

2110531 Data Science and Data Engineering Take-Home Midterm Exam
Paper Classification Report (2024/1)

Krissada Sarawit

6770215021

October 2024

สารบัญ

1	บทนำ	2
1.1	การกำหนดปัญหา	2
1.2	วัตถุประสงค์	2
2	การสำรวจข้อมูล	3
2.1	Dataset Description	3
2.2	Data Statistics	4
2.3	Data Preprocessing	4
3	Model	6
3.1	Model Architecture	6
3.2	Training Configuration	6
4	ผลลัพธ์	8
4.1	Kaggle Submission	8
5	การอภิปราย	9
5.1	การเพิ่มประสิทธิภาพ	9
6	สรุปผล	10

บทที่ 1

บทนำ

1.1 การกำหนดปัญหา

ในโลกของการวิจัยทางวิชาการ การแบ่งประเภทและการจำแนกวรรณกรรมมีบทบาทสำคัญในงานวรรณกรรม เพื่อผู้อ่านสามารถค้นหาผลงานทางวิชาการที่เกี่ยวข้องได้นั้น จำเป็นต้องมีวิธีการแบ่งประเภทวรรณกรรมอย่างถูกต้องและแม่นยำ ด้วยฐานข้อมูลอย่าง Scopus ที่มีบทความนับล้านบทความ การแยกประเภทด้วยการตีความโดยใช้มนุษย์จึงเป็นเรื่องที่ท้าทายและต้องใช้แรงงานอย่างมาก วิธีแก้ปัญหานี้มีแนวโน้มที่ดีอย่างหนึ่งคือ การจำแนกข้อความแบบหลายป้ายกำกับ หรือ Multi-label text classification

ปัญหา Multi-label text classification คือการจำแนกหมวดหมู่ของข้อความ โดยที่ทั้งความหนังสามารถเป็นไปได้หลายหมวดหมู่ ปัญหาที่ได้รับในโจทย์จะเป็นการจำแนกบทความทางวิชาการในด้านวิศวกรรมศาสตร์ โดยแบ่งเป็น 18 สาขา ได้แก่ civil, environmental, biomedical, petroleum, metallurgical, mechanical, electrical, computer, optical, nano, chemical, materials, agricultural, education, industrial, safety, "mathematics and statistics", and material science.

1.2 วัตถุประสงค์

วัตถุประสงค์ของการทดลองในครั้งนี้คือการพัฒนาโมเดลการจำแนกข้อความหลายป้ายกำกับที่มีประสิทธิภาพซึ่งสามารถจัดการณ์หัวข้อต่าง ๆ ได้อย่างแม่นยำจากเอกสารทุกดีย่อและเพื่องานวิจัย

เนื่องจากเอกสารแต่ละฉบับสามารถอ่านในสาขาวิชกรรมศาสตร์ที่แตกต่างกัน 18 สาขาแต่ละเอกสารสามารถอ่านได้มากกว่า 1 สาขา โนเดลจัดหมวดหมู่จึงจำเป็นต้องเรียนรู้รูปแบบที่ซับซ้อนจากข้อความและกำหนดป้ายกำกับที่เหมาะสมสมช่องท่อนถึงลักษณะสาขาวิชาการของเนื้อหา ซึ่งจะทำให้สามารถจัดหมวดหมู่วรรณกรรมทางวิชกรรมศาสตร์ได้อย่างแม่นยำ ซึ่งจะช่วยให้นักวิจัยจัดระเบียบและค้นหาผลงานวิจัยที่เกี่ยวข้องในโดเมนวิชกรรมศาสตร์ที่หลากหลายได้ถูกต้อง เป็นไปจะสิทธิภาพ

บทที่ 2

การสำรวจข้อมูล

2.1 Dataset Description

Dataset สำหรับการ Train มีทั้งหมด 454 แถว 3 คอลัมน์ ประกอบด้วย:

1. Title ชื่อของงานวิจัย
2. Abstract บทคัดย่อ
3. Classes ป้ายกำกับสาขาของงานวิจัยนั้น

ทำการโหลดไฟล์ train.json เข้า Pandas Dataframe

	Title	Abstract	Classes
1	Activated carbon derived from bacterial cellul...	© 2019 Elsevier B.V.Activated carbon derived f...	[CHE, MATENG]
2	The algorithm of static hand gesture recogniti...	© Springer International Publishing AG 2018.Te...	[CPE]
3	Alternative Redundant Residue Number System Co...	© 2018 IEEE.Residue number system (RNS) is a n...	[EE]
4	Comparative study of wax inhibitor performance...	© Published under licence by IOP Publishing Lt...	[PE, ME, CHE]
5	Undrained lower bound solutions for end bearing...	© 2019 John Wiley & Sons, Ltd.The undrained be...	[CE, MATSCI]

ภาพที่ 2.1: ตัวอย่าง Dataframe

```

    ✓ 0s df.info()

→ <class 'pandas.core.frame.DataFrame'>
Index: 454 entries, 1 to 454
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Title       454 non-null    object 
 1   Abstract    454 non-null    object 
 2   Classes     454 non-null    object 
dtypes: object(3)
memory usage: 30.4+ KB

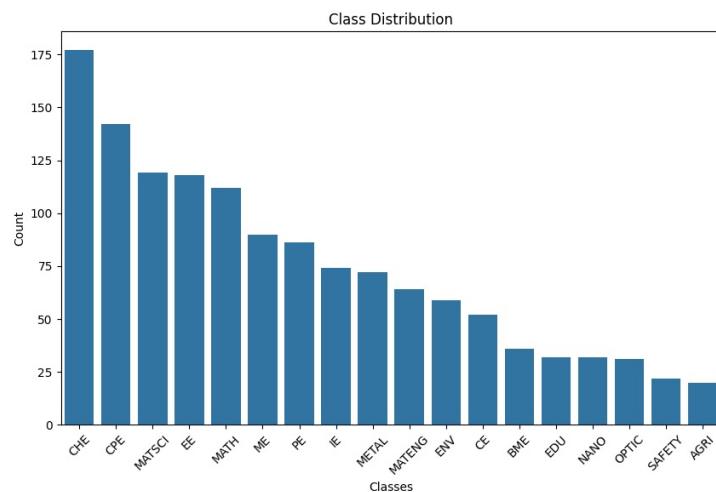
```

ภาพที่ 2.2: รายละเอียด Dataframe

รายละเอียด Dataframe ประกอบด้วย 454 แถว และ 3 คอลัมน์

2.2 Data Statistics

ถูกการกระจายตัวของข้อมูลแต่ละคลาส พบรากคลาสไม่สมดุลกัน โดยที่คลาสที่มีมากที่สุดคือ CHE มีจำนวนมากถึง 177 แถวและคลาสที่มีจำนวนน้อยที่สุดคือ AGRI มีเพียง 20 แถว ดังนั้นจะต้องคำนวณ class weight เพื่อทำไปใช้ใน loss function



ภาพที่ 2.3: การกระจายตัวของแต่ละ class

2.3 Data Preprocessing

การเตรียมข้อมูลประกอบด้วย แปลงให้ตัวอักษรเป็น lower case ลบตัวอักษรพิเศษ ลบเลข ลบช่องว่างที่ไม่จำเป็น และลดรูปคำให้เหลือเพียงคำฐาน โดยใช้ spacy library [1]

หลังจากนั้นทำการแปลง Label ของข้อมูลให้เป็นรูปแบบ Binary ด้วย MultiLabelBinarizer จาก library scikit-learn [2]

```

nlp = spacy.load("en_core_web_sm")

def preprocess_text(text):
    text = text.lower()
    # Remove © symbols
    text = text.replace('©', '')
    # Remove numbers
    text = re.sub(r'\d+', ' ', text)
    # Remove special characters
    text = re.sub(r'[^w\s]', ' ', text)
    # Remove extra whitespace
    text = re.sub(r'\s+', ' ', text).strip()
    return text

def lemmatize_text(text):
    # Word lemmatize
    doc = nlp(text)
    return ' '.join([token.lemma_ for token in doc])

def pre_process_text(text):
    text = preprocess_text(text)
    text = lemmatize_text(text)
    return text

```

ภาพที่ 2.4: การ Preprocess Data

	Title	Abstract	Classes	Text
1	Activated carbon derived from bacterial cellulose...	© 2019 Elsevier B.V.Activated carbon derived f...	[CHE, MATENG]	activate carbon derive from bacterial cellul...
2	The algorithm of static hand gesture recogniti...	© Springer International Publishing AG 2016.T...	[CPE]	the algorithm of static hand gesture recogni...
3	Alternative Redundant Residue Number System Co...	© 2018 IEEE.Residue number system (RNS) is a...	[EE]	alternative redundant residue number system co...
4	Comparative study of wax inhibitor performance...	© Published under licence by IOP Publishing Li...	[PE, ME, CHE]	comparative study of wax inhibitor performance...
5	Undrained lower bound solutions for end bearing...	© 2019 John Wiley & Sons, Ltd.The undrained be...	[CE, MATSCI]	undraine low bind solution for end bear capac...
...
450	A portable USB-controlled potentiostat for pap...	© 2018 IEEEThis paper presents a portable and ...	[CPE, CHE]	a portable usb control potentiostat for paper ...
451	Literature reviews on applying artificial intell...	Copyright © 2019 for this paper by its authors...	[CPE, EDU]	literature review on apply artificial intellig...
452	A multi-parameterized water quality prediction...	© 2019 The authors and IOS Press. All rights r...	[ENV, EE, CHE]	a multi parameterized water quality prediction...
453	Semantic Segmentation on Medium-Resolution Sat...	© 2018 IEEE.Semantic Segmentation is a fundame...	[EE, CPE, OPTIC, EDU]	semantic segmentation on medium resolution sat...
454	Reducing the defects of a-pillar stamping part...	© 2019 IEEE.This research aims to reduce defec...	[METAL, EDU, MATSCI]	reduce the defect of a pillar stamp part in th...

ภาพที่ 2.5: ข้อความที่อยู่ในคอลัมน์ Text เป็นข้อความที่ผ่านการ Preprocess แล้ว

บทที่ 3

Model

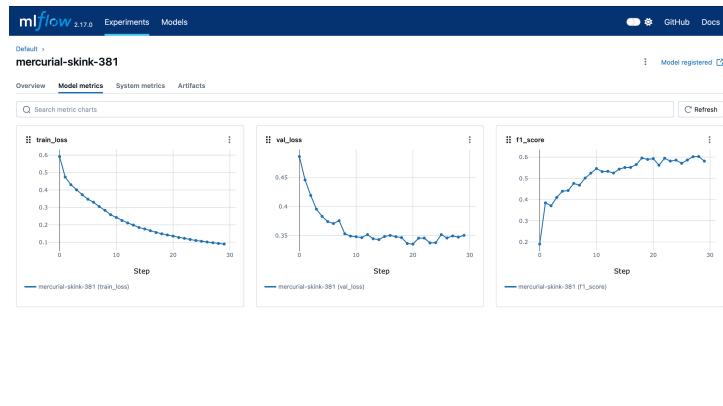
3.1 Model Architecture

ใช้ DistilBERT ซึ่งเป็น Pre-trained transformer model ด้วยเทคนิค distillation เป็นโมเดลที่มีขนาดเล็ก แต่ยังคงประสิทธิภาพส่วนใหญ่ไว้อยู่ [3] เนื่องจากข้อมูลมีจำนวนน้อย การ fine-tune โมเดลที่มีปริมาณ parameter น้อยจะสามารถทำได้ง่ายกว่า

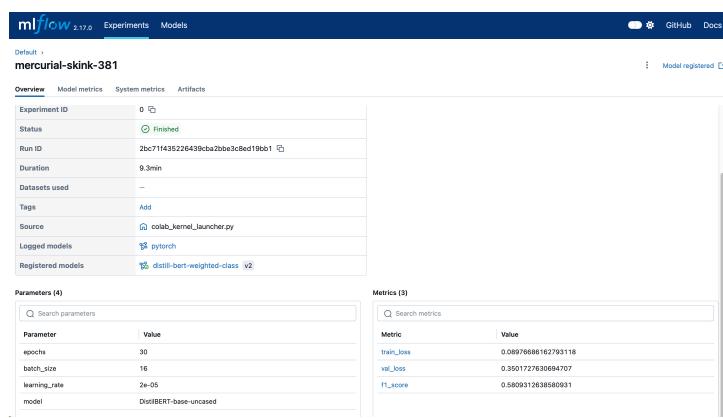
สถาปัตยกรรมโมเดลประกอบด้วย pre-trained DistilBERT ตามด้วย fully connected layer ที่ใช้ sigmoid เป็น activation function ที่จะทำหน้าที่แปลง output embeddings เป็นคะแนนความน่าจะเป็นของแต่ละคลาสทั้ง 18 คลาส

3.2 Training Configuration

- Optimizer: AdamW, learning rate 2e-5
- Loss function: BCEWithLogitsLoss, pos_weight
- Batch size: 16
- Max length: 512
- Number of epochs: 30



ภาพที่ 3.1: Model training graph



ภาพที่ 3.2: Model metrics from MLflow

โมเดลมี F1 Validation อยู่ที่ 0.5809 และค่า val_loss เริ่มคุ้นเคยที่ Epoch 10

บทที่ 4

ผลลัพธ์

4.1 Kaggle Submission

ผู้อ่านมีผลลัพธ์การ Submission ใน Kaggle ดังนี้

1. Public Score 0.4268

2. Private Score 0.4646

ภาพที่ 4.1: ผลการ Submission ใน Kaggle

Rank	Name	Submission	Private Score	Public Score	Team	Days Ago
5	- 3	67702406K7_Thansedch	0.5901	40	4d	
6	- 1	Pongob Krairaveeraj	0.5755	20	4d	
7	- 2	Psith Thanapapisarn	0.5745	28	4d	
8	- 6	Karanat Vettanasikul	0.5472	38	4d	
9	- 1	6770299621_Surewut_Kritikalwan	0.5278	19	4d	
10	- 1	6770238421_Nuttawat_Chantrapha	0.4968	21	4d	
11	--	6672100721_Chewin Hanssuda	0.4079	9	4d	
12	- 1	Krissada Sarawut	0.4646	7	6d	
13	- 1	6670150572_LNext	0.4537	8	4d	
14	--	lower-bound.csv	0.3538			
		Krissada Kwekai	0.3397	1	8d	

ภาพที่ 4.2: Leaderboard ใน Kaggle

บทที่ 5

การอภิปราย

5.1 การเพิ่มประสิทธิภาพ

จากการทดลองจะเห็นได้ว่า โมเดลไม่สามารถทำนาย Label ได้ดีนัก หากสามารถเพิ่มปริมาณข้อมูลที่ใช้ Fine-tune อาจเพิ่ม Performance ของโมเดลได้ โดยอาจทำได้โดย

1. แปลภาษาข้อมูลเป็นภาษาที่สองแล้วแปลกลับมาเป็นภาษาเดิม (Back translation) เพื่อเพิ่มปริมาณข้อมูล
2. การทำ Oversampling เพื่อลดปัญหา Imbalanced class
3. การทำ Early stopping เพื่อป้องกันโมเดล Overfitting
4. ใช้ Model ภาษาเฉพาะทางเช่น SciBERT ซึ่งเป็น BERT ที่ Train ด้วยข้อความในโดเมนวิทยาศาสตร์ ซึ่งจะทำให้โมเดลเข้าใจข้อความในบริบทและความทางวิศวกรรมได้ดีขึ้น

บทที่ 6

สรุปผล

การทำนายหมวดหมู่ของเอกสารบทความทางวิศวกรรมโดยใช้โมเดลภาษาในงานนี้ โดยการปรับแต่งโมเดล (fine-tuning) DistilBERT เพื่อรับรู้การจำแนกหมวดหมู่แบบหลายคลาส (multi-label classification) สำหรับหมวดหมู่ทางวิศวกรรมทั้ง 18 หมวดหมู่ โมเดลสามารถทำนายหมวดหมู่ของเอกสารความทางวิศวกรรมได้ถูกและมีประสิทธิภาพน้อยและไม่สมดุล

อย่างไรก็ตามยังพบข้อจำกัดในการทำนายหมวดหมู่ ซึ่งการปรับปรุงโมเดลในอนาคตอาจพิจารณาการใช้เทคนิคการสุ่มตัวอย่างข้อมูลเพิ่มเติม (data augmentation) หรือการเพิ่มข้อมูลเพื่อแก้ไขปัญหาความไม่สมดุล หรือปรับโมเดลให้มีความซับซ้อนมากขึ้น ซึ่งอาจจะช่วยให้โมเดลสามารถทำนายได้อย่างแม่นยำมากขึ้น

References

- [1] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.” To appear, 2017.
- [2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter,” *CoRR*, vol. abs/1910.01108, 2019.