รายงานโครงงานวิชา CS653: Big Data Engineering

หัวข้อโครงงาน: ผลิตภัณฑ์และการวิเคราะห์เชิง Sentiment

สมาชิกกลุ่ม

1. ปณิศรา เกษสาคร 6609036063
2. กฤษญา ช้างสอน 6709036062
3. รชต เอกรัตน์ 6709036161

a. ที่มาและความสำคัญ เป้าหมายของโปรเจค

ในยุคที่พฤติกรรมผู้บริโภคเปลี่ยนแปลงอย่างรวดเร็ว การฟังเสียงของลูกค้า (Voice of Customer) กลายเป็นสิ่งจำเป็นสำหรับองค์กรในการพัฒนาสินค้าและบริการให้ตอบสนองความต้องการได้อย่างตรงจุด โดยเฉพาะอย่างยิ่งในยุคดิจิทัลที่ลูกค้าแสดงความเห็นผ่านช่องทางออนไลน์อย่างกว้างขวาง เช่น เว็บบล็อก หรือร้านค้าออนไลน์ การรวบรวมและวิเคราะห์ข้อมูลฟีดแบคเหล่านี้จึงเป็นแหล่งข้อมูลที่ทรงคุณค่า (valuable data source) ที่ช่วยให้องค์กรเข้าใจความต้องการ ประสบการณ์ และความรู้สึกของลูกค้าในระดับเชิงลึก

ด้วยเหตุการพัฒนาระบบที่สามารถรวบรวมความคิดเห็นของลูกค้าจากแหล่งต่างๆ แล้วนำมาวิเคราะห์เพื่อหาข้อมูลเชิงลึก (insight) ที่สามารถนำไปใช้ในการปรับปรุงผลิตภัณฑ์ การบริการ และกลยุทธ์ทางการตลาดให้มีประสิทธิภาพมากยิ่งขึ้นจึงเป็นเรื่องสำคัญ โดยมีเป้าหมาย ดังนี้

1. เพื่อรวบรวมความคิดเห็นของลูกค้าที่มีต่อผลิตภัณฑ์ของบริษัทจากหลากหลายช่องทางออนไลน์ เช่น เว็บบล็อกของบริษัท และร้านค้าออนไลน์อย่าง eBay และ Amazon
2. เพื่อวิเคราะห์และสกัดข้อมูลเชิงลึกจากความคิดเห็นของลูกค้าใน 2 ประเด็นหลัก ได้แก่
    - 2.1 การกล่าวถึงสินค้า (Product Mention): โดยเฉพาะในแหล่งที่ไม่ได้มีการระบุชื่อสินค้าอย่างชัดเจน เช่น เว็บบล็อก จำเป็นต้องใช้เทคนิคการวิเคราะห์ข้อความเพื่อจับคู่กับชื่อสินค้าที่เกี่ยวข้อง
    - 2.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis): เพื่อตรวจสอบทัศนคติของลูกค้าที่มีต่อสินค้าในเชิงบวก กลาง หรือเชิงลบ
3. เพื่อใช้ข้อมูลที่ได้จากการวิเคราะห์ในการปรับปรุงสินค้าและสร้างประสบการณ์เชิงบวกให้กับลูกค้า เพิ่มความภักดีในแบรนด์ และสร้างความได้เปรียบในการแข่งขันทางธุรกิจ

b. ชุดข้อมูลและรายละเอียดของชุดข้อมูลที่จะใช้ ใช้อย่างไรใน high-level

การวิเคราะห์นี้ใช้ 3 ข้อมูลที่สอดคล้องกับการศึกษา Voice of Customer ในช่องทางออนไลน์ ได้แก่

1. ecommerce_reviews (Amazon, eBay)
   - แหล่งที่มา: Kaggle
   - รูปแบบไฟล์: CSV (amazon_reviews.csv และ ebay_reviews.csv)
   - ลักษณะ: Structured + Unstructured
   - ข้อมูลประกอบด้วย: review_id, product_id, review_text, rating, category, timestamp
   - ลักษณะโครงสร้าง: ข้อมูลชุดนี้มีทั้งรูปแบบ Structured และ Unstructured ซึ่งส่วนที่เป็น Structured ประกอบด้วย product_id, rating และ category ซึ่งสามารถใช้วิเคราะห์ได้ทันที และ Unstructured คือ review_text ซึ่งเป็นข้อความข้อคิดเห็นที่ต้องใช้เทคนิค NLP ในการวิเคราะห์ความรู้สึกและความหมาย
   - การใช้งาน : ใช้สำหรับ *Sentiment Analysis* และเชื่อมโยงกับ products_catalog เพื่อระบุรายละเอียดสินค้า
2. products_catalog
   - แหล่งที่มา: Kaggle
   - รูปแบบไฟล์: JSON /products.json
   - Schema: { product_id: [product_name, category_code, price] }
   - ลักษณะโครงสร้าง: Structured เพราะอยู่ในรูปแบบ key-value ที่ชัดเจน และสามารถเข้าถึงข้อมูลผ่านรหัสสินค้าได้โดยตรง
   - การใช้งาน: ใช้สำหรับ enrich ข้อมูลรีวิว และช่วยตรวจจับชื่อสินค้าที่ปรากฏในข้อความบน social media ซึ่งไม่ระบุรหัสสินค้าโดยตรง
3. social_media_posts
   - แหล่งที่มา: Kaggle
   - รูปแบบไฟล์: JSON product_reviews_final.json
   - Schema : { "TextItem": { "review_id": ["ข้อความ"] } }
   - ลักษณะโครงสร้าง: Semi-structured เพราะแม้จะอยู่ในรูปแบบ JSON แต่ไม่มี schema แน่นอน ข้อมูลข้อความถูกแยกเก็บไว้เป็น list และต้องใช้การ parsing และเทคนิค NLP เพื่อดึงเนื้อหาและวิเคราะห์เชิงความหมายได้
   - การใช้งาน: ใช้ตรวจจับการกล่าวถึงสินค้า (Product Mention Detection) และวิเคราะห์อารมณ์ (Sentiment Analysis) จากข้อความที่ไม่อยู่ในระบบ e-commerce

c. การวิเคราะห์คุณลักษณะ 5V ในการประมวลผลชุดข้อมูล

1. Volume (ปริมาณข้อมูล)

ข้อมูลต้นฉบับมีขนาดประมาณ 1.8 GB (กว่า 76 ล้านรายการ) ซึ่งเกินขีดความสามารถของ AWS Learner Lab ที่มีข้อจำกัดด้านทรัพยากรการประมวลผล หากใช้เกินงบประมาณที่กำหนด ระบบจะระงับการใช้งานโดยในครั้งก่อน ทางกลุ่มเคยถูกระงับบัญชีมาแล้วแม้ใช้ไฟล์ที่ลดขนาดเหลือเพียง 20 MB ต่อไฟล์ จึงต้องเริ่มต้นสร้าง Pipeline ใหม่อีกครั้ง จึงต้องลดขนาดข้อมูลจาก Amazon และ eBay ให้เหลือเพียงแหล่งละประมาณ 100 รายการ เพื่อให้เหมาะสมกับการประมวลผล NLP ที่ต้องใช้ทรัพยากรมาก แม้ปริมาณข้อมูลจะลดลง แต่ยังคงเพียงพอสำหรับ

การทดสอบกระบวนการวิเคราะห์ เช่น Sentiment Analysis และ Product Matching นอกจากนี้ไฟล์ products.json ยังคงมีขนาดประมาณ 21 MB ซึ่งครอบคลุมข้อมูลสินค้าที่หลากหลาย ช่วยให้สามารถทดลองรวมข้อมูลกับรีวิวได้อย่างมีประสิทธิภาพในสภาพแวดล้อมจำลอง

ภาพแจ้งเตือนการใช้งานเกิน budget ที่กำหนดที่ทีมได้พบเจอ

2. **Variety (ความหลากหลายของข้อมูล)**

ข้อมูลอยู่ในหลายรูปแบบ ประกอบด้วย .csv (structured) และ .json (semi-structured/structured) ทำให้ต้องออกแบบการรวมข้อมูล (data integration) อย่างระมัดระวัง

3. **Velocity (ความเร็วของข้อมูล)**

การวิเคราะห์นี้ใช้ batch processing ในการรวบรวมข้อมูลและประมวลผลเป็นรอบ โดยแม้จะไม่ได้รับข้อมูลแบบ real-time แต่สามารถประมวลผลใหม่ตามรอบเวลา เช่น รายสัปดาห์หรือรายเดือน เพื่อรักษาความทันสมัยของผลวิเคราะห์ (freshness) และลด latency ของระบบ

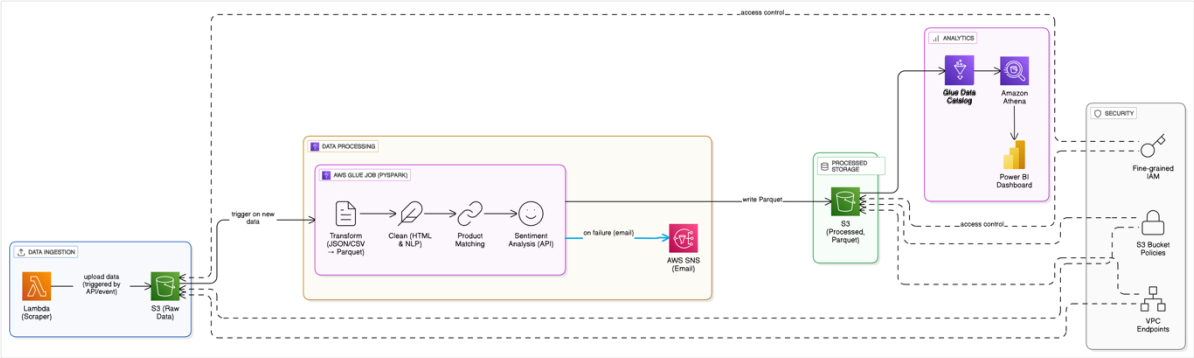
4. **Veracity (ความน่าเชื่อถือของข้อมูล)**

ความคิดเห็นจาก e-commerce มักเชื่อถือได้มากกว่าข้อมูลจาก social media เพราะมีการจัดกลุ่มประเภทจากระบบต้นทางแล้ว แต่ยังคงมีความเสี่ยงจากข้อความเทียม (spam), ความคิดเห็นที่ลำเอียง หรือไม่เกี่ยวข้องอยู่ เปรียบเทียบกับข้อมูลจาก Social media ที่ต้องใช้การกรอง (filtering) และ preprocessing เช่น การทำ keyword matching หรือ การทำ stopword removal

5. **Value (คุณค่าทางธุรกิจ)**

ข้อมูลสามารถนำไปสกัดเชิงลึกได้ เช่น การตรวจจับ Product Mention จากข้อความที่ไม่ระบุรหัสสินค้า ช่วยในการติดตามเทรนด์ได้ และการวิเคราะห์ Sentiment ของลูกค้า สามารถนำไปใช้พัฒนาสินค้าและสร้างประสบการณ์ที่ดีขึ้น

d. สถาปัตยกรรม Data Pipeline พร้อมการประเมินการออกแบบซึ่งอิงจาก AWS Well Architected Framework: Data Analytics Lens



AWS Well-Architected Framework: Data Analytics Lens ประเมินความสามารถของระบบข้อมูลใน 5 มิติหลัก คือ

1. **Operational Excellence**

ระบบถูกออกแบบให้แยกขั้นตอนการทำงานอย่างชัดเจนตั้งแต่การจำลองการดึงข้อมูล จนถึงการแสดงผลบน Dashboard ซึ่งช่วยให้สามารถติดตาม วิเคราะห์ และตรวจสอบได้ง่าย นอกจากนี้ยังมีการแจ้งเตือนเมื่อเกิดข้อผิดพลาดของ Glue Job ผ่าน AWS SNS ทำให้สามารถตอบสนองต่อปัญหาได้อย่างรวดเร็ว อย่างไรก็ตาม ควรมีการเสริมการเก็บ Log แบบรวมศูนย์ผ่าน CloudWatch เพื่อเพิ่มประสิทธิภาพในการดูแลระบบ

2. **Security**

ระบบมีแนวทางความปลอดภัยที่เหมาะสมในระดับโครงสร้าง เช่น การแยกเก็บข้อมูลเป็นขั้นตอนใน S3 เพื่อควบคุมการเข้าถึง แต่สามารถเพิ่มการติดตาม Logging บน S3 และ Athena เพื่อการตรวจสอบย้อนหลัง และหากข้อมูลมีลักษณะเป็นข้อมูลส่วนบุคคล ควรพิจารณาการปกปิดข้อมูลตามหลัก PDPA/GDPR เพิ่มเติม

3. **Reliability**

การแบ่งผลลัพธ์ของแต่ละขั้นใน S3 และการใช้ Glue Job ที่เป็นอิสระต่อกัน ช่วยให้สามารถ rerun เฉพาะขั้นที่ผิดพลาดได้ โดยไม่กระทบทั้ง pipeline อีกทั้งการแปลงข้อมูลและจัดการ schema ทำให้ระบบมีความยืดหยุ่นต่อการเปลี่ยนแปลงของข้อมูล อย่างไรก็ตาม การเพิ่มการตรวจสอบข้อมูลนำเข้า (Input Validation) และการใช้ Glue Bookmark จะช่วยลดความเสี่ยงจากข้อมูลซ้ำซ้อนหรือข้อมูลตกหล่น

4. **Performance Efficiency**

การเลือกใช้บริการ serverless เช่น AWS Lambda, Glue และ Athena ทำให้ระบบสามารถปรับตัวตามภาระงานได้อย่างเหมาะสม โดยเฉพาะการใช้ Parquet format ซึ่งช่วยลดปริมาณข้อมูลที่สแกนและเพิ่มความเร็วในการสืบค้นข้อมูล อีกทั้งยังมีการแยกข้อมูลตามขั้นตอน ซึ่งช่วยให้ประสิทธิภาพโดยรวมของระบบสูงขึ้น

5. **Cost Optimization**

ระบบใช้โมเดลแบบ pay-per-use ที่เหมาะสมกับการควบคุมต้นทุน เช่น การคิดค่าบริการตามจำนวน query ใน Athena และเวลาใช้งานของ Glue Job อย่างไรก็ตาม ควรเสริมการตั้ง Budget, การแจ้งเตือนการใช้งบผ่าน AWS Budgets และการตรวจสอบผ่าน Cost Explorer เพื่อป้องกันค่าใช้จ่ายเกินความจำเป็น โดยเฉพาะในขั้นตอนที่มีการเรียกใช้งาน ML API หรือการ query ข้อมูลขนาดใหญ่

e. รายละเอียดของ Data Pipeline Implementation ในแต่ละขั้นตอน

1. **Lambda (Scraper)**

ใช้ AWS Lambda จำลองการดึงข้อมูลจากภายนอก โดยทำการอัปโหลดไฟล์ข้อมูล JSON และ CSV ที่จัดเตรียมไว้ล่วงหน้า (จาก Kaggle) ขึ้นไปยัง Amazon S3 เพื่อใช้ในขั้นตอนถัดไป

2. **จัดเก็บไฟล์ต้นฉบับลงใน S3**

ข้อมูลต้นฉบับจาก Lambda ถูกจัดเก็บไว้ใน S3 ภายใต้โฟลเดอร์ /raw/ โดยแยกตามประเภทข้อมูล เช่น social media หรือ e-commerce เพื่อความเป็นระเบียบและสะดวกต่อการเข้าถึง

3. **ประมวลผลด้วย AWS Glue Job (PySpark)**

ใช้ AWS Glue Job ประมวลผลข้อมูลหลายขั้นตอน เช่น แปลงไฟล์เป็น Parquet, ทำความสะอาดข้อความ (NLP), จับคู่ข้อความกับสินค้า, วิเคราะห์อารมณ์ด้วยโมเดล Sentiment Analysis ผ่าน API โดยแต่ละ Job มีการแจ้งเตือนผ่าน SNS หากเกิดข้อผิดพลาด

4. **บันทึกผลลัพธ์ลง S3**

ผลลัพธ์จากการประมวลผลจะถูกบันทึกใน S3 ในรูปแบบไฟล์ Parquet แยกโฟลเดอร์ตามขั้นตอน เพื่อให้ง่ายต่อการตรวจสอบและเชื่อมต่อกับบริการอื่น

5. **วิเคราะห์ข้อมูลด้วย Amazon Athena**

ใช้ Amazon Athena ร่วมกับ Glue Data Catalog ในการ query ข้อมูลจาก S3 ด้วย SQL เพื่อวิเคราะห์และรวมข้อมูลจากหลายแหล่งเข้าสู่รูปแบบที่พร้อมใช้งาน

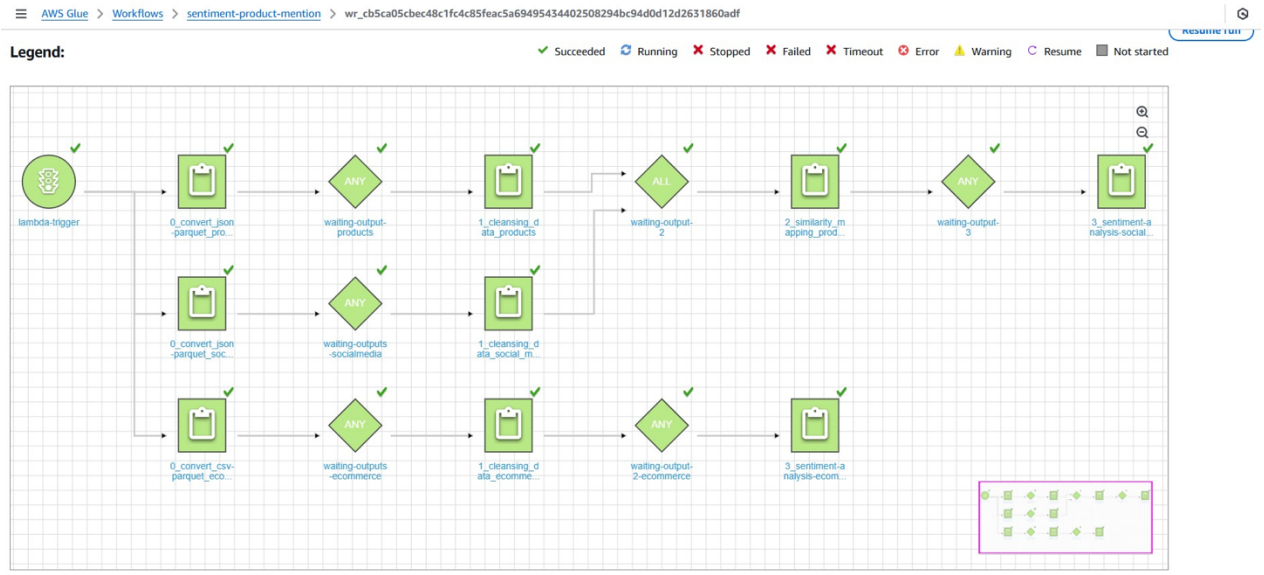
6. **สร้าง Dashboard ด้วย Power BI**

Power BI เชื่อมต่อกับ Athena เพื่อแสดงข้อมูลที่ผ่านการวิเคราะห์แล้วในรูปแบบ Dashboard ช่วยให้ผู้ใช้งานเข้าใจข้อมูลเชิงลึกได้ง่ายและสะดวกขึ้น เช่น ความคิดเห็นต่อสินค้า หรือแนวโน้มในแต่ละช่วงเวลา

f. ผลลัพธ์ที่ได้ อภิปรายและสรุปผล พร้อมภาพแสดงตัวอย่าง Data Pipeline ที่ Validate และ Deploy แล้ว

จากการออกแบบและพัฒนา Data Pipeline พบว่าสามารถดำเนินกระบวนการได้ครบถ้วน โดยเริ่มจากการนำเข้าข้อมูลรีวิวจาก Amazon, eBay และโซเชียลมีเดีย ผ่านการจำลองด้วย AWS Lambda และจัดเก็บลง S3 โดยแยกตามประเภทข้อมูลอย่างเป็นระบบ โดยข้อมูลเหล่านี้จะถูกประมวลผลด้วย AWS Glue Job ซึ่งทำหน้าที่แปลงข้อมูลเป็นไฟล์ Parquet, ทำความสะอาดข้อความด้วยเทคนิค NLP, จับคู่ข้อความกับสินค้า และวิเคราะห์ความคิดเห็นด้วย Sentiment Analysis ผลลัพธ์แต่ละขั้นถูกจัดเก็บแยกใน S3 และสามารถเรียกดูผ่าน Amazon Athena ที่เชื่อมโยงกับ Glue Data Catalog และในขั้นตอนสุดท้ายจะใช้ Power BI สำหรับแสดงผลในรูปแบบ Dashboard ทำให้สามารถเห็นภาพรวมของความคิดเห็นลูกค้าและการกล่าวถึงสินค้าได้อย่างชัดเจน

จากการดำเนินการของ Pipeline แสดงให้เห็นว่าสามารถนำแนวคิดและวิธีการไปใช้ในระบบวิเคราะห์ข้อมูลจริงได้อย่างมีประสิทธิภาพ สามารถมองเห็น insight สำคัญ เช่น สินค้าใดได้รับความคิดเห็นเชิงลบมากที่สุด หรือแพลตฟอร์มใดที่มีเสียงสะท้อนเชิงลบสูงได้ ซึ่งช่วยให้วิเคราะห์ปัจจัยที่ส่งผลต่อประสบการณ์ผู้ใช้ทั้งทางตรงและทางอ้อม และนำไปสู่การปรับปรุงสินค้าและบริการได้อย่างมีประสิทธิภาพ

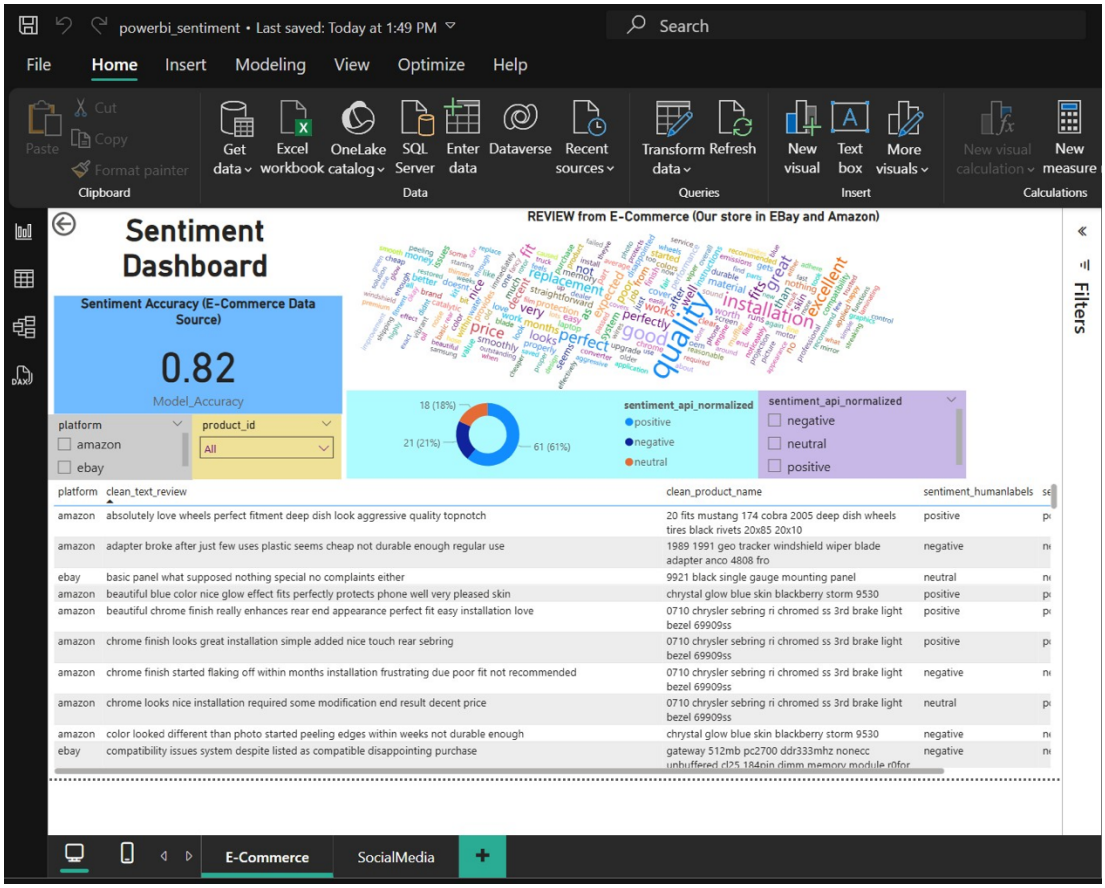


ภาพ Data Pipelineที่ Validate และ Deploy แล้ว

ผลลัพธ์จากการพัฒนา Data Pipeline สามารถนำไปใช้ในขั้นตอนการวิเคราะห์ข้อมูลและการแสดงผล โดยเฉพาะกับข้อมูลจาก E-Commerce และโซเชียลมีเดีย โดยในกรณีของ E-Commerce ระบบมีข้อมูลความคิดเห็นที่ถูกกำกับด้วย Human label อยู่แล้ว ทำให้สามารถเปรียบเทียบผลลัพธ์จากโมเดล Sentiment API กับข้อมูลจริง เพื่อประเมินความแม่นยำของโมเดล (Model Accuracy) ได้อย่างชัดเจน ซึ่งช่วยในการ validate ประสิทธิภาพของโมเดล

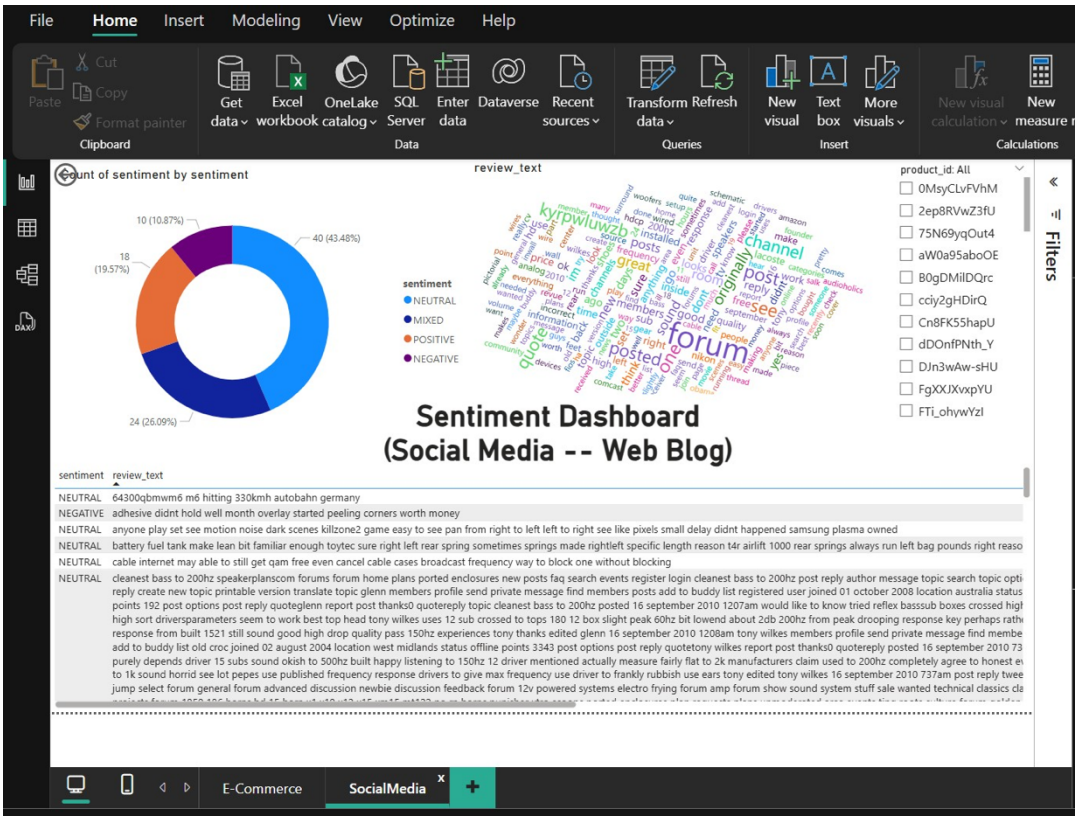
แต่ในทางกลับกัน ข้อมูลจากโซเชียลมีเดียไม่มี label กำกับไว้ล่วงหน้า เราจึงไม่สามารถรู้ได้ว่าแต่ละข้อความควรจัดอยู่ในประเภทใด (เช่น บวก ลบ หรือกลาง) ทำให้ไม่สามารถวัดความแม่นยำของโมเดลได้โดยตรงเหมือนกับข้อมูล E-commerce ดังนั้น เราจึงต้องใช้วิธีอื่นแทน เช่น การอ่านตัวอย่างข้อความบางส่วนด้วยตนเอง (การสุ่มตรวจสอบ) หรือการประเมินภาพรวมของผลลัพธ์ที่โมเดลวิเคราะห์ออกมา ว่าสอดคล้องกับความเป็นจริงหรือไม่ วิธีเหล่านี้เรียกว่าเป็นการวิเคราะห์เชิงคุณภาพ ซึ่งแม้จะให้แนวทางได้ดีในระดับหนึ่ง แต่ไม่สามารถวัดความถูกต้องเป็นตัวเลขได้ชัดเจนเหมือนกรณีที่มีข้อมูล label ครบถ้วน

ความแตกต่างนี้สะท้อนให้เห็นว่า การมีข้อมูลอ้างอิงที่ถูกต้องและมี label ที่เหมาะสม ช่วยให้สามารถวัดผลและปรับปรุงโมเดลได้อย่างเป็นระบบและแม่นยำ ซึ่งเป็นองค์ประกอบสำคัญต่อการพัฒนาโมเดลให้มีประสิทธิภาพในระยะยาว



platform	clean_text_review	clean_product_name	sentiment_humanlabels
amazon	absolutely love wheels perfect fitment deep dish look aggressive quality topnotch	20 fits mustang 174 cobra 2005 deep dish wheels tires black rivets 20x85 20x10	positive
amazon	adapter broke after just few uses plastic seems cheap not durable enough regular use	1989 1991 geo tracker windshield wiper blade adapter anco 4808 fro	negative
ebay	basic panel what supposed nothing special no complaints either	9921 black single gauge mounting panel	neutral
amazon	beautiful blue color nice glow effect fits perfectly protects phone well very pleased skin	chrystal glow blue skin blackberry storm 9530	positive
amazon	beautiful chrome finish really enhances rear end appearance perfect fit easy installation love	0710 chrysler sebring ri chromed ss 3rd brake light bezel 69909ss	positive
amazon	chrome finish looks great installation simple added nice touch rear sebring	0710 chrysler sebring ri chromed ss 3rd brake light bezel 69909ss	positive
amazon	chrome finish started flaking off within months installation frustrating due poor fit not recommended	0710 chrysler sebring ri chromed ss 3rd brake light bezel 69909ss	negative
amazon	chrome looks nice installation required some modification end result decent price	0710 chrysler sebring ri chromed ss 3rd brake light bezel 69909ss	neutral
amazon	color looked different than photo started peeling edges within weeks not durable enough	chrystal glow blue skin blackberry storm 9530	negative
ebay	compatibility issues system despite listed as compatible disappointing purchase	gateway 512mb pc2700 ddr333mhz nonecc	negative

ภาพการ Visualization ผลลัพธ์ของ Data Analysis ข้อมูลจาก E-Commerce



ภาพการ Visualization ผลลัพธ์ของ Data Analysis ข้อมูลจาก E-Commerce

g. อภิปรายสิ่งที่ได้เรียนรู้และแนวทางในการพัฒนาต่อยอด

จากการทำโครงงานนี้ได้เรียนรู้กระบวนการออกแบบ Data Pipeline บน AWS อย่างครบถ้วน ตั้งแต่การจัดเก็บข้อมูล การประมวลผลด้วย AWS Glue การวิเคราะห์ผ่าน Amazon Athena และการนำเสนอผลด้วย Power BI รวมถึงการประยุกต์ใช้เทคนิค NLP เช่น Sentiment Analysis และ Product Matching ในการวิเคราะห์ความคิดเห็นของลูกค้าจากหลากหลายแพลตฟอร์ม ในส่วนของแนวทางในการพัฒนาต่อยอดในอนาคต ได้แก่ การปรับระบบให้รองรับการประมวลผลแบบ Real-Time และการพัฒนาโมเดลวิเคราะห์ความคิดเห็นให้มีความแม่นยำยิ่งขึ้น เพื่อให้สามารถตอบสนองต่อความคิดเห็นของลูกค้าได้อย่างทันท่วงทีและมีประสิทธิภาพมากยิ่งขึ้น

h. Github Repository

https://github.com/supremerryDS/Sentiment-Product-Mention