้ง 2 กลุ่มด้วยแบบจำลองเชิงผสมระหว่างแบบจำลอง Naïve Bayes และแบบจำลองเชิงลึก ดังนั้นการทดลองทั้งสามจึงใช้ชุดข้อมูลทดสอบที่ต่างกัน



รูปที่ 10 การทดลองทั้งหมดในงานวิจัย

ชุดข้อมูลสำหรับฝึกและทดสอบ (Training and test dataset) ของแต่ละการทดลองมาจากชุดข้อมูลเวชระเบียนที่ร้อยละ 80 เป็นชุดข้อมูลสำหรับฝึก และร้อยละ 20 เป็นชุดข้อมูลสำหรับทดสอบซึ่งชุดข้อมูลทดสอบนี้จะถูกใช้ในการทดลองที่ 3 ส่วนชุดข้อมูลสำหรับฝึกจะถูกแบ่งออกเป็น 2 กลุ่มคือ ชุดข้อมูลบันทึกเวชระเบียนของรหัสไอซีดีที่พบยาก (Infrequent ICD dataset) สำหรับใช้ในการทดลองที่ 1 โดยแบ่งเป็นชุดข้อมูลฝึกร้อยละ 80 กับชุดข้อมูลทดลองร้อยละ 20 และชุดข้อมูลบันทึกเวชระเบียนของรหัสไอซีดีทั่วไป (Frequent ICD dataset) สำหรับใช้ในการทดลองที่ 2 โดยแบ่งเป็นชุดข้อมูลฝึกร้อยละ 80 กับชุดข้อมูลทดลองร้อยละ 20 เช่นกัน รายละเอียดได้แสดงในรูปที่ 11



รูปที่ 11 แผนภาพการแบ่งชุดข้อมูลสำหรับแต่ละการทดลอง

ซึ่งรายละเอียดของแต่ละการทดลองจะอธิบายผ่านแต่ละหัวข้อของแบบจำลองที่ใช้ในการวิจัยนี้

4.2.4.1. ตัวแปรควบคุม

- ความยาวของข้อความฝึก

เรากำหนดที่ความยาว 1000 ซึ่งเราได้ข้อสรุปจากการทดลองเปรียบเทียบที่ความยาวแตกต่างกันได้ผลตามตารางที่ 20 ซึ่งเราทดลองด้วยแบบจำลอง Bi-LSTM บนชุดข้อมูลบันทึกการรักษาที่นำมาเฉพาะบันทึกของรหัส 50 อันดับแรก เหตุผลที่เราใช้พิจารณาเลือกคือค่า Precision Recall และเวลาในการประมวลผล ในเรื่องของเวลาจะเพิ่มขึ้นตามความยาวของประโยคโดยธรรมชาติ แต่เราจะเห็นว่าความยาวที่ 1000 ให้ผลลัพธ์ที่ดีกว่าทั้งใน Precision และ Recall

ตารางที่ 20 เปรียบเทียบประสิทธิภาพของแบบจำลองที่ความยาวของข้อความต่างกัน

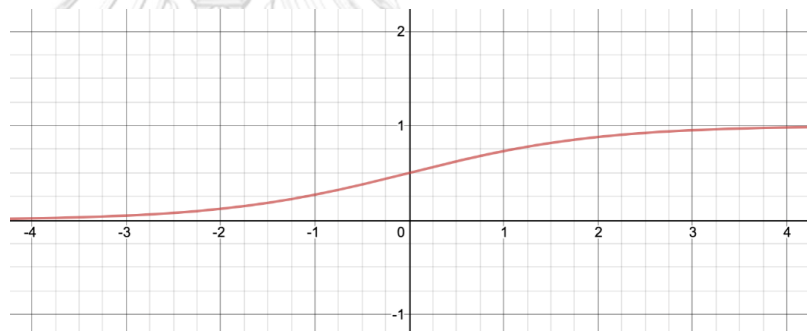
ความยาวของข้อความ	Precision	Recall
100	0.2762	0.6594
300	0.3044	0.7268
500	0.3085	0.7366
1000	0.3143	0.7506
5000	0.3129	0.7479

เราสร้างแบบจำลองการเรียนรู้เชิงลึกด้วย TensorFlow ผ่านการใช้ไลบรารี Keras API เมื่อฝึกแบบจำลองแต่ละแบบเรากำหนดใช้

- Optimizer เป็น Adaptive moment estimation (Adam optimizer) ด้วยอัตราการเรียนรู้ (Learning rate) ที่ 0.001

- Loss function ใช้ Binary cross entropy เนื่องจากแบบจำลองของเรา จำแนกประเภทแบบหลายผลาก ซึ่งจะพิจารณาที่แต่ละผลากว่าเป็นหรือไม่ เป็นรหัสไอบิตนี้จึงมีลักษณะการตัดสินใจแบบไบนารี ดังนั้นเราจึงเลือกใช้ Binary cross entropy
- Prediction layer เราใช้ Dense layer หรือ Fully connected layer ที่มี จำนวน unit เท่ากับจำนวนของรหัสไอบิตในชุดที่พิจารณา แล้วคำนวณโดยใช้ Sigmoid activation function เพื่อให้คำตอบของทุกรหัส แต่ละรหัส อยู่ในช่วง 0 ถึง 1 เท่านั้นตามสมการ (6)

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (6)$$



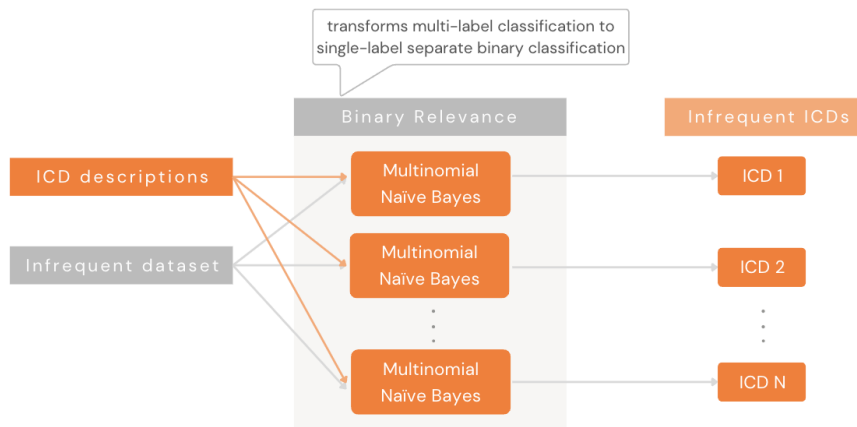
รูปที่ 12 แสดงของเขตของช่วงค่าคำตอบจากการคำนวณด้วยฟังก์ชัน sigmoid (อ้างอิงรูปจาก <https://guopai.github.io/ml-blog16.html>)

4.2.4.2. Multinomial Naïve Bayes

ในการทดลองที่ 1 ต้องการวัดประสิทธิภาพการจำแนกรหัสไอบิตกลุ่มหายากด้วยแบบจำลองทางสถิติซึ่งคือ แบบจำลอง Multinomial Naïve Bayes ที่นิยมใช้สำหรับปัญหาการจำแนกข้อความ (Text classification)

การวิจัยนี้ได้ดัดแปลงแบบจำลอง Multinomial Naïve Bayes จากไลบรารีของ Scikit-learn ร่วมกับวิธี Binary relevance จากไลบรารีของ Scikit-multilearn [36] เปลี่ยนให้เป็นการจำแนกประเภทหลายผลาก โดยแบบจำลอง

Multinomial Naïve Bayes จะจำแนกแต่ละฉลากแบบไบนารีและวิธี Binary relevance จะรวบรวมคำตอบกลับมาอยู่ในรูปแบบการจำแนกแบบหลายฉลาก รูปที่ 13 แสดงภาพโครงสร้างของแบบจำลอง Multinomial Naïve Bayes ของเรา



รูปที่ 13 แสดงภาพโครงสร้างของแบบจำลอง Multinomial Naïve Bayes

แบบจำลองนี้จำแนกชุดข้อมูลบันทึกเวชระเบียนของรหัสไอซีดีที่พบยาก (1,047 ICD-10-TM และ 2,208 ICD-9-CM) แบบจำลอง Naïve Bayes จะคำนวณค่าความน่าจะเป็นของแต่ละรหัส j (c_j) ที่กำหนดเวกเตอร์นำเข้า $[x_1, x_2, \dots, x_n]$ ตามสมการ (7)

$$P(c_j|[x_1, x_2, \dots, x_n]) \propto P(c_j) \prod_{i=1}^n P(x_i|c_j) \quad (7)$$

เมื่อ $P(x_i|c_j)$ คือความน่าจะเป็นของค่าที่ i ปรากฏในตัวอย่างข้อมูลฝึกของ c_j และ $c_j \in C$ เมื่อ C แทนรหัสไอซีดีที่พบยากทั้งหมด สำหรับการฝึกแบบจำลองนี้ เราฝึกด้วยชุดข้อมูลฝึกบันทึกเวชระเบียนของรหัสไอซีดีกลุ่มหายาก ชุดข้อมูลคำบรรยายรหัส (ICD description) และชุดคลังศัพท์เพิ่มเติมที่เกี่ยวข้องกับรหัสไอซีดีที่พิจารณา และเราหลีกเลี่ยงการเกิดความน่าจะเป็นเป็นศูนย์ในค่าที่อยู่นอกคลังศัพท์โดยใช้วิธี Laplace smoothing [37]

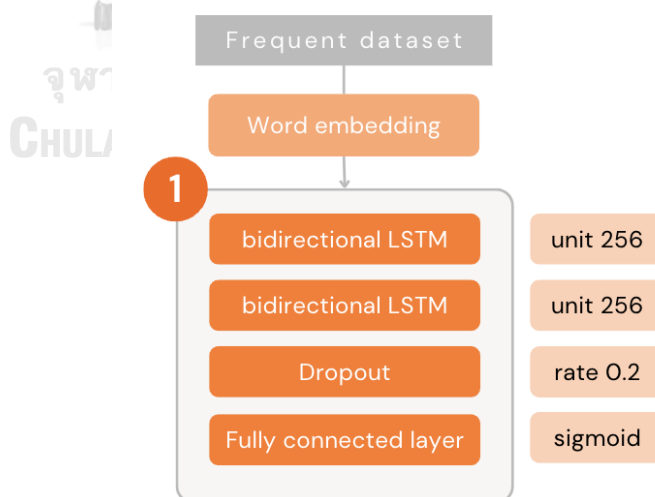
4.2.4.3. Bidirectional Long-short term memory (Bi-LSTM)

สำหรับการทดลองที่ 2 ที่วัดประสิทธิภาพการจำแนกรหัสไอซีดีกลุ่มทั่วไป ด้วยแบบจำลองเชิงลึก แบบจำลองแรกที่ใช้คือ แบบจำลอง Bidirectional Long-short term memory (Bi-LSTM)

Bi-LSTM ประกอบด้วย LSTM ที่เรียนรู้จากซ้ายไปขวา (forward) และ จากขวาไปซ้าย (backward) ส่วนของ LSTM พัฒนาต่อมาจาก Recurrent neural network เพื่อแก้ปัญหาค่าเกรเดียนต์หายไปของค่าเกรเดียนต์ (gradient vanishing) ในข้อความที่มีความยาวมาก ๆ โดยประดิษฐ์ให้มีหน่วยความจำที่คอยตัดสินใจให้ แบบจำลองจำหรือลืมข้อมูลในข้อความ

เนื่องจากเรากำหนดให้แบบจำลองมีการเรียนรู้แบบสองทิศทาง (Bidirectional) ทำให้แบบจำลองสามารถเรียนรู้ข้อมูลฝึกทั้งแบบเดินหน้าและถอยหลัง เรียนรู้ตำแหน่งลำดับและบริบทของคำไปพร้อม ๆ กันได้

โครงสร้างของแบบจำลอง Bi-LSTM ในรูปที่ 14 ประกอบด้วย Bi-LSTM layer ที่มี ขนาดยูนิต (Units) เท่ากับ 256 ทั้งหมด 2 ชั้น ต่อเข้ากับ Dropout layer ด้วย อัตราลดเท่ากับ 0.2 ตามด้วย fully connected layer เป็นชั้นสุดท้ายสำหรับ ทำนายรหัสไอซีดี



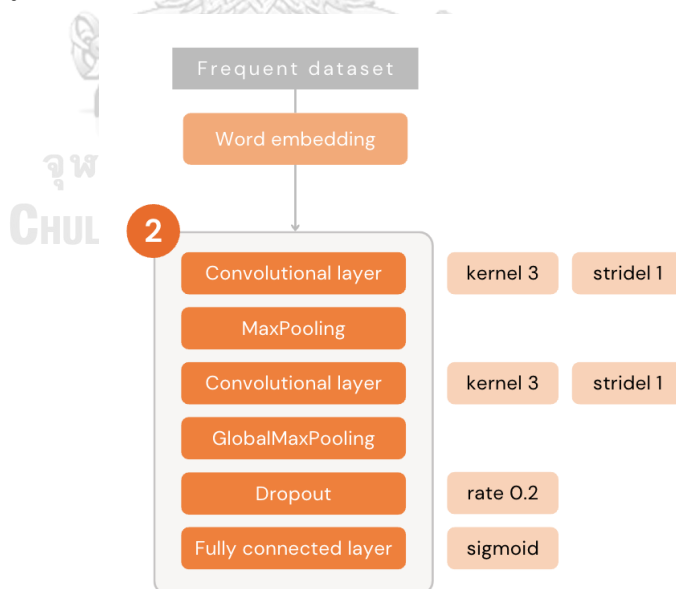
รูปที่ 14 โครงสร้างของแบบจำลอง Bidirectional LSTM

4.2.4.4. Convolutional neural network (CNN)

แบบจำลองที่ 2 ที่ใช้สำหรับวัดประสิทธิภาพการจำแนกรหัสไอซีดีกลุ่ม
ทั่วไปคือ แบบจำลอง

Convolutional neural network

แบบจำลองนี้ประกอบด้วย 6 layers ดังรูปที่ 15 หลังจากที่ผ่านมาข้อความ
เข้า embedding layer จนได้เมทริกซ์ฝังคำ (Embedded matrix) เมทริกซ์นี้จะ
เข้าสู่ชั้นที่1 คือ 1D-Convolutional layer ที่มีขนาดของ Kernel เท่ากับ 2 ค่า
strides เท่ากับ 1 ขนาดเอาต์พุตเท่ากับ 300 และใช้ tanh activation function
ตามด้วยชั้นที่2 คือ Max pooling layer จากนั้นชั้นที่3 เป็น 1D-Convolutional
layer ที่มีขนาดของ Kernel เท่ากับ 2 ค่าstrides เท่ากับ 1 ขนาดเอาต์พุตเท่ากับ
256 และใช้ tanh activation function ตามด้วยชั้นที่ 4 เป็น Global Max
Pooling และชั้นที่5 Dropout layer ด้วยอัตราลดเท่ากับ 0.2 สุดท้ายเข้า fully
connected layer ด้วย sigmoid activation function ซึ่งให้เวกเตอร์คำตอบที่มี
ค่าอยู่ในช่วง [0-1] และมีขนาดมิติเท่ากับจำนวนของรหัสไอซีดีทั่วไป (พบบ่อย)



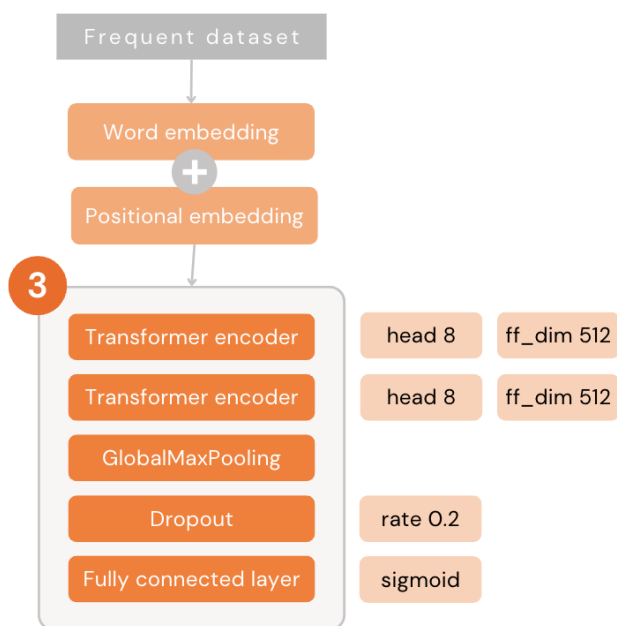
รูปที่ 15 โครงสร้างของแบบจำลอง Convolutional neural network

4.2.4.5. Transformer encoder

BERT เป็นแบบจำลองที่นิยมสำหรับงานที่เกี่ยวข้องกับการจำแนกข้อความด้วยขนาดที่ใหญ่และมีความซับซ้อนมาก BERT จึงทำงานได้อย่างมีประสิทธิภาพ แต่เนื่องจากข้อมูลของเรามีจำนวนและความหลากหลายไม่มากเมื่อเทียบกับขนาดข้อมูลอื่นที่ใช้ในการฝึก BERT และเรามีวัตถุประสงค์ที่จะลดการใช้ resource และเวลาในการดำเนินการ (Execution time) เราจึงเลือกใช้องค์ประกอบภายในของ BERT นั่นคือ Transformer encoder เราจึงเลือกแบบจำลองนี้เป็นแบบจำลองสุดท้ายสำหรับใช้จำแนกรหัสไอซีดีกลุ่มทั่วไป

Transformer encoder ประกอบด้วย Multi-head attention และ feed-forward neural network [38] ซึ่ง Multi-head attention เป็น self-attention layer ที่พยายามทำความเข้าใจความสัมพันธ์ของคำที่พิจารณาอยู่ตอนนี้ต่อคำข้างเคียงอื่น ๆ ภายในประโยคเดียวกันและจำนวน head ที่ต่างกันจะระบุจำนวนรูปแบบจุดโฟกัสที่แตกต่างกันด้วย ทั้งนี้ Transformer ยังเพิ่มการเรียนรู้รายละเอียดของตำแหน่งคำในประโยคโดยการเพิ่มการฝังตำแหน่งคำ (Positional embedding) ที่จะเป็นเวกเตอร์ขนาดเท่ากับเวกเตอร์ฝังคำ (Word embedding) เราจึงนำทั้งสองเวกเตอร์มาบวกกันแบบ dot product ได้เป็นเวกเตอร์สุทธิตั้งหน้าเข้าไปฝึกในแบบจำลอง Transformer encoder ต่อไป

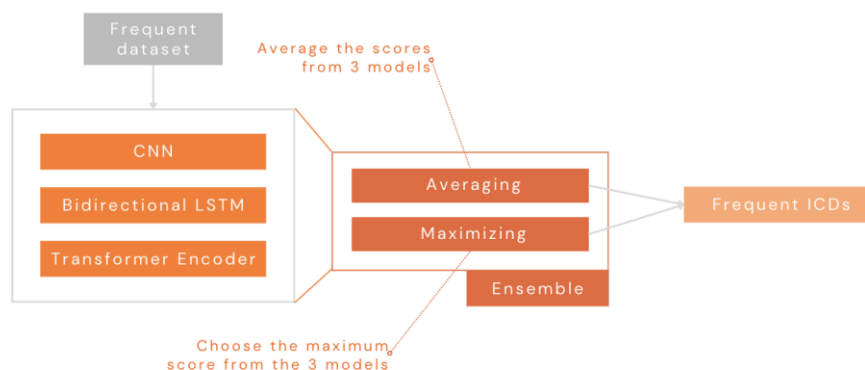
โครงสร้างแบบจำลองของเราประกอบด้วย Transformer encoder 2 ชั้นต่อกัน แต่ละชั้นกำหนดค่า intermediate dimension เท่ากับ 256 และใช้ 4 Multi-head attention ตามด้วย Global Max Pooling และ Dropout layer ด้วยอัตราลดเท่ากับ 0.2 ดังแสดงในรูปที่ 16



รูปที่ 16 โครงสร้างของแบบจำลอง Transformer encoder

4.2.4.6. Ensemble method

วิธีการรวม (Ensemble method) จะรวมชุดรหัสที่ทำนายจากทั้ง 3 แบบจำลองการเรียนรู้เชิงลึกด้วยการเลือกค่าทำนายตามรูปแบบการตัดสินใจ ในงานของเราเปรียบเทียบประสิทธิภาพของการรวมแบบจำลองระหว่างด้วยวิธีการเลือกค่ามากที่สุดกับวิธีการเฉลี่ยวิธีการรวมด้วยการตัดสินใจเลือกค่าความมั่นใจสูงสุด (Maximizing ensemble method) คือเราเลือกค่าทำนายที่มากที่สุดจากทั้ง 3 แบบจำลองจะถูกเลือกเป็นค่าที่ใช้ตัดสินใจในแต่ละรหัสไอซีดี ส่วนวิธีการเฉลี่ย (Averaging ensemble method) คือการนำค่าทำนายจากทั้ง 3 แบบจำลองมาเฉลี่ยกันเป็นค่ากลางในการตัดสินใจแต่ละรหัสไอซีดี โครงสร้างของแบบจำลองแสดงอยู่ในรูปที่ 17



รูปที่ 17 โครงสร้างของการรวมแบบจำลองด้วยวิธีการ Ensemble

4.3. การปรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter tuning)

เพื่อให้ได้แบบจำลองที่เข้ากับชุดข้อมูลของเรามากที่สุด เราจึงต้องปรับค่าไฮเปอร์พารามิเตอร์ซึ่งเป็นตัวแปรที่เราสามารถกำหนดเองได้ การปรับค่าไฮเปอร์พารามิเตอร์เป็นกระบวนการหาค่าของไฮเปอร์พารามิเตอร์ (Optimization) และเลือกชุดตัวแปรที่เหมาะสมที่สุดกับแบบจำลองของเรา

เราเลือกใช้ Hyperband Tuner [39] จากไลบรารี Keras Tuner [40] ในการปรับหาค่าไฮเปอร์พารามิเตอร์ จุดเด่นของ Hyperband คือมีการปรับปรุงเกี่ยวกับ Resource allocation และ early-stopping จากวิธีการปรับค่าตัวแปรแบบ RandomSearch โดยหลักการคือสุ่มประมวลผล Configuration ของไฮเปอร์พารามิเตอร์เป็นรอบ ๆ โดยในรอบถัด ๆ ไปจะเลือกหยาบเฉพาะ configuration ที่ให้ค่าประสิทธิภาพดีมาประมวลผลปรับค่าต่อไป วิธีการนี้ช่วยลดเวลาในการประมวลผลจากที่ต้องประมวลเท่าจำนวน epoch ที่กำหนดทีเดียวในทุกชุดตัวแปรก็เลือกแบ่งช่วง epoch ในการปรับแต่ละรอบการฝึกก็เลือกชุดตัวแปรที่ได้ออกมาฝึกต่อจาก epoch เดิมที่ค้างไว้ต่อไป ตารางที่ 22 ถึง 24 แสดงผลการปรับค่าไฮเปอร์พารามิเตอร์ที่ดีที่สุด 4 ตำแหน่งของแบบจำลอง CNN Bi-LSTM และ Transformer encoder ตามลำดับ

ตารางที่ 21 ชุดค่าไฮเปอร์พารามิเตอร์ของ CNN หลังการปรับ

#	#Conv	Filters		Kernel	Strides	Last pooling	Activ ¹	F1-score	
1	2	300	256	2	1	Global Max	tanh	0.7821	
2	3	300	256	256	2	1	Global Max	tanh	0.7149
3	2	300	256		3	1	Global Max	tanh	0.7066
4	3	300	256	256	3	1	Global Max	tanh	0.4975

¹Activation function

ตารางที่ 22 ชุดค่าไฮเปอร์พารามิเตอร์ของ Bi-LSTM หลังการปรับ

#	#Units		Dropout	Drop rate	F1-score
	Forward	Backward			
1	256	256	True	0.2	0.4047
2	256	256	True	0.15	0.3601
3	128	128	True	0.2	0.3507
4	128	128	True	0.15	0.3313

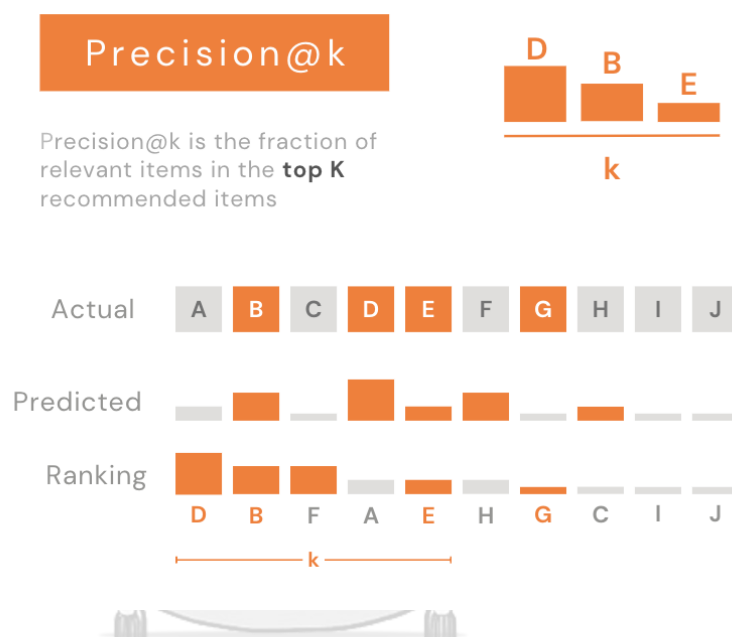
ตารางที่ 23 ชุดค่าไฮเปอร์พารามิเตอร์ของ Transformer encoder หลังการปรับ

#	#Encoder	#Head	#FF_dim	Pooling	Dropout rate	F1
1	2	8	512	Global Max	0.1	0.3311
2	2	4	256	Global Max	0.1	0.3167
3	2	4	512	Global Max	0.2	0.3127
4	2	4	512	Global Max	0.1	0.3107

4.4. การวัดผลแบบจำลอง (Evaluation)

เราเลือกวัดผลแบบจำลองด้วยค่า Precision@k (รูปที่ 18) และ Jaccard index (รูปที่ 19) บนชุดข้อมูลทดสอบเดียวกัน

ค่า Precision คือค่าสัดส่วนของผลเฉลยในค่าทำนาย k อันดับแรก ซึ่งสามารถอธิบายได้จากคำตอบทำนายรหัสไอซีดีที่มั่นใจที่สุด K อันดับแรกนั้นตอบถูกต้องจำนวนเท่าไรโดยในที่นี้เราเลือกพิจารณาที่อันดับ 3 5 และ 8

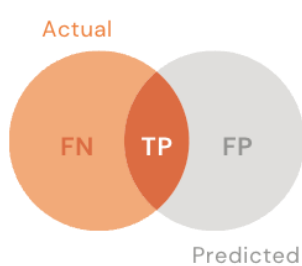


รูปที่ 18 ตัวอย่างการคำนวณของค่า Precision@k

ค่า Jaccard index มาจากการคำนวณจำนวนรหัสไอซีดีที่ทำนายถูกหารด้วยจำนวนรหัสไอซีดีที่ได้จากการทำนายและจากผลเฉลยจริง ซึ่งค่านี้เปรียบเหมือนนำวงคำตอบของเรามาทับซ้อนกับวงของผลเฉลยดังแสดงในรูปที่ 17 ยิ่งค่า Jaccard index เข้าใกล้ 1 มากนั้นหมายความว่าคำตอบของเรานั้นตอบเหมือนกับผลเฉลยมาก

Jaccard index

Jaccard index is the size of the **intersection** of the predicted labels and the true labels divided by the size of the union of the predicted and true labels



รูปที่ 19 ภาพประกอบการอธิบายการซ้อนทับของคำตอบ



5. ผลการทดลอง

บทนี้จะรายงานผลการปรับค่าไฮเปอร์พารามิเตอร์และค่าผลการทดลองจากทั้ง 3 การทดลองของชุดข้อมูลการวินิจฉัยโรคและชุดข้อมูลเหตุการณ์ตามลำดับ

การวิจัยแบ่งการทดลองออกเป็น 3 การทดลอง การทดลองที่ 1 วัดประสิทธิภาพการจำแนกรหัสไอซีดีกลุ่มหายากด้วยแบบจำลอง Naïve Bayes การทดลองที่ 2 วัดประสิทธิภาพการจำแนกรหัสไอซีดีกลุ่มพบได้ทั่วไปด้วยแบบจำลองการเรียนรู้เชิงลึก (Deep learning) และการทดลองที่ 3 วัดประสิทธิภาพการจำแนกรหัสไอซีดีทั้ง 2 กลุ่มด้วยแบบจำลองเชิงผสมระหว่างแบบจำลอง Naïve Bayes และแบบจำลองเชิงลึก ดังนั้นการทดลองทั้งสามจึงใช้ชุดข้อมูลทดสอบที่ต่างกัน

5.1. ผลการทดลองแบบจำลองเชิงสถิติ (Experiment 1)

การทดลองที่ 1 ใช้ตัวจำแนกรหัสไอซีดีด้วยแบบจำลองทางสถิติคือ Multinomial Naïve Bayes สำหรับจำแนกรหัสไอซีดีกลุ่มหายาก (Rare codes) ที่มีจำนวนตัวอย่างของแต่ละรหัสไม่ถึง 100 ตัวอย่าง ซึ่งแบ่งใช้สำหรับฝึกร้อยละ 80 และใช้สำหรับทดสอบร้อยละ 20 ดังนั้นผลการทดลองที่แสดงในตารางต่อไปนี้จะทำนายผลเฉพาะรหัสไอซีดีกลุ่มหายากเท่านั้น

5.1.1. ผลการทดลองบนชุดข้อมูลการวินิจฉัยโรค (Course notes)

ผลการทดลองส่วนของแบบจำลองทางสถิติหรือ Multinomial Naïve Bayes ในตารางที่ 24 พบว่าค่า Precision@k ค่อยข้างต่ำนั้นคือในช่วง 3 ถึง 8 รหัสแรกที่ทำนายมากไม่ค่อยตรงกับผลเฉลยจริง แต่ค่า Jaccard index แสดงว่าการทำนายรหัสไอซีดีของแบบจำลองนี้มีความถูกต้อง 20%

ตารางที่ 24 ผลการทดลองแบบจำลองทางสถิติด้วยชุดข้อมูลการวินิจฉัยโรค

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
MNB ¹	0.092	0.057	0.037	0.205

¹Multinomial Naïve Bayes

5.1.2. ผลการทดลองบนชุดข้อมูลการให้ยาและหัตถการ (Medication notes)

จากการทดลองทำนายรหัส ICD-9-CM ด้วยแบบจำลอง Multinomial Naïve Bayes ในตารางที่ 25 วัดผลค่า Jaccard index อธิบายสัดส่วนรหัสไอซีดีที่ทำนายถูกต้องเทียบกับจำนวนรหัสไอซีดีจากการทำนายและผลเฉลี่ยได้ผลเป็น 20.14% และผลค่า Precision@k ที่ k เป็น 3 5 และ 8 ไม่สามารถทำนายรหัสไอซีดีที่ถูกต้องได้ในช่วง 8 อันดับแรก

ตารางที่ 25 ผลการทดลองแบบจำลองทางสถิติด้วยชุดข้อมูลการให้ยาและหัตถการ

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
MNB ¹	0.1097	0.0680	0.0430	0.2014

¹Multinomial Naïve Bayes

5.2. ผลการทดลองแบบจำลองเชิงลึก (Experiment 2)

การทดลองที่ 2 ใช้ตัวจำแนกรหัสไอซีดีด้วยแบบจำลองเชิงลึก ได้แก่ Bidirectional LSTM, Convolutional neural network และ Transformer encoder สำหรับจำแนกรหัสไอซีดีกลุ่มทั่วไป (Common codes) ที่มีจำนวนตัวอย่างของแต่ละรหัสตั้งแต่ 100 ตัวอย่างขึ้นไป และเพิ่ม Ensemble method สำหรับเลือกคำตอบสุดท้ายจากแบบจำลองทั้งสามด้วยวิธีเฉลี่ยและวิธีเลือกค่ามากที่สุด ซึ่งแบ่งชุดข้อมูลใช้สำหรับฝึกร้อยละ 80 และใช้สำหรับทดสอบร้อยละ 20 ดังนั้นผลการทดลองที่แสดงในตารางต่อไปนี้จะทำนายผลเฉพาะรหัสไอซีดีกลุ่มทั่วไปเท่านั้น

5.2.1. ผลการทดลองบนชุดข้อมูลการวินิจฉัยโรค (Course notes)

สำหรับผลการทดลองในส่วนของแบบจำลองการเรียนรู้เชิงลึกทั้ง 3 บนชุดข้อมูลของรหัสไอซีดีทั่วไปในตารางที่ 26 นั้นให้ผลลัพธ์ค่า Precision@k ได้ดี จากการทำนาย 3 อันดับแรก (P@3) ทำนายได้ถูกต้องโดยเฉลี่ย 2 รหัสอย่างแน่นอนและมีโอกาสทำนายถูกเพิ่มขึ้นอีก 1 รหัสในช่วงไม่เกิน 8 รหัสแรก (P@8) แบบจำลอง Bi-LSTM ให้ประสิทธิภาพในการทำนายรหัส ICD-10-TM มากกว่าแบบจำลอง CNN และ Transformer encoder ในทุกตัววัดผล โดยเฉพาะ Jaccard index ที่มีความถูกต้องถึง 79.2% ซึ่งมากกว่า CNN และ Transformer encoder อยู่ 15% และ

17.2% ตามลำดับ ด้านแบบจำลอง Ensemble method จะเห็นว่าการรวมด้วยวิธีเลือกค่ามากที่สุดจะให้ผลลัพธ์การทำนายที่ใกล้เคียงกับผลเฉลี่ยมากกว่า 6.7% แต่หากวัดการทำนายรหัสได้ถูกต้องในช่วง 8 อันดับแรก (P@8) ทั้ง 2 แบบมีประสิทธิภาพใกล้เคียงกันซึ่งจะเห็นว่าค่าใกล้เคียงกับ Bi-LSTM ด้วยนั้นอาจเป็นไปได้ว่าการทำนายที่ถูกต้องในอันดับแรก ๆ มาจากการทำนายของแบบจำลอง Bi-LSTM อย่างไรก็ตามความถูกต้องของการทำนายรหัสไอซีตีทั่วไปด้วยแบบจำลองแบบวิธีรวมก็ยิ่งน้อยกว่าการทำนายด้วยแบบจำลองเดี่ยวของ Bi-LSTM ที่ค่า Jaccard index อยู่ที่ 3.2%

ตารางที่ 26 ผลการทดลองแบบจำลองการเรียนรู้เชิงลึกด้วยชุดข้อมูลการวินิจฉัยโรค

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
CNN	0.732	0.575	0.413	0.642
Bi-LSTM	0.813	0.654	0.477	0.792
Transformer	0.750	0.595	0.432	0.620
AvgEnsemble ²	0.812	0.654	0.478	0.693
MaxEnsemble ³	0.813	0.655	0.478	0.760

²Average Ensemble of CNN, Bi-LSTM and Transformer encoder

³Maxmizing Ensemble of CNN, Bi-LSTM and Transformer encoder

CHULALONGKORN UNIVERSITY

5.2.2. ผลการทดลองบนชุดข้อมูลการให้ยาและหัตถการ (Medication notes)

สำหรับผลการทดลองในส่วนของแบบจำลองการเรียนรู้เชิงลึกในตารางที่ 27 นั้น แบบจำลอง Bi-LSTM มีประสิทธิภาพมากกว่าแบบจำลองเดี่ยวอื่นในทุกตัววัดผล ค่า Precision@k ของแบบจำลองเชิงลึกมีแนวโน้มความถูกต้องในการทำนายรหัสไอซีตีที่คล้ายและให้ค่าใกล้เคียงกันคือ ที่ k เป็น 3 และ 5 มีแนวโน้มทำนายถูกต้องอย่างน้อย 1 และ 2 รหัสตามลำดับ สำหรับแบบจำลองเชิงลึกร่วมกับวิธีรวมด้วยการเฉลี่ย (AvgEnsemble) ให้ค่า Jaccard index ที่ 96.26% ซึ่งสูงกว่าแบบวิธีรวมด้วยการเลือกค่าสูงสุด (MaxEnsemble) อยู่ 3% แต่มีค่า Precision@k ใกล้เคียงกันและมีแนวโน้มคล้ายกันที่ k แต่ละช่วงซึ่งใกล้เคียงกับแบบจำลองเชิงลึกทั้ง 3 แบบเช่นกัน

ตารางที่ 27 ผลการทดลองแบบจำลองการเรียนรู้เชิงลึกด้วยชุดข้อมูลการให้ยาและหัตถการ

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
CNN	0.6020	0.4214	0.2836	0.8850
Bi-LSTM	0.6075	0.4259	0.2869	0.9364
Transformer	0.5896	0.4104	0.2760	0.8161
AvgEnsemble ²	0.6082	0.4272	0.2886	0.9626
MaxEnsemble ³	0.6080	0.4270	0.2884	0.9304

²Average Ensemble of CNN, Bi-LSTM and Transformer encoder

³Maxmizing Ensemble of CNN, Bi-LSTM and Transformer encoder

5.3. ผลการทดลองแบบจำลองผสมผสาน (Experiment 3)

การทดลองสุดท้ายใช้ตัวจำแนกรหัสไอซีดีด้วยแบบผสมผสานโดยใช้แบบจำลอง Multinomial Naïve Bayes สำหรับจำแนกรหัสไอซีดีกลุ่มพวยาก (Rare codes) และใช้แบบจำลองเชิงลึกทั้ง 3 แบบจำลองสำหรับจำแนกรหัสไอซีดีกลุ่มทั่วไป (Common codes) ซึ่งแบ่งชุดข้อมูลใช้สำหรับฝึกร้อยละ 80 และใช้สำหรับทดสอบร้อยละ 20 ดังนั้นผลการทดลองที่แสดงในตารางต่อไปนี้จะทำนายผลเฉพาะรหัสไอซีดีทั้งสองกลุ่ม

5.3.1. ผลการทดลองบนชุดข้อมูลการวินิจฉัยโรค (Course notes)

การทดลองสุดท้ายเราทำการต่อแบบจำลองทั้ง 2 เข้าด้วยกันในรูปแบบต่าง ๆ ผลการทดลองในตารางที่ 28 พบว่าแบบจำลอง Maximizing Ensemble method ให้ค่า Jaccard index มากที่สุดเท่ากับ 13.8% แต่ค่า Precision@k กลับน้อยกว่าแบบจำลอง CNN ใน 5 อันดับแรกและขึ้นมาเท่ากับในช่วง 8 แรก หากนับเป็นจำนวนรหัสจริงทั้งสองแบบจำลองก็ทำนายถูกอย่างน้อย 1 รหัสเท่ากับในช่วง 5 อันดับแรก (P@5)

ตารางที่ 28 ผลการทดลองแบบจำลองผสมผสานด้วยชุดข้อมูลการวินิจฉัยโรค

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
CNN + MNB ⁴	0.284	0.221	0.169	0.111
Bi-LSTM + MNB	0.245	0.182	0.136	0.113
Transformer + MNB	0.245	0.190	0.146	0.110
AvgEnsemble+MNB	0.186	0.145	0.113	0.104
MaxEnsemble+MNB	0.281	0.218	0.169	0.138

⁴Combined model between Deep learning model and MNB

5.3.2. ผลการทดลองบนชุดข้อมูลการให้ยาและหัตถการ (Medication notes)

สำหรับแบบจำลองที่รวมกันเพื่อทำนายรหัสไอซีดีทั้งสองกลุ่ม ผลการทดลองในตารางที่ 29 ที่ได้คือแบบจำลอง Maximizing ensemble method ให้ประสิทธิภาพที่สุดบนค่า Jaccard index รองลงมาคือแบบจำลอง Bi-LSTM สำหรับ Precision@k แบบจำลอง Bi-LSTM ทำนายได้ดีกว่าแบบจำลองอื่นเฉลี่ย 1.5% ในช่วงทำนายไม่เกิน 5 อันดับแรกจนมาช่วง 8 อันดับแรกที่แบบจำลอง Averaging ensemble method ทำนายได้ดีกว่า อย่างไรก็ตามประสิทธิภาพของแบบจำลองต่อกันนี้ก็ได้มีโอกาสการทำนายรหัสให้ถูกต้องอย่าง 1 รหัสจากการทำนาย 8 อันดับแรกเลยสักแบบจำลอง

ตารางที่ 29 ผลการทดลองแบบจำลองผสมผสานด้วยชุดข้อมูลการให้ยาและหัตถการ

แบบจำลอง	Precision@k			Jaccard index
	P@3	P@5	P@8	
CNN + MNB ⁴	0.1127	0.0763	0.0543	0.0701
Bi-LSTM + MNB	0.1343	0.0914	0.0637	0.0904
Transformer + MNB	0.1187	0.0823	0.0597	0.0838
AvgEnsemble+MNB	0.1298	0.0910	0.0666	0.0735
MaxEnsemble+MNB	0.1297	0.0909	0.0663	0.0942

⁴Combined model between Deep learning model and MNB

6. อภิปรายและสรุปผล

6.1. การอภิปรายผล (Discussion)

6.1.1. ความสัมพันธ์ของรหัสแนะนำ (Recommended ICD relationship)

ส่วนนี้เราจะแสดงตัวอย่างของผลการทำงานของแบบจำลองของเรา โดยแบบจำลองของเรานำเสนอรหัสไอซีดีที่เกี่ยวข้องกับกลุ่มอาการและโรคที่รับมามากที่สุด ในที่นี้เราแนะนำรหัสไอซีดีที่แบบจำลองเสนอมา 10 อันดับแรก (predicted codes) เทียบกับที่ผู้เข้ารหัสลง (actual codes)

จากตารางที่ 30 ตัวอย่างเป็นกลุ่มอาการโรคหืดที่เป็นโรคเกี่ยวกับระบบทางเดินหายใจ จะเห็นว่าแบบจำลองของเราแนะนำรหัสที่เกี่ยวข้องกับระบบทางเดินหายใจในมาในอันดับ 1 และ 2 และนำเสนอรหัสที่ถูกต้องในอันดับที่ 7 สำหรับกลุ่มโรคที่เกี่ยวข้องกับความดันโลหิตสูง โรคเบาหวาน หรือโรคที่เกี่ยวข้องกับระบบการเผาผลาญไขมันนั้นจะเป็นกลุ่มโรคที่มีอาการอ่อนเพลีย เหนื่อยง่าย หายใจหอบ ซึ่งคล้ายคลึงกับอาการของโรคหืดเช่นกัน โรคที่เกี่ยวกับสมดุลกรด-เบส หรือของเหลวในร่างกายผู้ป่วยจะมีอาการเหนื่อยหอบง่ายเช่นกัน ซึ่งจะชัดเจนในผู้ป่วยที่ของเหลวไม่สมดุลอาจนำไปสู่ภาวะน้ำท่วมปอดหรืออวัยวะอื่น ๆ ด้านกลุ่มโรคไตเรื้อรังหรือปัสสาวะไม่ออกก็มีอาการเหนื่อยง่าย อ่อนเพลีย ในกลุ่มโรคหัวใจ ในเด็ก เด็กก็มีการเหนื่อยหอบ หายใจสั้น (shortness of breath) คล้ายกับอาการของโรคหืดด้วย

สำหรับตารางที่ 31 ตัวอย่างคือกลุ่มอาการของโรคกรดไหลย้อน ซึ่งผู้ป่วยจะมีอาการหลัก ๆ ตามที่ได้บรรยายไว้ในตาราง จะเห็นว่าแบบจำลองของเราแนะนำโรคกรดไหลย้อนอยู่ที่อันดับที่ 6 ในส่วนของอันดับที่ 1 นั้นคือโรคหอบหืดซึ่งเป็นโรคที่พบมากเช่นกันและมีส่วนเกี่ยวข้องกับช่องอก (chest) จึงมีโอกาสที่แบบจำลองจะแนะนำโรคนี้อันดับแรกได้ สำหรับกลุ่มโรคตับและทางเดินอาหารที่แบบจำลองแนะนำนั้นเป็นโรคที่เกี่ยวข้องกับการกิน (eating) หรือผลจากการกิน สำหรับโรครูมาตอยด์นั้น เนื่องจากอาการที่บรรยายมีการแสดงอาการปวดตอนกลางคืนหรือตอนนอนซึ่งคล้ายคลึงกับอาการของโรครูมาตอยด์

ตารางที่ 30 ตัวอย่างการให้รหัสไอซีดีจากแบบจำลองของเรา (1)

A persistent cough, especially at night wheezing when exhaling and sometimes when inhaling shortness of breath or difficulty breathing chest tightness, making it difficult to breathe deeply	
รหัสไอซีดีจากแบบจำลองของเรา	รหัสไอซีดีที่ผู้เข้ารหัสลง
<ol style="list-style-type: none"> 1. R06 Abnormalities of breathing 2. J98 Other respiratory disorders 3. I10 Essential (primary) hypertension 4. E78 Disorders of lipoprotein metabolism and other lipidaemia 5. Q25 Congenital malformations of great arteries 6. E11 Type 2 diabetes mellitus 7. J45 Asthma 8. E87 Other disorders of fluid, electrolyte and acid-base balance 9. N18 Chronic kidney disease 10. R33 Retention of urine 	<ol style="list-style-type: none"> 1. J45 Asthma

ตารางที่ 31 ตัวอย่างการให้รหัสไอซีดีจากแบบจำลองของเรา (2)

<p>แสบร้อนลิ้นปี่ เรอเปรี้ยว burning sensation in chest, usually after eating, which might be worse at night or while lying down</p>	
<p>รหัสไอซีดีจากแบบจำลองของเรา</p>	<p>รหัสไอซีดีที่ผู้เข้ารหัสลง</p>
<ol style="list-style-type: none"> 1. J45 Asthma 2. K75 Other inflammatory liver diseases 3. Y83 Surgical operation and other surgical procedures as the cause of abnormal reaction of the patient 4. K70 Alcoholic liver disease 5. K62 Other diseases of anus and rectum 6. K21 Gastro-oesophageal reflux disease 7. K76 Other diseases of liver 8. K91 Postprocedural disorders of digestive system, not elsewhere classified 9. M06 Rheumatoid nodule 10. Z11 Special screening examination for infectious and parasitic diseases 	<ol style="list-style-type: none"> 1. K21 Gastro-oesophageal reflux disease

จากการที่แบบจำลองของเราสามารถนำเสนอโรคที่มีกลุ่มอาการคล้ายคลึงหรือใกล้เคียงกันจะเป็นการช่วยแพทย์วินิจฉัยโรคในเบื้องต้นได้ (differential diagnosis) โดยนำเสนอโรคที่ผู้ป่วยอาจจะผ่านกลุ่มอาการที่ได้รับมา ช่วยเพิ่มประสิทธิภาพการวินิจฉัยให้ครอบคลุมยิ่งขึ้น

6.1.2. ความแตกต่างระหว่างให้รหัสกับวินิจฉัยโรคในเบื้องต้นได้ผ่านการนำเสนอรหัสไอซีดี

จากการอภิปรายความสัมพันธ์ของรหัสแนะนำก่อนหน้า งานวิจัยนี้สามารถแนะนำการให้รหัสไอซีดีที่ใกล้เคียงกับกลุ่มอาการที่ได้รับมาได้ แต่สมบัตินี้แตกต่างจากงานวิจัยสำหรับวินิจฉัยโรคในเบื้องต้นได้ (differential diagnosis) เนื่องจากลักษณะข้อมูลสำหรับฝึกของงานวิจัยเราที่เน้นการให้รหัสไอซีดีนั้นเป็นการรวบรวมรายละเอียดอาการตลอดการพักรักษาตัวและรวมไปถึงผลการวินิจฉัยของแพทย์ ซึ่งไปผ่านการวิเคราะห์อาการและระบุวิธีการรักษาโดยแพทย์เป็นที่เรียบร้อยแล้ว ซึ่งต่างกับชุดข้อมูลสำหรับงานวิจัยการวินิจฉัยโรคในเบื้องต้นได้ที่มีลักษณะอธิบายอาการไม่เจาะจงซึ่งมาจากการที่พยาบาลสอบถามอาการเบื้องต้นก่อนพบแพทย์

ด้วยลักษณะชุดข้อมูลที่แตกต่างกันนี้ ทำให้ไม่สามารถเปรียบเทียบงานวิจัยนี้กับการวิจัยก่อนหน้าได้อย่างตรงไปตรงมา เราเพียงแค่นำเสนอความสัมพันธ์ของกลุ่มอาการกับรหัสไอซีดีที่แนะนำ

ทั้งนี้เราได้ทำการเปรียบเทียบความสามารถในการให้รหัสจากข้อมูลสรุปการวินิจฉัยโรคที่ระบุโรคหลัก (Principal diagnosis) โรคร่วม (Comorbidity) และโรคแทรก (Complication) ที่ได้รับการให้รหัสจากผู้ให้รหัสและผ่านการตรวจสอบความถูกต้องจากทีมผู้ตรวจสอบเรียบร้อยแล้วจำนวน 35 ตัวอย่างซึ่งมีจำนวนรหัสไอซีดีโดยเฉลี่ย 3.9 รหัสต่อบันทึก แบบจำลองของเราให้รหัสไอซีดีถูกต้อง 88 รหัสจากทั้งหมด 137 รหัส มีความถูกต้อง 2.5 รหัสต่อบันทึกและคิดเป็นร้อยละ 64.2

6.1.3. เปรียบเทียบการทำงานกับ ChatGPT (Comparison with ChatGPT)

ระหว่างการวิจัยของเรา ChatGPT กลายมาเป็นบริการการตอบโต้อัตโนมัติ (Automated chat service) ที่เป็นที่รู้จักและได้รับความสนใจอย่างมากเพราะความสามารถด้านการสื่อสารเสมือนมนุษย์บนหัวข้อที่หลากหลาย ไม่ว่าจะเป็นด้านปรัชญา ด้านดนตรี หรือค้นหาและเสนอคำตอบจากคำถามที่ผู้ใช้งานระบุไว้ เราจึงสำรวจความสามารถของ ChatGPT เกี่ยวกับรหัสไอซีดี ผลคือ ChatGPT รู้จักและสามารถอธิบายความหมายของรหัสไอซีดีได้อย่างถูกต้อง รวมไปถึงความแตกต่างระหว่าง ICD-10 และ ICD-10-TM ได้แม้ว่าจะไม่สามารถอธิบายถึงรายละเอียดได้ก็ตาม เราได้ทำการทดสอบเพิ่มเติมโดยให้ข้อความบันทึกเวชระเบียนและให้ ChatGPT ตอบรหัสไอซีดีที่เกี่ยวข้องกลับมา พบว่า ChatGPT สามารถให้รหัสไอซีดีที่ถูกต้องได้สำหรับข้อความบันทึกเวชระเบียนที่มีเนื้อหาไม่ซับซ้อน มีคำสำคัญที่ชัดเจน และมีโรคที่เกี่ยวข้องไม่เกิน 2 โรค สำหรับข้อความบันทึกเวชระเบียนที่มีเนื้อหาซับซ้อน อาการที่หลากหลาย และเกี่ยวข้องกับโรคมากกว่า 2 โรค ChatGPT เลือกที่จะแนะนำรหัสเป็นช่วงกว้าง เช่น E00-E07 หรือแสดงรหัสที่คิดว่าเกี่ยวข้องออกมาเสนอผู้ใช้งานทั้งหมด ตัวอย่างเช่น บันทึกการรักษาที่มีคำว่า *Neoplasm* แล้วให้คำตอบมาเป็นรหัส C71, D32, D35, D49 เป็นต้น

ในทางกลับกัน งานของเราเน้นเป็นการแนะนำรหัสไอซีดีที่เกี่ยวข้องให้ได้อย่างถูกต้องโดยแสดงให้เห็นจากการที่แบบจำลองของเราไม่แนะนำรหัสไอซีดีที่มีจำนวนตัวอย่างฝึกมากที่สุดออกมาในทุก ๆ การทำนาย นอกจากนี้แบบจำลองของเรายังรองรับบันทึกเวชระเบียนที่มีความหลากหลายของอาการและโรค รวมถึงภาษาที่ใช้บันทึกทั้งไทยและอังกฤษ

6.1.4. วิเคราะห์ต้นทุน (Cost Analysis)

เมื่อเปรียบเทียบระยะเวลาและปริมาณทรัพยากรที่ใช้ต่อการฝึก 1 ครั้ง ระหว่างแบบจำลอง BERT กับแบบจำลองของเรานั้น แบบจำลอง BERT ใช้ปริมาณ ทรัพยากรเนื่องจากขนาดของแบบจำลอง BERT ที่มีจำนวนตัวแปร (parameters) มากกว่าแบบจำลองของเรา ในด้านเวลาในการปรับ (tuning) แบบจำลอง BERT จะใช้เวลานานกว่า ซึ่งแบบจำลองของเรานั้นเป็นการแบ่งฝึกแบบจำลองออกเป็น CNN Bi-LSTM และ Transformer ก่อนจะมารวมกันทำใช้ลดระยะเวลาในการฝึก และขนาดของแบบจำลองที่เล็กกว่าทำให้ใช้ทรัพยากรไม่มากเท่ากับแบบจำลอง BERT

6.1.5. การประยุกต์ใช้งานร่วมกับระบบสารสนเทศโรงพยาบาล (Hospital Information System: HIS)

งานวิจัยนี้สร้างขึ้นเพื่อเป็นตัวช่วยแนะนำรหัสไอซีดีที่เกี่ยวข้องกับกลุ่มอาการของผู้ป่วย ซึ่งสามารถใช้งานได้ทั้งกลุ่มผู้ป่วยในและผู้ป่วยนอก โดยเฉพาะผู้ป่วยใน การลงรหัสไอซีดีนี้เป็นขึ้นส่วนสำคัญสำหรับการเบิกจ่ายเงินค่ารักษาคืนแก่โรงพยาบาล แนวทางการเชื่อมต่อแบบจำลองของเราเข้ากับระบบ HIS ของโรงพยาบาลนั้นสามารถทำได้โดยการเชื่อมต่อ API ของแบบจำลองเข้ากับระบบของโรงพยาบาล โดยระบบของโรงพยาบาลจะทำการส่งกลุ่มอาการและโรคมายังแบบจำลองผ่านทาง API แบบจำลองของเราจะเสนอรหัสไอซีดีที่เกี่ยวข้องกับอาการที่ได้รับมาและส่งกลับไปยังระบบของโรงพยาบาล ซึ่งจะช่วยให้ผู้เกี่ยวข้องกับการลงรหัสไอซีดีสามารถดำเนินการได้อย่างครบถ้วนและรวดเร็วเพื่อให้สอดคล้องกับจำนวนค่ารักษาที่โรงพยาบาลสมควรได้รับ

6.1.6. แนวทางการพัฒนาต่อ (Future work)

ความคาดหวังจากงานวิจัยนี้คือการนำไปใช้จริงให้เกิดประโยชน์สูงสุด ทั้งนี้ แนวทางการพัฒนาต่อจึงเป็นการนำเสนอแนวทางเพิ่มเติมที่ช่วยเพิ่มประสิทธิภาพของแบบจำลองและงานวิจัยนี้ให้ดียิ่งขึ้น

1. เปรียบเทียบแบบจำลองของเรากับแบบจำลองเชิงเดี่ยวในขนาดที่สามารถจัดการกับปัญหาการกระจายตัวแบบหางยาวของชุดข้อมูลโดยไม่ละทิ้งประชากรส่วนน้อยของข้อมูล จุดนี้เป็นแนวทางการปรับปรุงความสามารถในการจัดการปัญหาเกี่ยวกับความไม่สมดุลของข้อมูล
2. เปรียบเทียบกระบวนการ pre-trained word embedding ของเรากับการใช้ pre-trained word embedding จาก BERT ที่ผ่านการฝึกด้วยข้อมูลทางการแพทย์แบบหลายภาษามาใช้ร่วมกับแบบจำลองของเรา ซึ่งอาจจะช่วยให้แบบจำลองของเรามีประสิทธิภาพมากขึ้นยิ่งขึ้น
3. เปรียบเทียบประสิทธิภาพการจำแนกรหัสไอซีดีจากแบบจำลอง BERT กับแบบจำลองของเรา เนื่องจาก BERT เป็นแบบจำลองยอดนิยมสำหรับงานจำแนกข้อความถือเป็น Baseline ใหม่ของงานวิจัยเราได้

6.1.7. ข้อจำกัดและแนวทางการพัฒนาต่อ (Limitations)

- 6.1.7.1. เวลาในการประมวลผลต่อรอบจะใช้เวลานาน เนื่องจากแบบจำลอง Naïve Bayes เป็นแบบจำลองทางสถิติที่ต้องคำนวณความน่าจะเป็นซึ่งเวลาในการคำนวณจะขึ้นกับปริมาณคลังศัพท์และจำนวนคำของประโยค ยิ่งประโยคมีความยาวมากเวลาในการประมวลผลก็จะนานขึ้น สำหรับการวิจัยต่อไปอาจมุ่งไปที่ optimization แบบจำลอง Naïve Bayes
- 6.1.7.2. แบบจำลองของเราฝึกบนข้อมูลจากแหล่งเดียว (โรงพยาบาลแห่งเดียว) หากได้รับข้อมูลจากหลายแหล่งมาฝึก แบบจำลองมีความครอบคลุมมากยิ่งขึ้นเมื่อได้เรียนรู้รูปแบบการบันทึกที่แตกต่างกัน
- 6.1.7.3. แบบจำลองเป็นเครื่องมือสำหรับแนะนำรหัสไอซีดีที่เกี่ยวข้องกับข้อความที่ได้รับมา จำเป็นต้องมีผู้ให้รหัสตรวจสอบและเลือกรหัส เครื่องมือเรามีส่วนช่วยให้การทำงานนั้นรวดเร็วขึ้นและลดโอกาสผิดพลาดลงได้
- 6.1.7.4. แบบจำลองนี้ครอบคลุมเฉพาะรหัสไอซีดีที่ปรากฏในชุดข้อมูลวิจัยเท่านั้น ซึ่ง ICD-10-TM พิจารณาเฉพาะรหัสส่วนหมวดหมู่ (3หลักแรก) และ ICD-9-CM ที่พิจารณาครบทุกตำแหน่ง

6.2. การสรุปผล (Conclusion)

เรานำเสนอแบบจำลองสำหรับจัดการกับปัญหาชุดข้อมูลที่มีการกระจายตัวแบบหางยาว ภายใต้การจำแนกประเภทแบบหลายคลาสพร้อมทั้งรองรับข้อความที่ผสมทั้งภาษาไทยและอังกฤษ ชุดข้อมูลของเราที่รวมทั้งคำไทยและอังกฤษจะถูกจัดการด้วยการฝังคำที่ผ่านการฝึกด้วยข้อมูลที่มีเนื้อหาเป็นภาษาไทยและอังกฤษปะปนกัน ข้อมูลคำอธิบายรหัสไอซีดีและคลังศัพท์เพิ่มเติม

การจำแนกรหัสไอซีดี เราแบ่งแบบจำลองออกเป็น 2 ส่วน แต่ละส่วนจัดการกับชุดข้อมูลของกลุ่มรหัสไอซีดีที่แตกต่างกัน ชุดข้อมูลของรหัสไอซีดีที่พบยาก (โรคที่ไม่ค่อยป่วยกัน) ทำนายด้วยแบบจำลอง Multinomial Naïve Bayes ซึ่งใช้หลักการของความน่าจะเป็นสำหรับชุดข้อมูลของรหัสไอซีดีทั่วไป (โรคที่พบได้ทั่วไป ผู้ป่วยเป็นกันมาก) ทำนายด้วยแบบจำลองการเรียนรู้เชิงลึกซึ่งประกอบด้วย CNN Bi-LSTM และ Transformer encoder

ผลการทดสอบบนชุดข้อมูลบันทึกการรักษา (Course notes) ของรหัสที่พบได้ทั่วไป Bi-LSTM ทำนายรหัสได้ดีกว่าทุกแบบจำลองอยู่ที่ประมาณ 0.79 ที่ค่า Jaccard index และแนะนำรหัสได้ถูกต้อง 3 ใน 5 อันดับแรก ต่อมาเราพยายามเพิ่มประสิทธิภาพของแบบจำลองด้วยการใช้วิธี Ensemble ผลที่ได้คือวิธี Ensemble ด้วยการเลือกค่าสูงสุด ทำนายได้ดีกว่า CNN และ Transformer encoder อยู่ประมาณ 0.14 แต่ยังไม่พอกว่า Bi-LSTM อยู่ประมาณ 0.03 บนการวัดค่า Jaccard index ในส่วนความถูกต้องในการแนะนำรหัสไอซีดีนั้น วิธีการ Ensemble ให้ผลที่ดีจากส่วนของแบบจำลอง Bi-LSTM และในการรวมแบบจำลองเข้ากับแบบจำลอง Naïve Bayes พบว่าวิธี Ensemble ด้วยการเลือกค่าสูงสุดทำนายได้ดีกว่าแบบจำลองอื่น ๆ บนการวัดด้วยค่า Jaccard index

จากการทดสอบบนชุดข้อมูลบันทึกการให้ยา (Medication note) ของรหัสไอซีดีทั่วไปพบว่าแบบจำลอง CNN ให้ค่าที่ดีกว่าแบบจำลองอื่น ๆ ในทุกการวัดของเรา รหัสที่แนะนำใน 5 อันดับแรกก็ทำนายถูก 2 รหัส แบบจำลองที่ให้ผลใกล้เคียงกับแบบจำลอง CNN คือวิธี Ensemble ด้วยการเฉลี่ย แต่เมื่อนำแบบจำลองมาเชื่อมกับแบบจำลอง Naïve Bayes แบบจำลองที่ให้ค่า Jaccard index กลับเป็นวิธี Ensemble ด้วยการเลือกค่าสูงสุด สำหรับ Precision@3,5 และ Precision@8 แบบจำลอง Bi-LSTM และ วิธี Ensemble ด้วยการเฉลี่ย ให้ผลลัพธ์ที่ดีที่สุดตามลำดับ

บรรณานุกรม

- [1] ป. สวัสดิ์มงคล, "ปัจจัยที่มีผลต่อความสำเร็จในการให้รหัสโรค หัตถการและการผ่าตัดของผู้ป่วยในที่โรงพยาบาลศิริราช (Factors influence the success of ICD coding for inpatients at Siriraj Hospital)," (in en), *Siriraj Medical Bulletin*, vol. 2, no. 3, pp. 121-130, 2009 2009.
- [2] ศ. พรรณารุโณทัย and S. Pannarunothai, "กลุ่มวินิจฉัยโรคร่วม : การพัฒนาและประโยชน์ในประเทศไทย," (in en_US), 2544 2544.
- [3] *PyThaiNLP: Thai Natural Language Processing in Python*. (2023). Zenodo. Accessed: 2023-06-22 06:35:14.
- [4] V. Sornlertlamvanich, "Word segmentation for Thai in machine translation system," *Machine Translation*, 1993-01-01 1993.
- [5] *DeepCut: A Thai word tokenization library using Deep Neural Network*. (2019). Zenodo. Accessed: 2023-06-22 06:22:35.
- [6] P. Chormai, P. Prasertsom, and A. T. Rutherford, "AttaCut: A Fast and Accurate Neural Thai Word Segmenter," (in en).
- [7] A. P. Noto and D. R. S. Saputro, "Classification data mining with Laplacian Smoothing on Naïve Bayes method," presented at the International Conference of Mathematics and Mathematics Education (I-Cmme) 2021, 2022.
- [8] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, "Explainable Prediction of Medical Codes from Clinical Text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018 2018, New Orleans, Louisiana: Association for Computational Linguistics, pp. 1101-1111, doi: 10.18653/v1/N18-1100.
- [9] T. Vu, D. Q. Nguyen, and A. Nguyen, "A Label Attention Model for ICD Coding from Clinical Text," in *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, 7/2020 2020, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, pp. 3335-3341, doi:

- 10.24963/ijcai.2020/461.
- [10] E. Alsentzer *et al.*, "Publicly Available Clinical," (in en), *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72-78, 2019 2019, doi: 10.18653/v1/W19-1909.
- [11] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus," (in en), 2020-09-11 2020.
- [12] K. S. Kalyan and S. Sangeetha, "SECNLP: A survey of embeddings in clinical natural language processing," *J Biomed Inform*, vol. 101, p. 103323, Jan 2020, doi: 10.1016/j.jbi.2019.103323.
- [13] A. Blanco, O. Perez-de-Vinaspre, A. Perez, and A. Casillas, "Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity," *Comput Methods Programs Biomed*, vol. 188, p. 105264, May 2020, doi: 10.1016/j.cmpb.2019.105264.
- [14] P. Jatunarpit, K. Piromsopa, and C. Charoanlap, "Development of thai text-mining model for classifying ICD-10 TM," in *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 6/2016 2016, Ploiesti, Romania: IEEE, pp. 1-6, doi: 10.1109/ECAI.2016.7861163.
- [15] P. Ketpupong and K. Piromsopa, "Applying Text Mining for Classifying Disease from Symptoms," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, 9/2018 2018, Bangkok: IEEE, pp. 467-472, doi: 10.1109/ISCIT.2018.8587993.
- [16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," (in en), 2013-09-06 2013.
- [17] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014 2014, Doha, Qatar: Association for Computational Linguistics, pp. 1532-1543, doi: 10.3115/v1/D14-1162.
- [18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," (in en), *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 12/2017 2017, doi:

- 10.1162/tacl_a_00051.
- [19] W. Yin and H. Schütze, "Learning Word Meta-Embeddings," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016 2016, Berlin, Germany: Association for Computational Linguistics, pp. 1351-1360, doi: 10.18653/v1/P16-1128.
- [20] P. Jawanpuria, S. D. N T V, A. Kunchukuttan, and B. Mishra, "Learning Geometric Word Meta-Embeddings," in *Proceedings of the 5th Workshop on Representation Learning for NLP, 2020* 2020, Online: Association for Computational Linguistics, pp. 39-44, doi: 10.18653/v1/2020.repl4nlp-1.6.
- [21] D. Bollegala and J. O' Neill, "A Survey on Word Meta-Embedding Learning," in *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, 7/2022 2022, Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, pp. 5402-5409, doi: 10.24963/ijcai.2022/758.
- [22] N. Sadoughi *et al.*, "Medical code prediction with multi-view convolution and description-regularized label-dependent attention," (in en), 2018-11-04 2018.
- [23] E. Moons, A. Khanna, A. Akkasi, and M.-F. Moens, "A Comparison of Deep Learning Methods for ICD Coding of Clinical Records," *Applied Sciences*, vol. 10, no. 15, 2020, doi: 10.3390/app10155262.
- [24] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview," (in en), *Frontiers of Computer Science*, vol. 12, no. 2, pp. 191-202, 4/2018 2018, doi: 10.1007/s11704-017-7031-7.
- [25] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright, "Supervised Topic Models for Diagnosis Code Assignment to Discharge Summaries," A. Gelbukh, Ed., 2018 2018, Cham: Springer International Publishing, pp. 485–497.
- [26] T. Baumel, J. Nassour-Kassis, R. Cohen, and M. Elhadad, "Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment," (in en).
- [27] "ศูนย์มาตรฐานรหัสและข้อมูลสุขภาพแห่งชาติ." (accessed).
- [28] "ICD-10 Version:2019." (accessed).
- [29] ก. สังขวาสี, อักษรย่อที่หมอใช้ 2017. บริษัท ศรีเมืองการพิมพ์ จำกัด, 2017.
- [30] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A

- survey of word embeddings for clinical text," *J Biomed Inform*, vol. 100S, p. 100057, 2019, doi: 10.1016/j.yjbinx.2019.100057.
- [31] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
- [32] *thai2fit: Thai language Implementation of ULMFit*. (2021). Zenodo. Accessed: 2023-07-05 14:02:27.
- [33] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," 2010/05/22/ 2010, Valletta, Malta: ELRA, pp. 45-50.
- [34] Watson, Matthew, Qian, Chen, J. Bischof, F. Chollet, and others, "KerasNLP," ed, 2022.
- [35] "scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation." (accessed.
- [36] P. Szymański and T. Kajdanowicz, "A scikit-based Python environment for performing multi-label classification," 2018-12-10 2018, doi: 10.48550/arXiv.1702.01460.
- [37] R. a. A. Ramadhani, F. Indriani, and D. T. Nugrahadi, "Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016-10 2016, pp. 287-292, doi: 10.1109/ICACSIS.2016.7872720.
- [38] A. Vaswani *et al.*, "Attention is All you Need," (in en).
- [39] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," (in en).
- [40] T. O'Malley *et al.*, "KerasTuner," ed, 2019.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ประวัติผู้เขียน

ชื่อ-สกุล	ณัฐชา ชาชุม
วัน เดือน ปี เกิด	14 มิถุนายน 2539
สถานที่เกิด	ระยอง ประเทศไทย
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์
ที่อยู่ปัจจุบัน	58/228 หมู่บ้านชื้อตรงสุวินทวงศ์ ถนนสุวินทวงศ์ แขวงลำผักชี เขตหนอง จอก กรุงเทพมหานคร 10530



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY