โครงงานเลขที่ วศ.คพ. P025-2/2566

เรื่อง

การวิเคราะห์เชิงเปรียบเทียบของโมเดลทางภาษาสำหรับการจำแนกประเภท ESG

โดย

นายสุภาค ไชยเนตรเกษม รหัส 630610769
นายธนิสร ไชยวุฒิ รหัส 630610738

โครงงานนี้
เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่
ปีการศึกษา 2566

PROJECT No. CPE P025-2/2566

Comparative Analysis of ESG-NLP classification model

Supak Chainetkasem 630610769
Thanisorn Chaiwut 630610738

A Project Submitted in Partial Fulfillment of Requirements
for the Degree of Bachelor of Engineering
Department of Computer Engineering
Faculty of Engineering
Chiang Mai University
2023

หัวข้อโครงงาน	: การวิเคราะห์เชิงเปรียบเทียบของโมเดลทางภาษาสำหรับการจำแนกประเภท ESG
	: Comparative Analysis of ESG-NLP classification model
โดย	: นายสุภาค ไชยเนตรเกษม รหัส 630610769
	นายธนิสร ไชยวุฒิ รหัส 630610738
ภาควิชา	: วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	: รศ.ดร. ปฏิเวธ วุฒิสารวัฒนา
ปริญญา	: วิศวกรรมศาสตรบัณฑิต
สาขา	: วิศวกรรมคอมพิวเตอร์
ปีการศึกษา	: 2566

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่ ได้อนุมัติให้โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต (สาขาวิศวกรรมคอมพิวเตอร์)

..... หัวหน้าภาควิชาวิศวกรรมคอมพิวเตอร์
(รศ.ดร. สันติ พิทักษ์กิจนุกูร)

คณะกรรมการสอบโครงงาน

..... ประธานกรรมการ
(รศ.ดร. ปฏิเวธ วุฒิสารวัฒนา)

..... กรรมการ
(ผศ.ดร. ณัฐนันท์ พรหมสุข)

..... กรรมการ
(ผศ.ดร. กานต์ ปทานุคม)

หัวข้อโครงงาน	: การวิเคราะห์เชิงเปรียบเทียบของโมเดลทางภาษาสำหรับการจำแนกประเภท ESG
	: Comparative Analysis of ESG-NLP classification model
โดย	: นายสุภาค ไชยเนตรเกษม รหัส 630610769
	นายธนิสร ไชยวุฒิ รหัส 630610738
ภาควิชา	: วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	: รศ.ดร. ปฏิเวธ วุฒิสารวัฒนา
ปริญญา	: วิศวกรรมศาสตรบัณฑิต
สาขา	: วิศวกรรมคอมพิวเตอร์
ปีการศึกษา	: 2566

บทคัดย่อ

ในปัจจุบันนักลงทุนได้ตระหนักว่าความพยายามในการทำกำไรเพื่อเอาชนะแนวโน้มของตลาดการลงทุนเป็นเรื่องที่ยาก นักลงทุนทั่วโลกจึงหันมาให้ความสนใจกับการลงทุนในบริษัทที่ให้ความสำคัญกับพัฒนาองค์กรอย่างยั่งยืน ซึ่งประกอบไปด้วยสามด้านหลักคือ ด้านสิ่งแวดล้อม(Environment), ด้านสังคม(Social) และด้านธรรมาภิบาล(Governance) โดยมีแนวคิดคือการลงทุนที่เน้นความยั่งยืนจะนำมาซึ่งผลตอบแทนที่มากขึ้นตามไปด้วย ในงานวิจัยนี้จะเน้นไปที่การเก็บรวบรวมรายงานผลประกอบการประจำปีจากบริษัทในดัชนีSETของตลาดหลักทรัพย์แห่งประเทศไทยนำมาแบ่งประโยคและจำแนกแต่ละประโยคเพื่อดูว่าแต่ละบริษัทมีประโยคที่เกี่ยวข้องกับด้านสิ่งแวดล้อม(Environment), ด้านสังคม(Social), ด้านธรรมาภิบาล(Governance) หรือไม่เกี่ยวข้องกับด้านใดเลย(Neutral) มากเพียงใดด้วยโมเดลด้านภาษาธรรมชาติ และหาว่าโมเดลด้านภาษาธรรมชาติแบบใดที่สามารถจำแนกประโยคในรายงานประจำปีว่าเป็นด้านสิ่งแวดล้อม(Environment), ด้านสังคม(Social) หรือด้านธรรมาภิบาล(Governance)หรือไม่เกี่ยวข้องกับด้านใดเลย(Neutral)ได้ดีที่สุด

Project Title	: Comparative Analysis of ESG-NLP classification model
Name	: Supak Chainetkasem 630610769
	Thanisorn Chaiwut 630610738
Department	: Computer Engineering
Project Advisor	: Assoc. Prof. Patiwet Wuttisarnwattana, Ph.D.
Degree	: Bachelor of Engineering
Program	: Computer Engineering
Academic Year	: 2023

ABSTRACT

Nowadays, investors have realized that it is difficult to make profitable attempts to overcome the trend of the investment market. Investors around the world have therefore turned their attention to investing in companies that place importance on sustainable organizational development. which consists of three main aspects: Environmental, Social, and Governance, with the idea that investments that emphasize sustainability will bring greater returns as well. In this research, we will focus on collecting annual reports from companies in the SET Index of the Stock Exchange of Thailand. Divide it into sentences and classify each sentence to see how many sentences in each company relate to the environment (Environment), society (Social), governance (Governance), or not related to any aspect at all(Neutral) with natural language models and find the best natural language model that can classify sentences in annual reports as environmental, social, or governance, or not related to any aspect at all (Neutral).

กิตติกรรมประกาศ

โครงงานนี้ได้รับความกรุณาจาก ผศ.ดร.ปฏิเวธ วุฒิสารวัฒนา อาจารย์ที่ปรึกษาที่ได้สละเวลาให้ความช่วยเหลือทั้งให้คำแนะนำให้ความรู้และแนวคิดต่างๆรวมถึง ผศ.ดร.กานต์ ปทานุคม และ ผศ.ดร.ณัฐนันท์ พรหมสุข ที่ให้คำปรึกษาจนทำให้โครงงานเล่มนี้เสร็จสมบูรณ์ไปได้

ขอบคุณคนใน Biomedical Imaging Lab ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ได้ให้คำปรึกษาการทำโครงงานและให้สนับสนุนทางด้านสื่อการศึกษาออนไลน์ขอขอบคุณเพื่อนๆที่ให้กำลังใจลอดการทำโครงงานที่ผ่านมา

นอกจากนี้ผู้จัดทำขอขอบพระคุณบิดา มารดาที่ได้ให้ชีวิต เลี้ยงดูสั่งสอน และส่งเสียให้กระผมได้ศึกษาเล่าเรียนจนจบหลักสูตรปริญญาตรี หลักสูตรวิศวกรรมศาสตรบัณฑิต ซึ่งท่านได้ให้กำลังใจตลอดมาจนทำให้โครงงานนี้สำเร็จ รวมทั้งขอขอบพระคุณอีกหลายๆท่านที่ไม่ได้เอ่ยนามมา ณ ที่นี้ ที่ได้ให้ความช่วยเหลือตลอดมา หากหนังสือโครงงานเล่มนี้มีข้อผิดพลาดประการใด กระผมขอน้อมรับด้วยความยินดี

นายสุภาค ไชยเนตรเกษม
นายธนิสร ไชยวุฒิ
29 มีนาคม 2567

สารบัญ

สารบัญรูป

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ที่มาของโครงงาน

ในปัจจุบันการลงทุนในหุ้นนั้นถือเป็นหนึ่งในการลงทุนที่ดีที่สุด เนื่องจากการลงทุนในหุ้นนั้นมีข้อดีมากมาย เช่น การมีสภาพคล่องและผลตอบแทนที่สูง ความสามารถในการสร้างกระแสเงินสดจากเงินปันผล นอกจากนั้นถ้าถือในระยะยาวพอจะทำให้โอกาสในการขาดทุนนั้นมีน้อยมาก แต่การลงทุนในหุ้นก็ไม่ได้เหมาะกับคนทุกคน เนื่องจากหุ้นนั้นเป็นสินทรัพย์ที่มีความผันผวนของราคาสูง ดังนั้นคนที่รับความเสี่ยงได้น้อยก็จะไม่เหมาะกับการลงทุนประเภทนี้

การลงทุนในหุ้นนั้นมีแนวทางมากมาย เช่น การลงทุนในหุ้นเติบโต การลงทุนในหุ้นคุณค่าหรือการลงทุนโดยเน้นลงทุนในหุ้นที่ให้เงินปันผลที่สูง โดยในโครงงานนี้จะนำเสนอแนวทางการลงทุนที่มีชื่อว่า การลงทุนอย่างยั่งยืน (Sustainable Investment) โดยการลงทุนอย่างยั่งยืนนี้ จะมุ่งเน้นการลงทุนไปยังบริษัทที่ให้ความสำคัญกับสามด้าน คือ สิ่งแวดล้อม (Environment), สังคม(Social) และธรรมาภิบาล (Governance) โดยจะมีการให้คะแนนทางด้านความยั่งยืนที่มีชื่อว่าคะแนน ESG

โครงงานฉบับนี้มีจุดประสงค์สองอย่าง คือ สร้างระบบที่สามารถบ่งบอกได้ว่าบริษัทแต่ละบริษัทมีประโยคที่พูดถึงเกี่ยวกับประเด็นที่เกี่ยวข้องกับ ESG อยู่มากน้อยเพียงใดจากการให้ระบบอ่านรายงานประจำปีของบริษัทต่างๆ และพัฒนาโมเดลด้านภาษาธรรมชาติที่ดีที่สุดจากโมเดลด้านภาษาธรรมชาติหลายๆโมเดล เพื่อจำแนกว่าประโยคแต่ละประโยคในรางงานประจำปีพูดถึงหัวข้อใดใน ESG หรือไม่พูดถึงหัวข้อใดเลย

1.2 วัตถุประสงค์ของโครงงาน

1. พัฒนาโปรแกรมเพื่อวิเคราะห์ประโยคว่าประโยคที่ปรากฏในรายงานประจำปีเป็นชนิด สิ่งแวดล้อม(Environment), สังคม (Social), ธรรมาภิบาล (Governance) หรือประโยคที่ไม่เข้าชนิดใดเลย (Neutral) โดยใช้โมเดลด้านภาษาธรรมชาติและค้นหาว่าโมเดลด้านภาษาธรรมชาติชนิดใดที่นำมาใช้กับงานนี้ได้ดีที่สุด
2. พัฒนาโปรแกรมเพื่อวิเคราะห์รายงานประจำปีเพื่อหาว่ารายงานประจำปีที่วิเคราะห์มีประโยคชนิดสิ่งแวดล้อม(Environment), สังคม (Social), ธรรมาภิบาล (Governance) และประโยคที่ไม่เข้าชนิดใดเลย (Neutral) อยู่มากน้อยเพียงใดโดยใช้โปรแกรมวิเคราะห์วิเคราะห์ประโยคที่พัฒนาจากโมเดลด้านภาษาธรรมชาติ

1.3 ขอบเขตของโครงงาน

1. ใช้ข้อมูลจากรายงานประจำปีที่เผยแพร่ในเว็บไซต์ของตลาดหลักทรัพย์แห่งประเทศไทยเท่านั้น
2. ใช้รายงานประจำปีฉบับภาษาอังกฤษเท่านั้น

1.4 ประโยชน์ที่ได้รับ

เนื่องจากข้อมูลเกี่ยวการพูดถึงประเด็นที่เกี่ยวข้องกับESGของบริษัทต่างๆมีการเปิดเผยข้อมูลที่น้อยมาก โดยถ้าอยากได้ข้อมูลที่มากขึ้นก็จะมีค่าใช้จ่ายเพิ่มเติมในการเปิดเผยข้อมูล ทางผู้จัดทำจึงพัฒนาโปรแกรมขึ้นมาเพื่ออ่านรายงานประจำปีของบริษัทเพื่อดูว่าบริษัทนั้นมีการพูดถึงเกี่ยวกับประเด็นESGมากน้อยเพียงใด เพื่อ

ที่จะไม่ต้องใช้แรงงานมนุษย์ในการหาข้อมูลจากรายงานประจำปี ดังนั้นจึงทำให้ต้นทุนในส่วนนี้หายไป ข้อมูลนี้จึงสามารถเปิดเผยข้อมูลทั้งหมดได้โดยไม่มีค่าใช้จ่ายใด ๆ

1.5 เทคโนโลยีและเครื่องมือที่ใช้

1.5.1 เทคโนโลยีด้านฮาร์ดแวร์

Notebook Acer Nitro 7 AN715-51 และ Notebook Acer Nitro 5 AN515-51 สำหรับงานทั้งหมดในโครงงานนี้

1.5.2 เทคโนโลยีด้านซอฟต์แวร์

ใช้Pythonในการเขียนโปรแกรมและใช้libraryดังนี้ Pandas ,NumPy ,Scikit Learn ,Tensorflow ,Keras ,Transformers ,Spacy ,PyTorch ,PyMuPDF ,Nltk

1.6 แผนการดำเนินงาน

ขั้นตอนการดำเนินงาน	ต.ค. 2566	พ.ย. 2566	ธ.ค. 2566	ม.ค. 2567	ก.พ. 2567	มี.ค. 2567
ศึกษาค้นคว้าข้อมูล						
เตรียมข้อมูลสำหรับพัฒนาโมเดลทางด้านภาษา						
พัฒนาโปรแกรมจำแนกประโยคและทดลองหาโมเดลทางด้านภาษาที่ดีที่สุด						
พัฒนาโปรแกรมอ่านรายงานประจำปีของบริษัท						

1.7 บทบาทและความรับผิดชอบ

ผู้จัดทำทั้งสองคนช่วยกันทำงานทั้งหมดทุกส่วน แต่จะแบ่งงานหลักๆที่แต่ละคนได้ทำเป็นส่วนใหญ่ได้ดังนี้ นายสุภาค ไชยเนตรเกษม: รับผิดชอบหน้าที่ในการใช้การประมวลผลภาษาธรรมชาติเพื่อพัฒนาโปรแกรมจำแนกประโยคและทดลองเพื่อหาโมเดลทางด้านภาษาที่ดีที่สุด และเตรียมข้อมูลสำหรับพัฒนาโมเดลทางด้านภาษา นายธนิสร ไชยวุฒิ: รับผิดชอบหน้าที่ในการเตรียมข้อมูลสำหรับพัฒนาโมเดลทางด้านภาษา ศึกษาหาข้อมูลอ้างอิงที่จะนำมาใช้ในงานส่วนต่างๆ และพัฒนาโปรแกรมอ่านรายงานประจำปีของบริษัท

1.8 ผลกระทบด้านสังคม สุขภาพ ความปลอดภัย กฎหมาย และวัฒนธรรม

ผลกระทบทางด้านสังคม: ระบบที่เป็นผลลัพธ์ของโครงงานนี้จะสามารถใช้เพื่อเป็นตัวช่วยสำหรับนักลงทุนทั่วไปที่สนใจในการลงทุนเชิงยั่งยืน เนื่องจากในปัจจุบันนี้ข้อมูลเกี่ยวกับผลการดำเนินงานทางด้าน ESG ของบริษัทยังมีไม่มากนัก จึงหวังว่าระบบที่ถูกพัฒนามานี้จะสามารถช่วยได้ไม่มากก็น้อย

ผลกระทบทางด้านสุขภาพ: ตัวระบบที่ถูกพัฒนาขึ้นมานั้นไม่ส่งผลกระทบต่อสุขภาพแต่อย่างใด

ผลกระทบทางด้านกฎหมาย: ข้อมูลที่ใช้ในโครงงานนี้รวมไปถึงซอฟต์แวร์ที่ใช้ในการพัฒนานั้นทั้งหมดล้วนเป็นสิ่งที่สามารถหาได้ทั่วไปโดยไม่ละเมิดลิขสิทธิ์ ดังนั้นการทำโครงงานนี้จึงไม่มีผลกระทบด้านกฎหมายอย่างแน่นอน

ผลกระทบทางด้านวัฒนธรรม: ตัวระบบที่ถูกพัฒนาขึ้นมานั้นไม่มีผลกระทบทางด้านวัฒนธรรมแต่อย่างใด

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

การทำโครงงาน เริ่มต้นด้วยการศึกษาค้นคว้า ทฤษฎีที่เกี่ยวข้อง หรือ งานวิจัย/โครงงาน ที่เคยมีผู้นำเสนอไว้แล้ว ซึ่งเนื้อหาในบทนี้ก็จะเกี่ยวกับการอธิบายถึงสิ่งที่เกี่ยวข้องกับโครงงาน เพื่อให้ผู้อ่านเข้าใจเนื้อหาในบทถัดๆ ไปได้ง่ายขึ้น

2.1 Data Cleaning

กระบวนการตรวจสอบ สะสาง แก้ไข หรือจัดรูปแบบข้อมูลให้อยู่ในสภาพที่พร้อมใช้งานที่สุด รวมไปถึงคัดกรองข้อมูลที่ไม่ถูกต้องหรือไม่จำเป็นออกไปจากชุดข้อมูลที่จะใช้วิเคราะห์หรือประมวลผล เพื่อให้ชุดข้อมูลที่จะใช้มีความสมบูรณ์ มีคุณภาพ พร้อมนำไปวิเคราะห์และใช้ประโยชน์

2.2 Bidirectional Encoder Representations from Transformers (BERT)

Bert มีชื่อเต็มว่า Bidirectional Encoder Representations from Transformersคือ โมเดลที่ต่อยอดมาจากโมเดลที่เรียกว่า Transformers ซึ่งถูกออกแบบมาให้เลือกใช้เฉพาะส่วนที่เป็น encoder ทำหน้าที่ในการแปลงคำในประโยคให้กลายเป็นเวกเตอร์ จากนั้นจึงใช้วิธีการฝึกโมเดลในรูปแบบที่ต่างออกไปจากโมเดลทางภาษาอื่นๆ ซึ่งการฝึกของ BERT จะแบ่งออกเป็นสองส่วนได้แก่

2.2.1 Masked Language Model

เป็นการฝึกโมเดลโดยคำในประโยคที่ป้อนเข้ามาเป็นอินพุตของระบบจะถูกลบออกไปบางส่วนและเรียกคำที่ถูกลบออกไปนี้ว่า (Masked words) โมเดลจะต้องพยายามเติมคำที่หายไปเหล่านี้ให้ถูกต้องซึ่งหากโมเดลจะสามารถเติมคำได้ถูกต้อง การเรียนรู้แบบนี้ช่วยให้โมเดลสามารถเรียนรู้ความสัมพันธ์และบริบทของคำในประโยคได้ดีโดยข้อมูลที่นำมาใช้ในการฝึกโมเดลเป็นคลังข้อมูลทางภาษาที่มีขนาดใหญ่ เช่น Wikipedia

2.2.2 Fine Tuning on Specific Tasks

เป็นการเพิ่ม Layer พิเศษเข้าไปในชั้นเอาท์พุตของโมเดล โดย Layer พิเศษเหล่านี้จะสามารถทำให้โมเดลมีความสามารถอื่นเพิ่มเติม เช่น สามารถประมวลผลได้ว่าประโยคที่รับเข้ามามีใจความที่เป็นแง่บวกหรือแง่ลบ, ความสามารถในการตอบคำถาม รวมถึงความสามารถในการแปลภาษา จากข้อมูลข้างต้น ด้วยความสามารถที่มากกว่าโมเดลทางภาษาแบบอื่นและความสะดวกสบายในการปรับใช้งานได้หลากหลายรูปแบบ ทำให้โมเดลทางภาษาแบบ BERT เป็นโมเดลทางภาษาที่ได้รับความนิยมเป็นอย่างมากและถูกใช้งานในหลายภาคส่วน เช่น ระดับอุตสาหกรรมและในเชิงวิชาการ

2.3 Recurrent Neural Network (RNN)

แบบหนึ่งที่ออกแบบมาแก้ปัญหาสำหรับงานที่ข้อมูลมีลำดับ Sequence โดยใช้หลักการ Feed สถานะภายในของโมเดลกลับมาเป็น Input ใหม่ คู่กับ Inputปกติ เรียกว่า Hidden State, Internal State, Memory ช่วยให้โมเดลรู้จำ Pattern ของลำดับ Input Sequence ได้

2.3.1 Long-Short Term Memory (LSTM)

จากปัญหาที่เกิดขึ้นใน RNNs เกี่ยวกับค่า gradient ที่มีค่าน้อยลงจากการทำงานของ backpropagation จึงได้มีการคิดค้น machine learning ตัวใหม่ที่ใช้หลักการคล้าย ๆ เดิม แต่เปลี่ยนตัวฟังก์ชันด้านในให้มีความเสถียรและมีประสิทธิภาพมากขึ้น ซึ่งนั่นก็คือ Long Short-Term Memoryหรือเรียกย่อๆว่า LSTMs สิ่งที่โดดเด่นขึ้นมานั้นก็คือการที่มันสามารถเลือกได้ว่าข้อมูลไหนที่ควรจะจดจำ ข้อมูลไหนที่ควรจะกำจัดทิ้งออกไปผ่านการลืมของสถานะใน node นั้น ๆ

2.4 ELECTRA

โมเดล ELECTRA เป็นโมเดลทางภาษาที่นำเสนอเป็นวิธีการ pretraining ใหม่ๆ โดยการฝึกโมเดล transformer สองตัว คือ generator และ discriminator โดยที generator มีบทบาทในการแทนที่โทเค็นในลำดับและจึงถูกฝึกเป็น masked language model ในขณะที่ discriminator ซึ่งเป็นโมเดลที่เราสนใจพยายามระบุว่าโทเค็นไหนถูกแทนที่โดย generator ในลำดับนั้นๆ

2.5 การลงทุนอย่างยั่งยืน

หมายถึง แนวคิดการลงทุนที่คำนึงถึงการดำเนินงานด้านสิ่งแวดล้อม สังคม และบรรษัทภิบาลของธุรกิจประกอบการพิจารณาตัดสินใจลงทุนควบคู่ไปกับการวิเคราะห์ข้อมูลทางการเงินของธุรกิจ เพื่อสร้างผลตอบแทนในระยะยาวและสร้างผลกระทบเชิงบวกหรือลดผลกระทบเชิงลบต่อสังคมและสิ่งแวดล้อม

2.6 ความรู้ตามหลักสูตรซึ่งถูกนำมาใช้หรือบูรณาการในโครงงาน

จากหลักสูตรที่ได้เรียนทั้งหมดที่ผ่านมาทำให้ได้ความรู้จากวิชา 261448 หรือ Data Mining For CPE ที่มีความรู้พื้นฐานเกี่ยวกับการเรียนรู้ของเครื่อง และวิชา 261456 หรือ Intro Computer Intelligence For CPE ที่มีความรู้เกี่ยวกับโครงข่ายประสาทเทียมและวิชา 261459 หรือ Deep Learning ที่มีความรู้เกี่ยวกับพื้นฐานเกี่ยวกับการเรียนรู้เชิงลึกและวิชา 261499 หรือ Natural Language Processing ที่มีความรู้เกี่ยวกับพื้นฐานเกี่ยวกับการประมวลผลภาษาธรรมชาติ นำมาใช้เป็นแนวคิดในการพัฒนาตัวโครงงานนี้

2.7 ความรู้นอกหลักสูตรซึ่งถูกนำมาใช้หรือบูรณาการในโครงงาน

ความรู้ในทางการเงินในการอ่านรายงานประจำปีและวิธีจำแนกประโยคว่าประโยคนั้นๆมีหัวข้อใดบ้างใน ESG เพื่อหาข้อมูลมาใช้ในการพัฒนาโมเดล

บทที่ 3
โครงสร้างและขั้นตอนการทำงาน

3.1 การเตรียมข้อมูล

3.1.1 ดาวน์โหลดข้อมูล

เก็บเอกสารรายงานประจำปีย้อนหลังไม่เกิน 5 ปี จากเว็ปไซต์ของตลาดหลักทรัพย์แห่งประเทศไทย โดยเอกสารจะอยู่ในรูปแบบไฟล์ PDF เพื่อเตรียมสำหรับจัดทำข้อมูลประโยคที่เกี่ยวข้องกับการดำเนินงาน ที่ให้ความสำคัญกับ สิ่งแวดล้อม(Environment), สังคม(Social) และธรรมาภิบาล(Governance)

3.1.2 แบ่งรายงานประจำปีเป็นประโยค

สร้างโปรแกรมแบ่งประโยคที่เมื่อนำรายงายประจำปีเข้าไปแล้วโปรแกรมจะตัดประโยคที่ไม่สมบูรณ์หรือพวกของที่ไม่จำเป็น เช่น ตาราง และหัวข้อ ออกไปและคัดมาให้แค่ประโยคที่สามารถใช้กับโมเดลได้ออกมาเป็นผลลัพธ์

3.1.3 เลือกข้อมูลที่จะนำไปใช้พัฒนาโปรแกรม

สุ่มอ่านประโยคในรายงานประจำปีของบริษัทที่เป็นผลลัพธ์ของโปรแกรมแบ่งประโยคที่เตรียมเอาไว้เพื่อเตรียมข้อมูลประโยคประเภทสิ่งแวดล้อม(Environment), สังคม(Social), ธรรมาภิบาล(Governance) และประโยคที่ไม่เข้าประเภทใดเลย(Neutral) โดยใช้เกณฑ์จากคู่มือของตลาดหลักทรัพย์แห่งประเทศไทย [5]

3.1.4 การทำความสะอาดข้อมูล

นำประโยคที่เลือกมาทำการทำความสะอาดข้อมูลก่อนนำไปใช้กับโมเดล โดยวิธีการที่ใช้ทำความสะอาดข้อมูลคือ การตัดเครื่องหมายหรือสัญลักษณ์พิเศษต่างๆออก, การตัด url ออก, การเปลี่ยนตัวอักษรทั้งหมดเป็นตัวพิมเล็ก และการลบคำหยุดออก

3.2 การพัฒนาโมเดลการเรียนรู้ของเครื่อง

พัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อนำมาทำนายชนิดของประโยคว่าเป็นหัวข้อใดในESGหรือไม่เกี่ยวข้องกับหัวข้อใดเลย โดยทดลองกับโมเดลหลายแบบซึ่งโมเดลที่เลือกได้แรงบันดาลใจมาจากงานวิจัย BERT goes sustainable : an NLP approach to ESG financing [8] ได้แก่โมเดล Bert [2], Distilbert [6], Albert-V2 [3], Electar-base [1], Electar-small [1], Roberta [4]และใช้ LSTM [7] เป็นโมเดลพื้นฐานการวัดผลแล้วเลือกโมเดลที่ดีที่สุดโดยวัดประสิทธิภาพจากค่า Accuracy

3.3 การพัฒนาโปรแกรมอ่านรายงานประจำปีของบริษัท

นำโมเดลการเรียนรู้ของเครื่องที่ดีที่สุดที่พัฒนามาได้มาสร้างโปรแกรมนับจำนวนประโยคแต่ละชนิดในรายงานประจำปีที่ใส่เข้าไปเป็นอินพุต เพื่อดูว่ารายงานประจำปีที่ใส่เข้าไปนั้นมีการพูดถึงประเกี่ยวกับสิ่งแวดล้อม, สังคม และธรรมาภิบาลมากน้อยเพียงใดและพูดประโยคที่ไม่เกี่ยวข้องกับประเด็นใดเลยในESGมากน้อยเพียงใด

บทที่ 4
การทดลองและผลลัพธ์

ผลลัพธ์จากโครงงานฉบับนี้จะแบ่งเป็นสองส่วนคือ ส่วนของระบบที่ใช้ในการอ่านรายงานและจำแนกชนิดประโยค และโปรแกรมอ่านรายงานประจำปีของบริษัทว่ามีประโยคที่พูดถึงเกี่ยวกับประเด็น ESG ของบริษัทและประโยคที่ไม่เกี่ยวข้องกับประเด็น ESG เลยมากน้อยเพียงใด

4.1 โมเดลที่ใช้ในการจำแนกชนิดประโยค

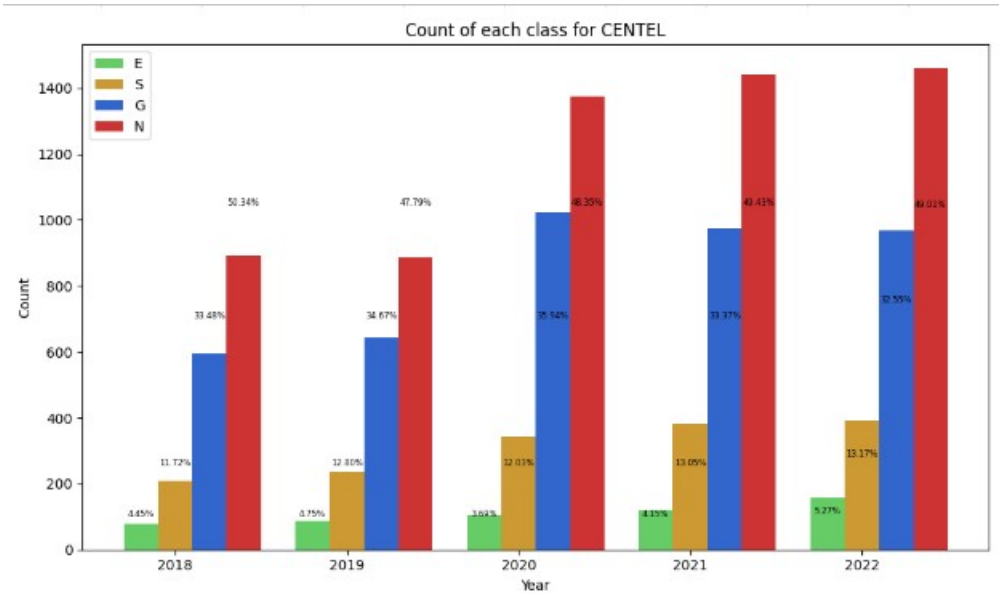
จากการเปรียบเทียบผลลัพธ์ของโมเดลทั้งหมดที่มี ทำให้เห็นได้ว่าโมเดลที่ค่าความแม่นยำสูงที่สุดคือโมเดล Albert-v2 โดยได้ค่าความแม่นยำอยู่ที่ 92.07 % ซึ่ง Albert-v2 ถือเป็นโมเดลขนาดเล็กหากเทียบกับโมเดลอื่นๆที่นำมาทดลอง โดยประโยคที่ทำนายผิดมักเป็นประโยคที่มีคำที่สอดคล้องกับประเด็นหนึ่งๆอยู่มากแต่แท้จริงแล้วความหมายไม่ได้เกี่ยวกับประเด็นนั้น เช่น ประโยคที่พูดถึงการจำหน่ายพลังงานหรือผลผลิตจากธรรมชาติ ซึ่งมีคำที่เกี่ยวข้องกับสิ่งแวดล้อมอยู่มากแต่แท้จริงแล้วประโยคนั้นไม่ได้กล่าวถึงการรักษาสิ่งแวดล้อม และประโยคที่มีความหมายที่ต้องตีความหลายชั้นถึงจะทราบความหมายที่แท้จริง และประโยคที่สามารถเป็นได้หลายประเด็นใน ESG ซึ่งเราได้จำแนกประเภทตามประเด็นที่ประโยคนั้นให้ความสำคัญมากกว่าหรือชัดเจนกว่าแต่โมเดลทำนายได้เป็นประเด็นที่มีความสำคัญน้อยกว่า ซึ่งผลลัพธ์ที่ทำนายผิดนั้นถือว่าสมเหตุผลและยอมรับได้เพราะประโยคที่ทำนายผิดมักเป็นประโยคที่แม้แต่มนุษย์ยังจำแนกได้ยาก

4.2 โปรแกรมอ่านรายงานประจำปีของบริษัท

เมื่อได้โมเดลที่ดีที่สุดสำหรับการการจำแนกชนิดประโยคซึ่งก็คือ Albert-v2 เราจึงทดลองนำโมเดลนั้นไปสร้างโปรแกรมสำหรับจำแนกประโยคของรายงานประจำปีทั้งเล่มในทุกๆปีย้อนหลัง5ปีของบริษัทที่ใส่เข้าเป็นอินพุต โดยเมื่อเราได้ทดลองใช้โปรแกรมนั้นกับหลายๆบริษัทแล้ว จากผลลัพธ์ที่ได้เราจึงได้ข้อมูลที่สรุปได้ว่าบริษัทส่วนมากมักมีแนวโน้มในการพูดถึงประโยคที่เกี่ยวข้องกับสิ่งแวดล้อมและสังคมเพิ่มมากขึ้นทุกปี ซึ่งอาจเป็นเพราะประเด็นของESGที่กำลังได้รับความนิยมในปัจจุบัน และประโยคส่วนมากเป็นประโยคที่เกี่ยวข้องกับธรรมาภิบาลและไม่เกี่ยวข้องกับประเด็นไหนเลย เพราะในรายงานประจำปีมักมีการพูดถึงเรื่องการบริหารจัดการในบริษัทเป็นจำนวนมากและเรื่องอื่นๆที่ไม่เกี่ยวข้องกับประเด็นESGก็มีอยู่มากเช่นกัน

Model	Accuracty
Roberta	85.98
Albert-v2	92.07
Bert base uncased	85.37
Electra base	89.02
Distilbert	87.8
LSTM	75.63
Electra small	86.59

รูปที่ 4.1: ความแม่นยำในการทำนายของโมเดลแต่ละแบบ



รูปที่ 4.2: ทดลองใช้โปรแกรมอ่านรายงานประจำปีของบริษัท CENTEL

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 สรุปผล

ในการทำโครงงานนี้สามารถพัฒนาระบบที่ใช้ในการจำแนกประเภทประโยคในรายงานประจำปีอัตโนมัติ โดยระบบนี้มีความแม่นยำถึง 92.07 % และประโยคที่ทำนายผิดก็มักเป็นประโยคที่จำแนกได้ยากจริงๆแม้จะให้มนุษย์มาจำแนกเองและได้ทำการทดลองนำไปใช้กับรายงานประจำปีทั้งเล่มในหลายๆปีและหลายๆบริษัทแล้วก็ได้พบว่ามีแนวโน้มที่หลายๆบริษัทมักมีเหมือนกันดังที่เห็นในผลลัพธ์จากหัวข้อ4.2

5.2 ปัญหาที่พบและแนวทางการแก้ไข

ในการทำโครงงานนี้ พบว่าเกิดปัญหาหลักๆ ดังนี้

1. การหาวิธีในการจำแนกประโยคในรายงานประจำปีของแต่ละบริษัทเพื่อใช้สำหรับพัฒนาโมเดลจำแนกประโยค
2. หาวิธีเลือกดึงประโยคที่สมบูรณ์เหมาะสำหรับในการใช้กับโมเดลจำแนกประโยคโดยอัตโนมัติ

5.3 ข้อเสนอแนะและแนวทางการพัฒนาต่อ

ข้อเสนอแนะเพื่อพัฒนาโครงงานนี้ต่อไป มีดังนี้

1. ทำให้โปรแกรมสามารถรับมือกับการจำแนกประโยคที่มีหลายประเด็นของ ESG ในหนึ่งประโยคได้
2. พัฒนาระบบที่ใช้กับรายงานฉบับภาษาไทยได้

บรรณานุกรม

[1] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] The Stock Exchange of Thailand. *Sustainability Reporting Guide for Listed Companies*. The Stock Exchange of Thailand, 2022.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[7] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[8] Paolo Torroni, Vanessa da Rocha Santos Almeida, and Nicolaas Ruberg. Bert goes sustainable: an nlp approach to esg financing. 2021.

ภาคผนวก

ภาคผนวก ก

The first appendix

ภาคผนวก ข
คู่มือการใช้งานระบบ

ข.1 คู่มือการดาวน์โหลดโปรแกรม

การใช้งานโปรแกรมขั้นแรก ผู้สนใจทดลองใช้โปรแกรมสามารถค้นหาและดาวน์โหลดโปรแกรมได้ที่ https://github.com/Gravitumn/ESG-classification

ข.2 คู่มือการฝึกสอนโมเดลเพิ่มเติม

ผู้ที่มีความสนใจจะฝึกสอนโมเดลเพิ่มเติมสามารถทำการเรียกใช้ฟังก์ชัน Train_model ซึ่งอยู่ภายในไฟล์ Train.py ได้ โดยมี Arguments ที่สามารถส่งเข้าไปในฟังก์ชันได้ดังนี้

1. bert_model เป็น argument ที่ใช้ในการระบุโมเดลที่ใช้ในการฝึกสอนโดยใช้ file path ในการระบุ หากต้องการใช้โมเดลที่ฝึกสอนโดยผู้จัดทำ สามารถใช้ไฟล์ model/albert2 ซึ่งเตรียมไว้ให้ในโฟลเดอร์ได้เลย แต่หากต้องการใช้โมเดลชนิดอื่นก็สามารถใส่ path ดังตัวอย่างด้านล่างได้เช่นกัน
2. tokenizer เป็นการระบุ tokenizer ที่จะใช้ในการแบ่งประโยคซึ่งโดยปกติจะใช้เป็น tokenizer ของโมเดลนั้นๆ ดังตัวอย่างด้านล่าง
3. device เป็นการระบุอุปกรณ์ที่ใช้ในการฝึกสอน ตามตัวอย่างด้านล่าง
4. file_path ของ dataset ที่เราจัดเตรียมไว้ โดยไม่จำเป็นต้องแบ่ง Train/test เนื่องจาก จะมีให้ระบุเป็น Argument
5. batch_size เป็นพารามิเตอร์ของการฝึกสอนโมเดลที่ช่วยให้โมเดลเรียนรู้ได้ดีมากขึ้นหากปรับได้อย่างเหมาะสม นอกจากนี้การปรับให้มี batch_size ที่สูงเกินไปยังส่งผลให้หน่วยความจำของคอมพิวเตอร์ไม่เพียงพออีกด้วย โดยมี default batch size = 8
6. Shuffle เป็นพารามิเตอร์ที่ใช้ระบุว่าต้องการสับเปลี่ยนการเรียงลำดับของ dataset หรือไม่ซึ่งมีค่า default เป็น True และแนะนำให้ตั้งค่าเป็น True เสมอ
7. lr หรือ learning_rate เป็นพารามิเตอร์ที่ใช้ในการระบุ learning rate ในการเรียนรู้ของคอมพิวเตอร์ มีค่า default = 1e-5
8. num_epochs เป็นพารามิเตอร์ที่ใช้ในการระบุว่าโมเดลจะได้รับการฝึกสอนเป็นจำนวนกี่ครั้ง โดยมีค่า default = 100
9. T_0 เป็นพารามิเตอร์สำหรับ Cosine annealing warm restart โดย T_0 จะเป็นตัวกำหนดว่าจะให้โมเดลทำการเรียนรู้กี่ epoch ก่อนจะเกิดการ restart ซึ่งในระหว่างนั้นจะมีการปรับ learning rate ให้ต่ำลงและเมื่อเกิดการ restart จะทำให้ learning rate กลับมาเท่ากับค่าเริ่มต้น ช่วยให้โมเดลสามารถเรียนรู้ต่อไปได้ มีค่า default= 2
10. T_mult เป็นพารามิเตอร์สำหรับ Cosine annealing warm restart เป็นตัวคูณซึ่งใช้ในการกำหนดจำนวนรอบก่อนจะเริ่ม restart อีกครั้งหนึ่ง ยกตัวอย่างเช่นหาก T_mult = 2 และ T_0 = 10 เมื่อเกิดการ restart จะต้องรออีก 20 ครั้งเพื่อจะ restart และหลังจากนั้นจะเพิ่มเป็น 40 ครั้ง และ 80 ครั้งจนกว่าจะพบเงื่อนไขการจบการฝึกสอน โดยมีค่า default = 2

11. eta_min เป็นพารามิเตอร์สำหรับกำหนดว่า learning_rate สามารถลดลงต่ำสุดได้มากเพียงใด เพื่อไม่ให้ learning_rate ลดต่ำจนเกินไป โดยมีค่า default = 1e-6
12. model_save_path เป็น argument สำหรับระบุ file path ที่จะใช้ในการบันทึกโมเดลเมื่อจบการฝึกสอน โดยมี default path เป็น /model/trained_model
13. early_stop_epoch เป็น argument สำหรับระบุจำนวนรอบสูงสุดที่ต้องการให้โมเดลทำการฝึกสอน โดยมีค่าตั้งต้น = 8 รอบ
14. return_result เป็น argument สำหรับระบุว่าผู้ใช้งานต้องการผลลัพธ์จากการฝึกสอนหรือไม่ โดยผลลัพธ์เหล่านี้จะเป็น confusion matrix ที่แสดงการทำนายประโยคของโมเดล และกราฟ accuracy ในแต่ละ epoch โดยมีค่า default = False
15. test_size เป็น argument สำหรับการทำ traintestsplit โดยมีค่าเริ่มต้น = 0.1 ซึ่งเป็นการระบุว่าใช้ test set 10% ของ dataset
16. random_state เป็น argument สำหรับใช้ระบุว่าจะทำการ traintestsplit ที่ random_state เท่าใด โดยมีค่าตั้งต้นเป็น 69

ประวัติผู้เขียน

นาย สุภาค ไชยเนตรเกษม เกิดเมื่อวันที่ 25 เมษายน 2545 ณ จังหวัดแพร่ สำเร็จการศึกษาระดับมัธยมจากโรงเรียนพะเยาพิทยายม เข้าศึกษาที่ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่ เมื่อปีการศึกษา 2563 โดยมีความสนใจในด้าน Artificial Intelligence และ Natural Language Processing

นาย ธนิสร ไชยวุฒิ เกิดเมื่อวันที่ 12 กุมภาพันธ์ 2545 ณ จังหวัดพะเยา สำเร็จการศึกษาระดับมัธยมจากโรงเรียนพะเยาพิทยายม เข้าศึกษาที่ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่ เมื่อปีการศึกษา 2563 โดยมีความสนใจในด้าน Artificial Intelligence และ Natural Language Processing