การวิเคราะห์ขั้วอารมณ์แบบผสมจากข่าวหลักทรัพย์ในประเทศไทย

MIXED SENTIMENT ANALYSIS FOR STOCK NEWS IN THAILAND

อาจารย์ที่ปรึกษา รศ.ดร.อัณณ์นุพันธ์ รอดทุกข์ ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยรามคำแหง
ผู้เชี่ยวชาญด้านภาษาศาสตร์ ผศ.ดร.อุมาภรณ์ สังขมาน ภาควิชาภาษาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
นำเสนอโดย 6405003317 นางสาวจุฑารัตน์ ภู่อมร และ 6205002246 นางสาวนฤมล คงโอ

AGENDA
ALWAYS

Introduction 1

Theory 2

Methodology 3

Results 4

INTRODUCTION

ที่มา

ประเทศไทยมีระบบเศรษฐกิจแบบผสมที่รัฐและเอกชนร่วมกันขับเคลื่อน โดยเอกชนมีบทบาทสำคัญในการลงทุนและระดมทุนผ่านตลาดหลักทรัพย์ ตามกฎเกณฑ์จาก ก.ล.ต. และ ตลท. การระดมทุนได้รับอิทธิพลจากหลายปัจจัย เช่น

- ปัจจัยพื้นฐาน (รายได้ กำไร เงินปันผล)
- ปัจจัยทางเทคนิค (ปริมาณการซื้อขาย ภาวะตลาด)
- ข่าวสารทั้งในและต่างประเทศทั้งทางบวก และ ลบ

ความสำคัญของปัญหา

เพื่อช่วยให้นักลงทุนสามารถเข้าใจอารมณ์ของข่าวสารที่มีปริมาณมากและซับซ้อน จึงต้องมีการนำ วิธีการวิเคราะห์ความรู้สึก (Sentiment Analysis) บนพื้นฐานของ การประมวลผลภาษาธรรมชาติ (NLP) มาใช้ ลดความเสี่ยงจากการตัดสินใจที่ใช้อารมณ์ของข้อความ ทำให้การวิเคราะห์ตลาดหุ้นและการลงทุนมีความแม่นยำ โปร่งใส และสนับสนุนการตัดสินใจเชิงข้อมูลมากขึ้น โดยมีการดำเนินการ ดังนี้

- จำแนกข้อความเชิงอารมณ์เดี่ยว (single sentiment) คือ บวก ลบ หรือกลาง
- จัดการกับข้อความที่มี หลายขั้วอารมณ์ (mixed sentiment) ซึ่งเป็นความท้าทายสูง

วัตถุประสงค์ที่ได้รับจากงานวิจัย

1. สร้างโมเดลที่มีประสิทธิภาพสามารถจำแนกความรู้สึกจากข่าวสารการลงทุนในตลาดหุ้น โดยใช้หลากหลายเทคนิคการเรียนรู้ สร้างโมเดลการจำแนกจากชุดฝึกข้อมูลที่มีความหลากหลายและซับซ้อน

2. โมเดลการเรียนรู้ที่มีประสิทธิภาพในการจำแนกความรู้สึกเชิงผสมของข้อมูล จะเป็นเครื่องมือทางเลือกที่ช่วยสนับสนุนการตัดสินใจของนักลงทุน หรือบุคคลทั่วไปที่มีความสนใจการลงทุนประเภท หลักทรัพย์

THEORY

ข้อมูลนำเข้า (Input Data)

ข้อมูลข่าวสารการเงินและบทวิเคราะห์ตลาดหุ้น

ตัวอย่าง:

- “วันนี้ SETHD ปิดที่ระดับ 1,160.56 จุด ลดลง 9.16 จุด หรือ 0.78 %”
- “เช้าวันนี้ SETCLMV ปิดที่ระดับ 632.69 จุด เพิ่มขึ้น 0.55 จุด หรือ 0.09 %”
- “สรุปภาวะตลาดต่างประเทศ ประจำวันที่ 23 ต.ค.67”

การตัดคำ (TOKENIZER)

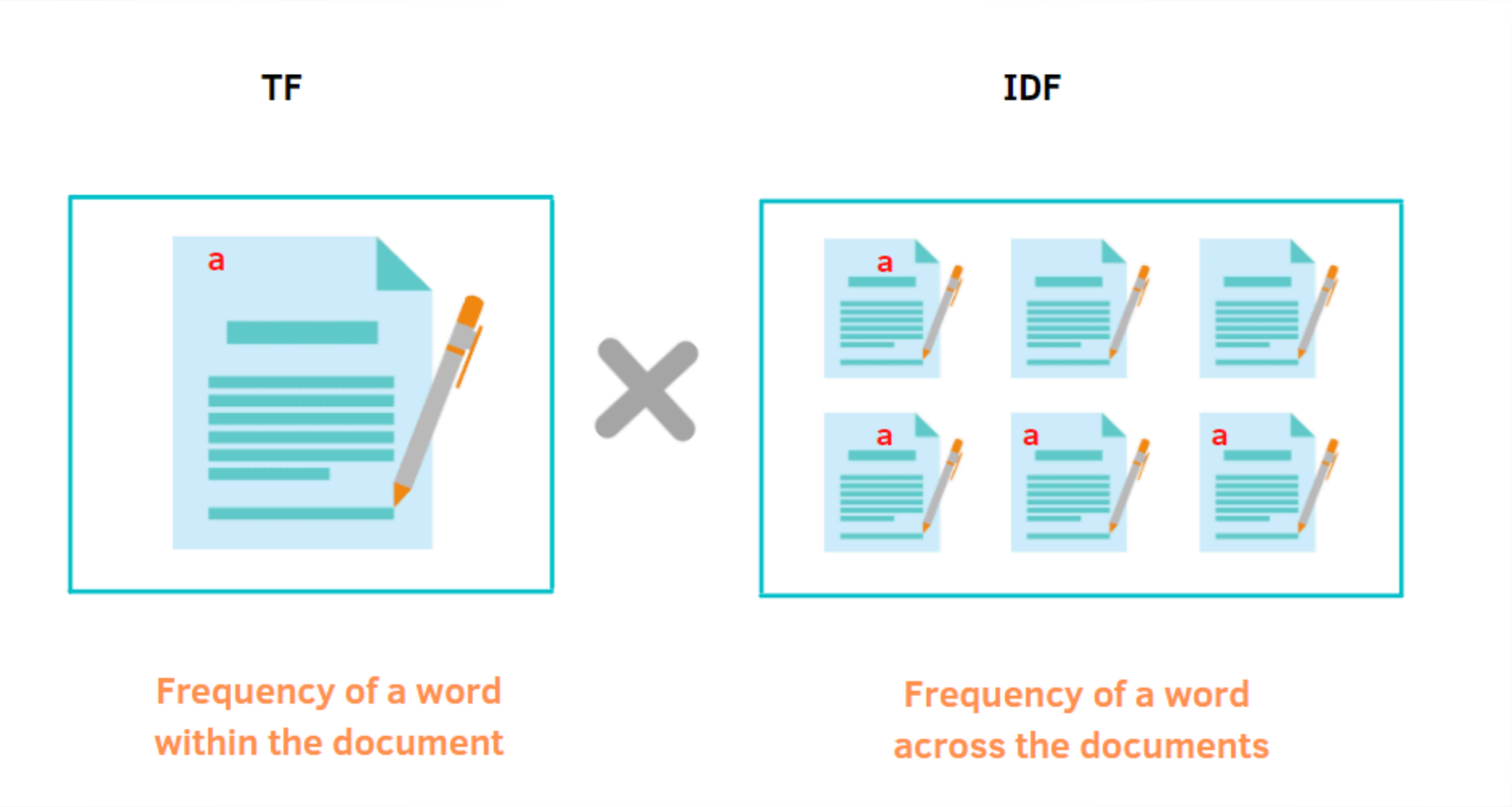
การตัดคำภาษาไทยด้วย PyThaiNLP ทำงานโดยใช้ค้นหาคำที่ยาวที่สุด ร่วมกับพจนานุกรมคำศัพท์ภาษาไทยขนาดใหญ่ เรียกเทคนิคการตัดคำชนิดนี้ว่า newmm

ตัวอย่าง:

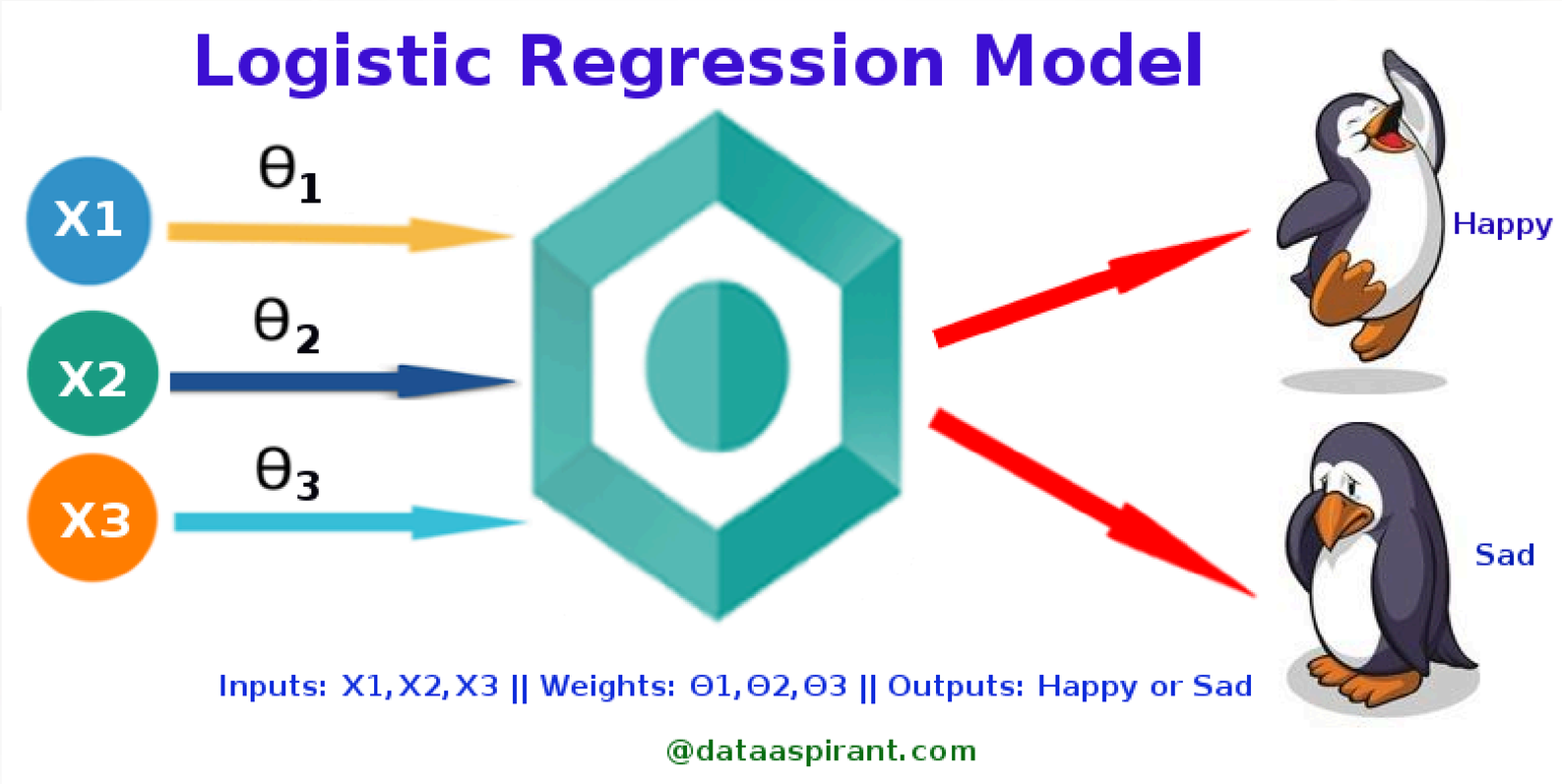
ข้อความ : "กทม.เป็นเมืองที่มีสถานที่ท่องเที่ยวมากมาย"

ผลลัพธ์ : ['กทม.', 'เป็น', 'เมือง', 'ที่', 'มี', 'สถานที่', 'ท่องเที่ยว', 'มากมาย']

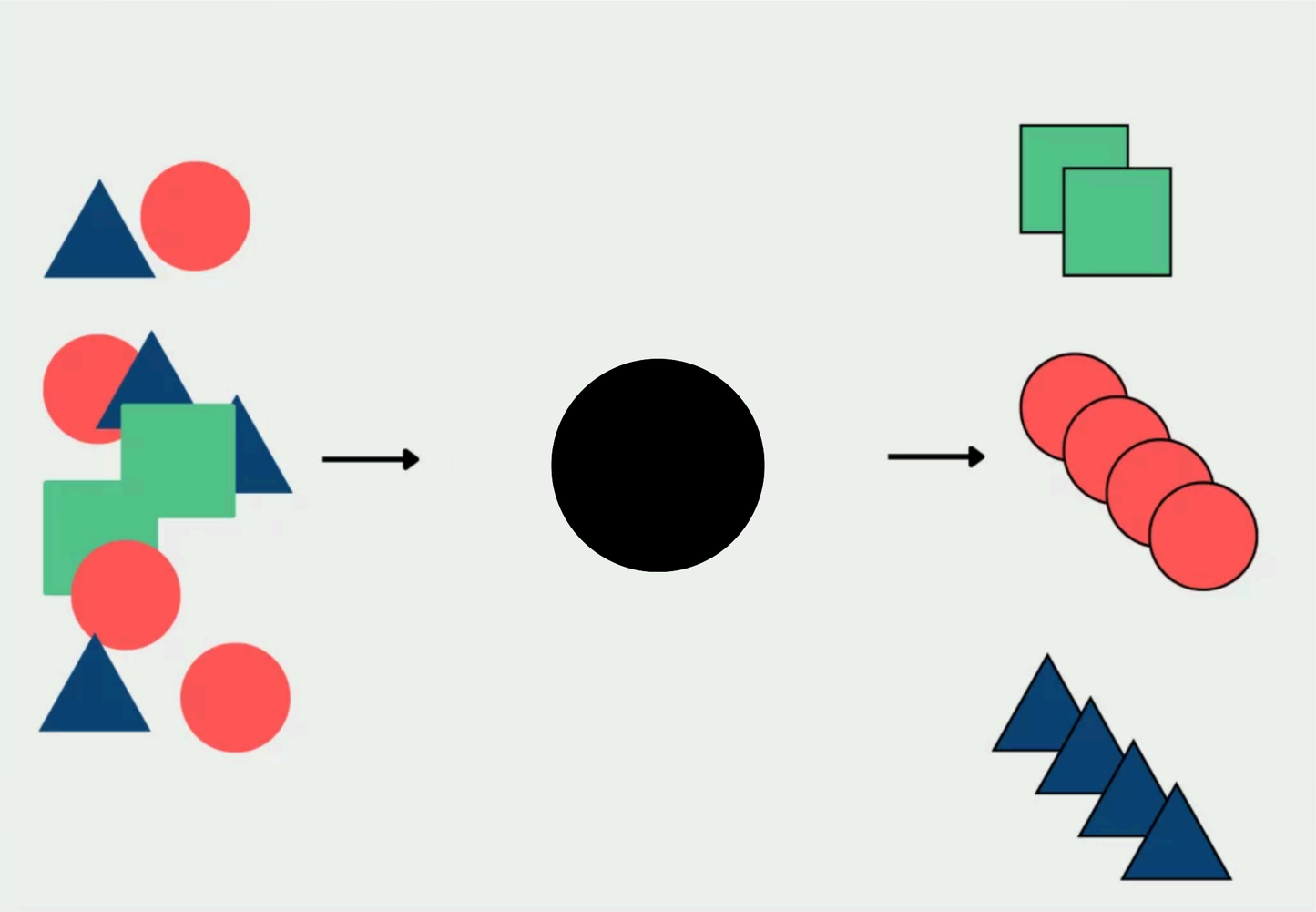
TF - IDF


TF
IDF
a
a
a
a
a
Frequency of a word
within the document
Frequency of a word
across the documents

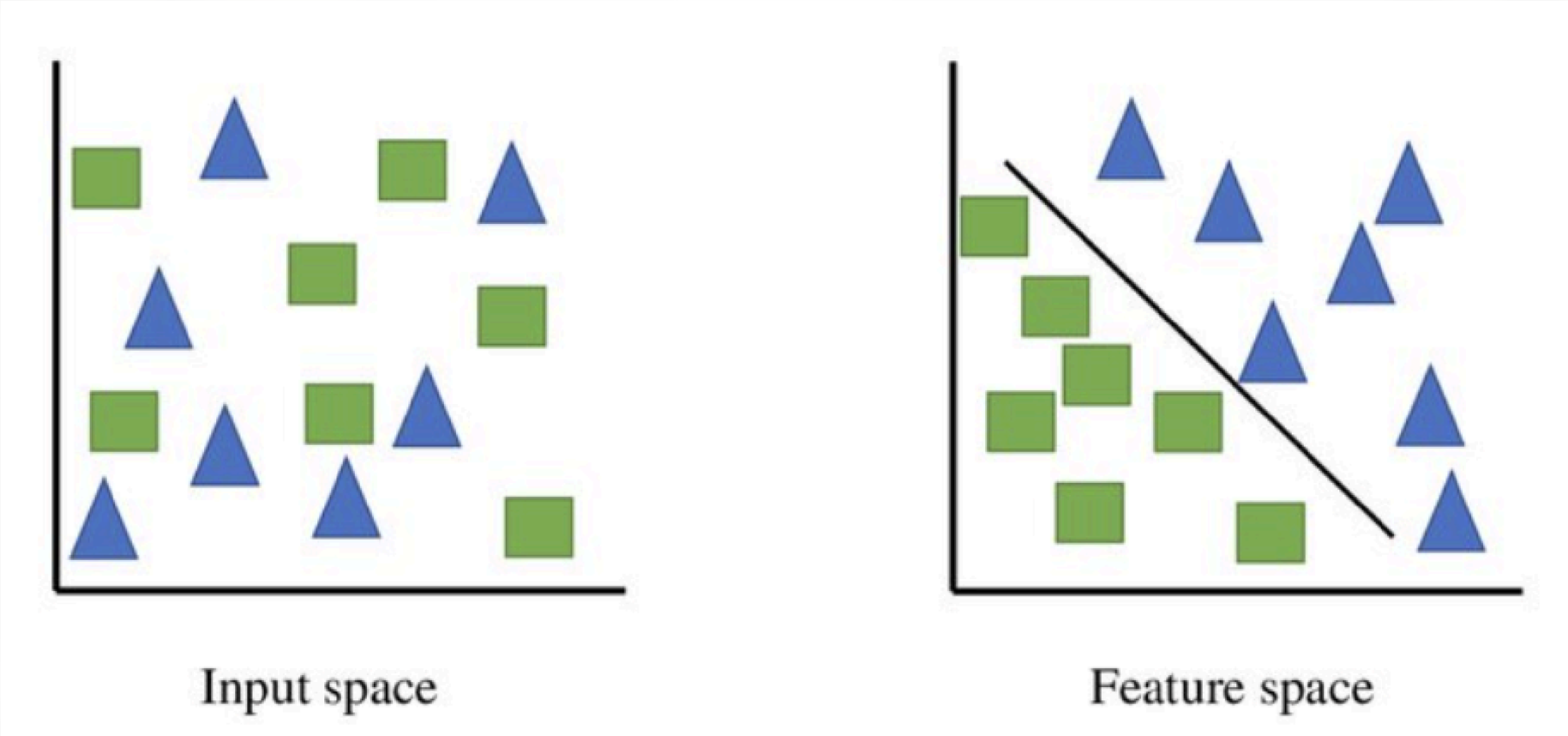
LOGISTIC REGRESSION



NAIVE BAYES



SUPPORT VECTOR MACHINE



ทำไม WANGCHANBERTA คือ ทางเลือกที่ดีที่สุดสำหรับ THAINLP

คุณสมบัติ	โมเดลสถิติ LR , NB , SVM	WangchanBERTa
การเตรียมข้อมูล	ใช้ TF-IDF เพื่อวัดความสำคัญของแต่ละคำในการจำแนกข้อความ	โมเดลเรียนรู้ด้วยตัวเอง จากการฝึกฝน (Pre-training) ไม่ทำ Feature Engineering อย่าง TF-IDF
การทำงาน	ใช้กฎทางคณิตศาสตร์และสถิติ ในการจำแนกข้อความ	ใช้โครงข่ายประสาทเทียม เรียนรู้จากชุดข้อมูลขนาดใหญ่ ทำให้เข้าใจภาษาไทยใกล้เคียงกับมนุษย์และสามารถรับรู้บริบทที่ละเอียดอ่อนได้
ประสิทธิภาพ	จำแนกข้อความได้ดีกับข้อมูลที่ไม่ซับซ้อนและมีขนาดเล็กถึงปานกลาง ใช้ทรัพยากรการประมวลผลน้อยกว่า แต่ความแม่นยำอาจมีข้อจำกัดเมื่อข้อมูลมีความซับซ้อน	ให้ ประสิทธิภาพที่เหนือกว่า สำหรับงาน NLP ที่ซับซ้อน โดยเฉพาะอย่างยิ่งในข้อความที่มีความหมายกำกวม (Mixed Sentiment) และมีความยืดหยุ่นสูง สามารถนำไป Fine-tune เพื่อใช้งานที่หลากหลายได้ แม้จะใช้ทรัพยากรในการประมวลผลสูงมากก็ตาม

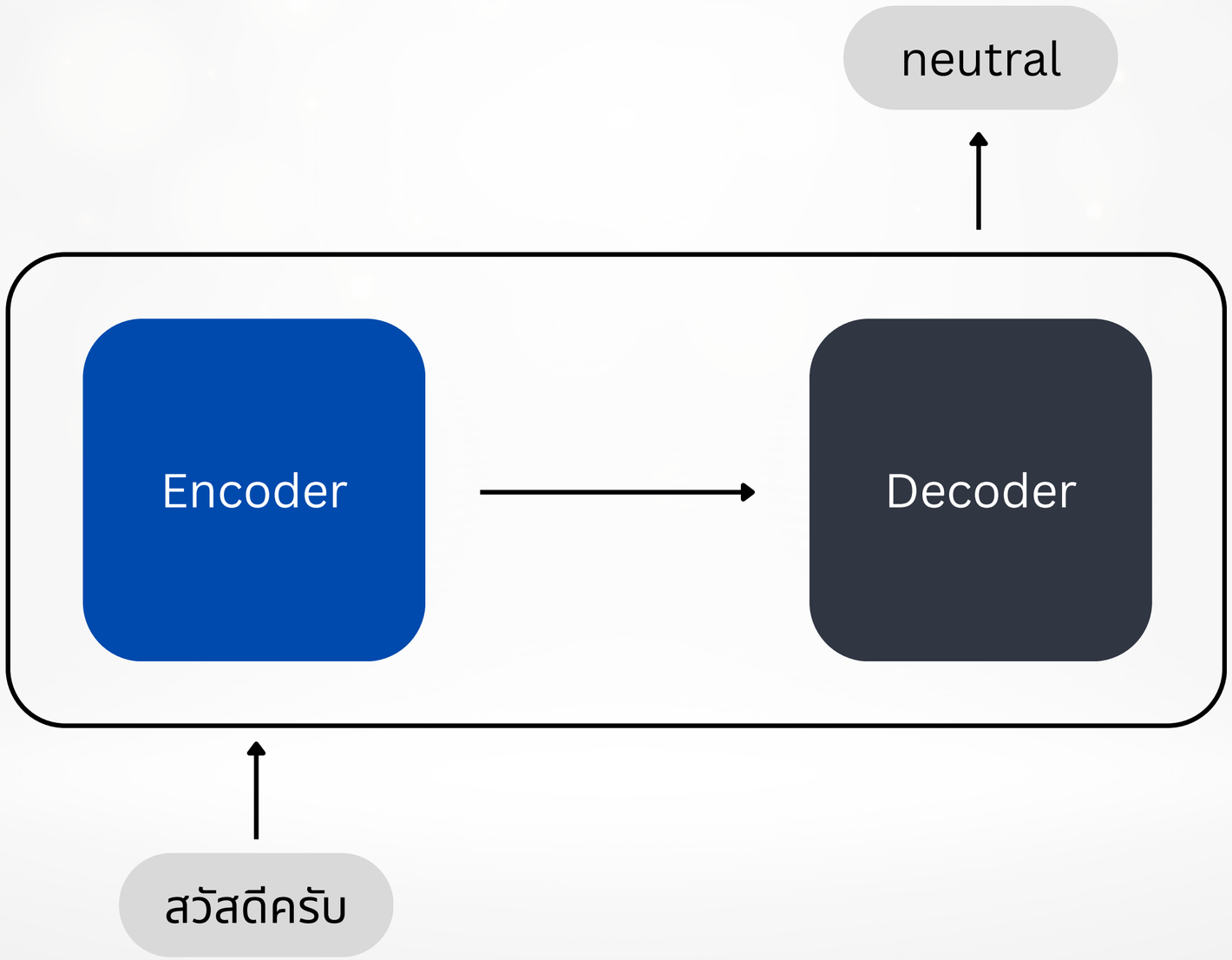
WANGCHANBERTA

Text	Label
สวัสดีครับ	neutral
รวยจัดพี่สุดยอดมากครับ	positive
กากชิบหาย	negative

→ **WangchanBERTa** →

Text	Label
สวัสดีครับ	neutral
รวยจัดพี่สุดยอดมากครับ	positive
กากชิบหาย	negative

TRANSFORMER

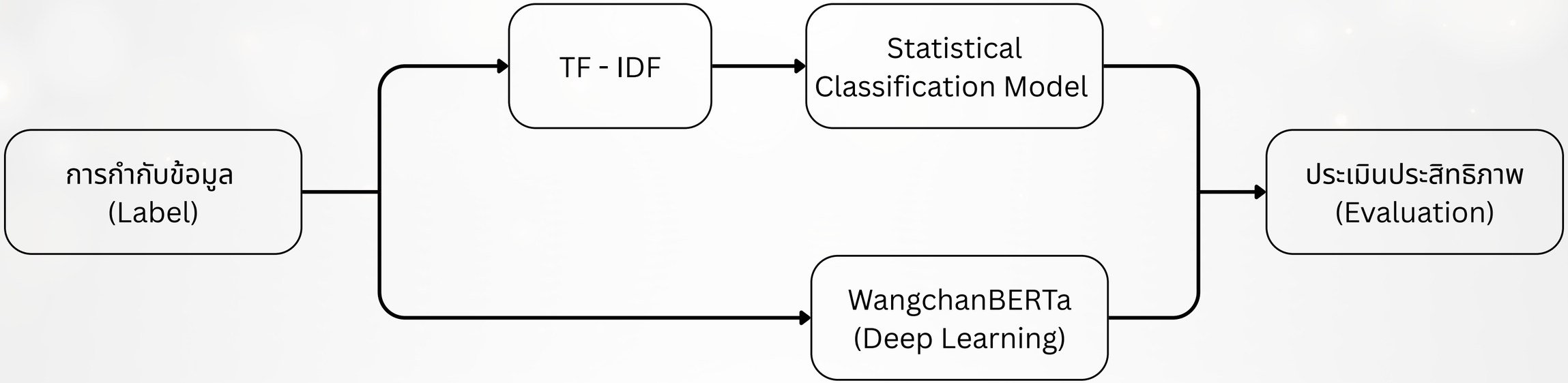


TOKENIZER

หน่วยคำ (Word)	วันนี้|ฉัน|สั่ง|กิวด้ง|มา|ทาน|ที่|บ้าน
หน่วยคำย่อย (Subword)	วันนี้|ฉัน|สั่ง|กิ|วด|◌้ง|มา|ทาน|ที่บ้าน (SentencePiece; XLMR)
หน่วยพยางค์ (Syllable)	วัน|นี้|ฉัน|สั่ง|กิว|ด้ง|มา|ทาน|ที่|บ้าน
หน่วยตัวอักษร (Character)	ว|◌ั|น|น|◌ี|◌้|ฉ|◌ั|น|ส|◌ั|◌่|ง|กิ|ว|ด|◌้|ง|ม|า|ท|า|น|ที่|◌ี|◌่|บ|◌้|า|น

METHODOLOGY

การออกแบบกระบวนการดำเนินการวิจัย



การออกแบบชุดฝึกการทดลอง

1. ข้อความที่แสดงความรู้สึกเดียว (Single Sentiment)

เชิงบวก (Positive)

เชิงลบ (Negative)

เป็นกลาง (Neutral)

ตัวอย่าง:

- “วันนี้ SETHD ปิดที่ระดับ 1,160.56 จุด ลดลง 9.16 จุด หรือ 0.78 %” (เชิงลบ)
- “เช้าวันนี้ SETCLMV ปิดที่ระดับ 632.69 จุด เพิ่มขึ้น 0.55 จุด หรือ 0.09 %” (เชิงบวก)
- “สรุปภาวะตลาดต่างประเทศ ประจำวันที่ 23 ต.ค.67” (เชิงกลาง)

การออกแบบชุดฝึกการทดลอง

2.ข้อความที่แสดงความรู้สึกผสม (Mixed Sentiments)

เชิงบวก (Positive)

เชิงลบ (Negative)

เป็นกลาง (Neutral)

ข้อความที่แสดงความรู้สึกผสม (Mixed Sentiment) ที่ผ่านการจัดกลุ่ม 6 รูปแบบ จากผู้เชี่ยวชาญ

1. Positive + คำเชื่อม + Negative → Class = Positive
ตัวอย่าง: “บริษัทมีกำไรเพิ่มขึ้น แต่มีหนี้สินสูงขึ้นเล็กน้อย”
2. Positive + คำเชื่อม + Negative → Class = Negative
ตัวอย่าง: “รายได้รวมเติบโต แต่กำไรลดลงมาก”
3. Negative + คำเชื่อม + Positive → Class = Negative
ตัวอย่าง: “บริษัทขาดทุนต่อเนื่อง แม้ว่ายอดขายจะเริ่มฟื้นตัว”
4. Negative + คำเชื่อม + Positive → Class = Positive
ตัวอย่าง: “บริษัทมีหนี้สูง แต่สามารถทำกำไรเพิ่มขึ้นได้”
5. Positive ติด Negative → Class = Positive
ตัวอย่าง: “ผลประกอบการของบริษัทในไตรมาสนี้กำไรลดลง กำไรลดลง” (positive word + negative word)
6. Negative ติด Positive → Class = Negative
ตัวอย่าง: “ผลประกอบการของบริษัทในไตรมาสนี้ขาดทุนเพิ่มขึ้น” (negative word + positive word)

ชุดข้อมูล

ประเภทชุดข้อมูล	จำนวนคลาส	รูปแบบคำผสม	รายละเอียด
1	2	เดี่ยว	ประโยคที่แสดงความรู้สึกเดียว ไม่มีข้อความเชิงกลาง
2	2	ผสม (การผสมคำ 4 รูปแบบ)	ประโยคที่แสดงความรู้สึกเดียว และ ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล
3	2	ผสม (การผสมคำ 6 รูปแบบ)	ประโยคที่แสดงความรู้สึกเดียว และ ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล
4	3	เดี่ยว	ประโยคที่แสดงความรู้สึกเดียว มีข้อความเชิงบวก เชิงลบ เชิงกลาง
5	3	ผสม (การผสมคำ 6 รูปแบบ)	ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล

WangchanBERTa Configuration

```
eval_strategy=12,
save_strategy=12,
learning_rate=2e-5,
per_device_train_batch_size=8,
per_device_eval_batch_size=8,
num_train_epochs=12,
logging_steps=50,
load_best_model_at_end=True,
metric_for_best_model="accuracy",
greater_is_better=True,
weight_decay=0.05,
gradient_accumulation_steps=2,
lr_scheduler_type="linear",
warmup_steps=500,
```

RESULT

ตัวชี้วัดประสิทธิภาพของการฝึกสอนโมเดล

- **Accuracy** เป็นการวัดการทำนายที่ถูกต้องทั้งหมด ไม่ว่าจะเป็นผลถูกต้องหรือไม่ถูกต้อง หารด้วยจำนวนทั้งหมดของการทำนาย

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** เป็นการวัดความแม่นยำ ในการระบุค่าที่สนใจ (เช่น โมเดลทำนายว่าเป็นข้อความเชิงบวก มักจะถูกต้อง)

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** เป็นการวัดความถูกต้องว่าโมเดลสามารถตรวจจับค่าที่สนใจ และมีโอกาสที่จะไม่พลาดกรณีที่เป็น "ค่าที่สนใจ" จริง ๆ

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score** ค่าที่ใช้เพื่อหาสมดุลระหว่าง Precision และ Recall โดยคำนวณจากค่าเฉลี่ยของสองค่าดังกล่าวเพื่อประเมินผลการทำนายโดยรวม

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

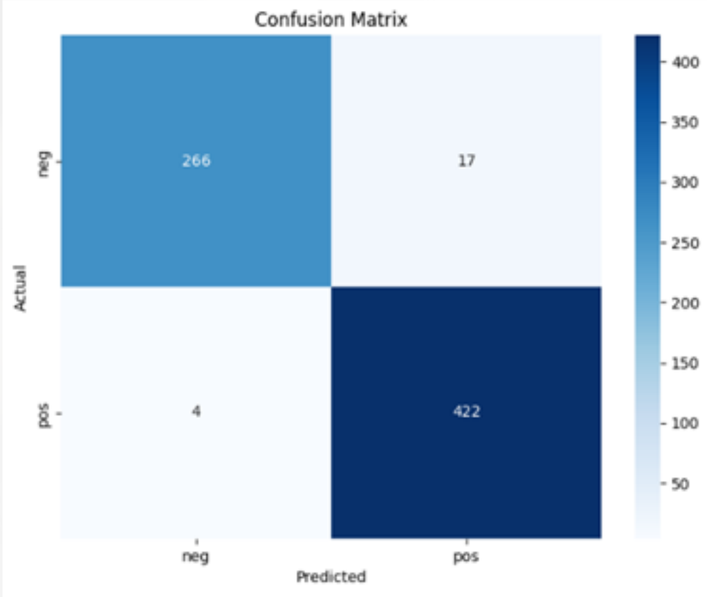
ประเภทชุดข้อมูล : ชุดที่ 1 **รูปแบบคำผสม : เดี่ยว** **จำนวนคลาส : 2**

รายละเอียด : ประโยคที่แสดงความรู้สึกเดียว ไม่มีข้อความเชิงกลาง

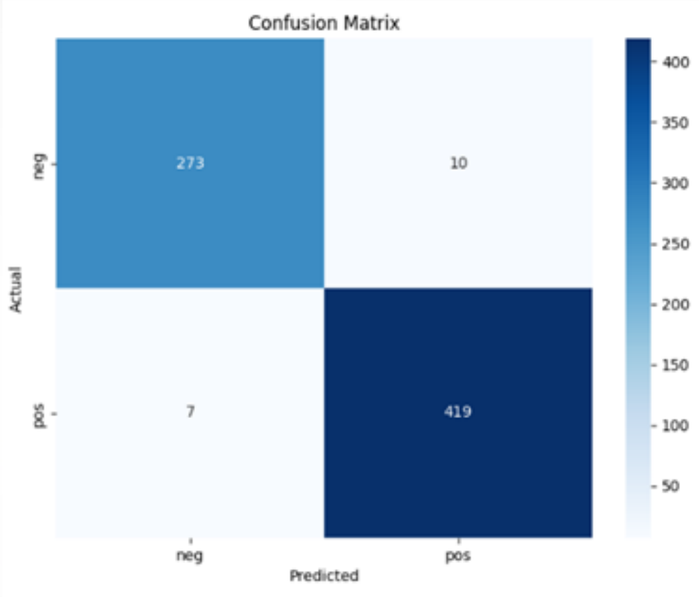
Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.97	0.99	0.94	0.96
			pos		0.96	0.99	0.98
		SVM	neg	0.98	0.97	0.96	0.97
			pos		0.98	0.98	0.98
		NB	neg	0.89	0.87	0.87	0.87
			pos		0.91	0.91	0.91
Group 2	-	Wangchan BERTa	neg	0.98	0.97	0.98	0.97
			pos		0.98	0.98	0.98

ตัวอย่างทั้งหมด
709 ตัวอย่าง
แบ่งเป็น
neg = 283
pos = 426

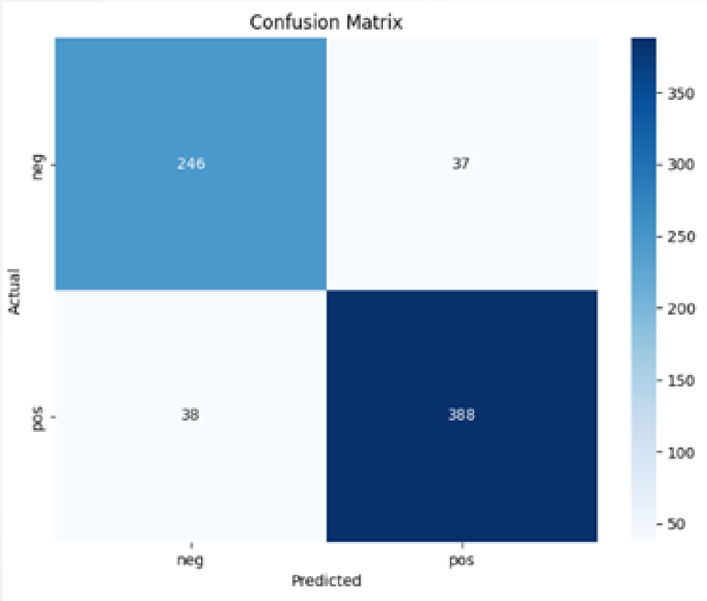
LR



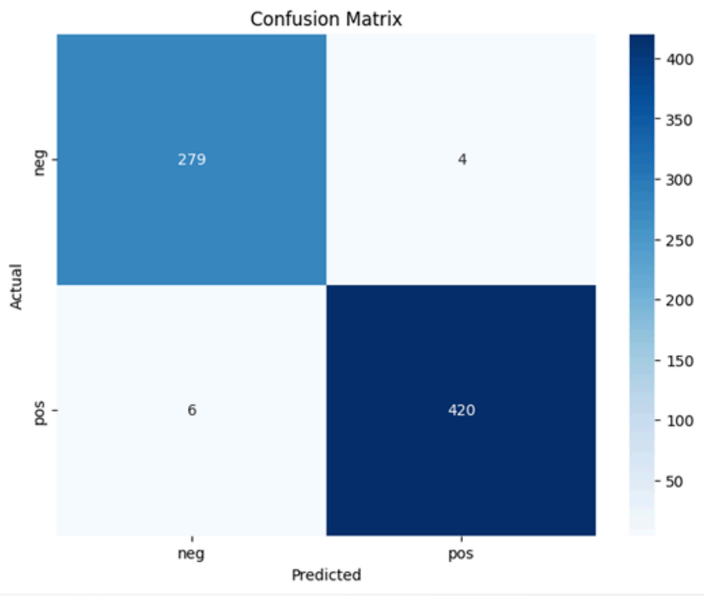
SVM



NB



WangchanBERTa



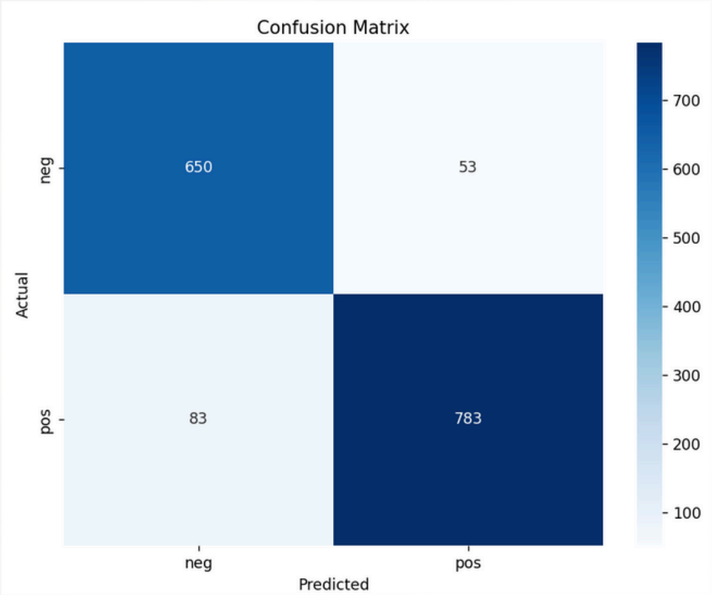
ประเภทชุดข้อมูล : ชุดที่ 2 **รูปแบบคำผสม :** ผสม (การผสมคำ 4 รูปแบบ)

จำนวนคลาส : 2 **รายละเอียด :** ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล

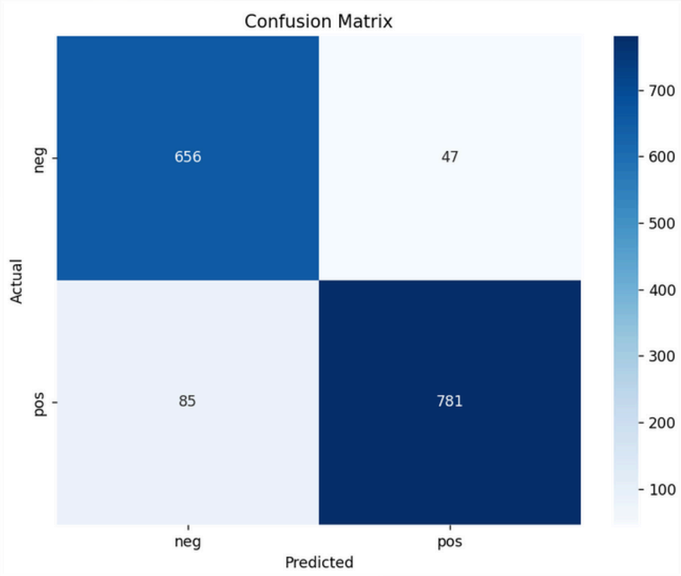
Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.91	0.89	0.92	0.91
			pos		0.94	0.9	0.92
		SVM	neg	0.92	0.89	0.93	0.91
			pos		0.94	0.9	0.92
		NB	neg	0.84	0.79	0.89	0.84
			pos		0.9	0.8	0.85
Group 2	-	WangchanBERTa	neg	0.98	0.99	0.98	0.98
			pos		0.98	0.99	0.99

ตัวอย่างทั้งหมด
1569 ตัวอย่าง
แบ่งเป็น
neg = 703
pos = 866

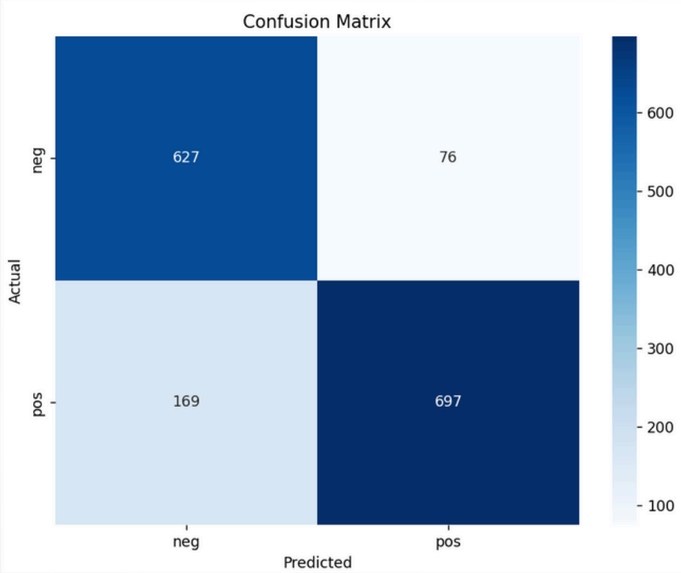
LR



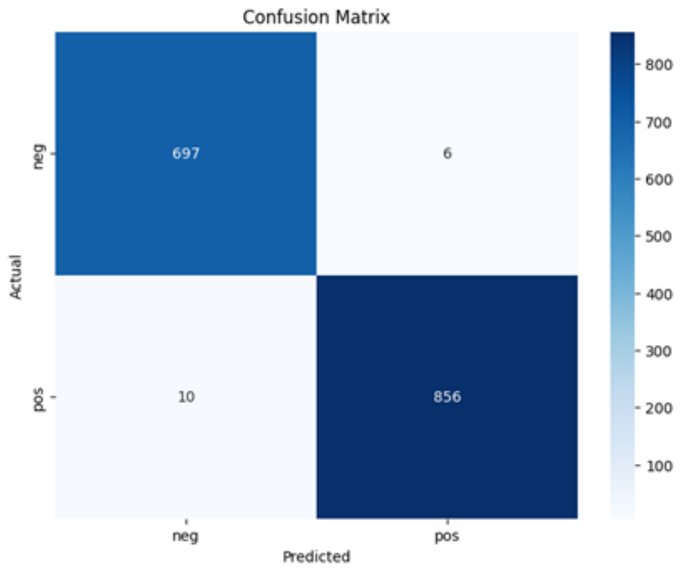
SVM



NB



WangchanBERTa



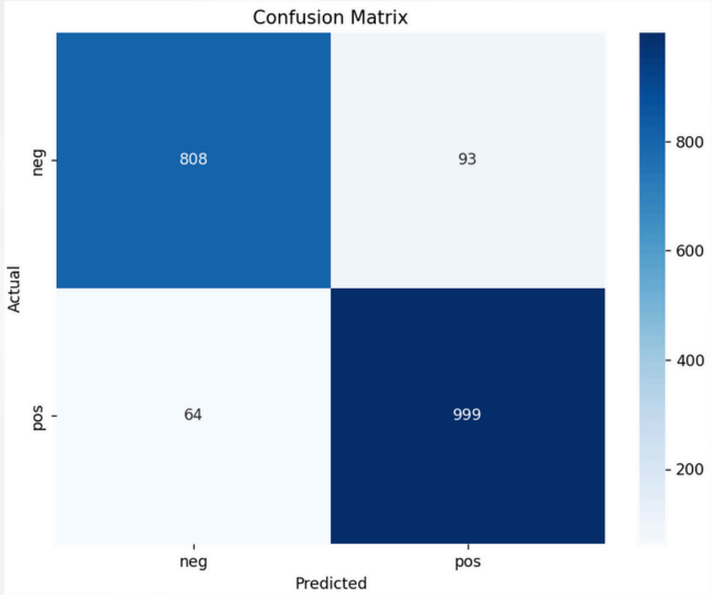
ประเภทชุดข้อมูล : ชุดที่ 3 **รูปแบบคำผสม :** ผสม (การผสมคำ 6 รูปแบบ)

จำนวนคลาส : 2 **รายละเอียด :** ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล

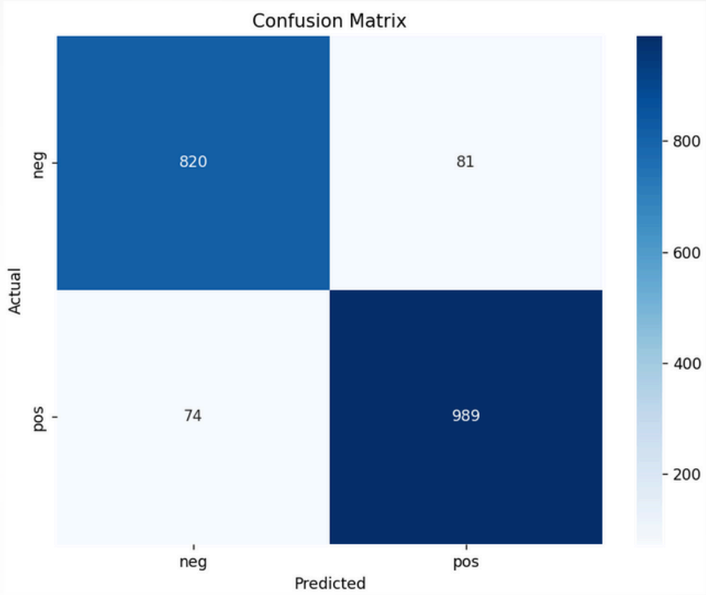
Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.92	0.93	0.9	0.91
			pos		0.91	0.94	0.93
		SVM	neg	0.92	0.92	0.91	0.91
			pos		0.92	0.93	0.93
		NB	neg	0.87	0.83	0.91	0.87
			pos		0.92	0.84	0.88
Group 2	-	WangchanBERTa	neg	0.98	0.98	0.97	0.98
			pos		0.97	0.98	0.98

ตัวอย่างทั้งหมด
1964 ตัวอย่าง
แบ่งเป็น
neg = 901
pos = 1063

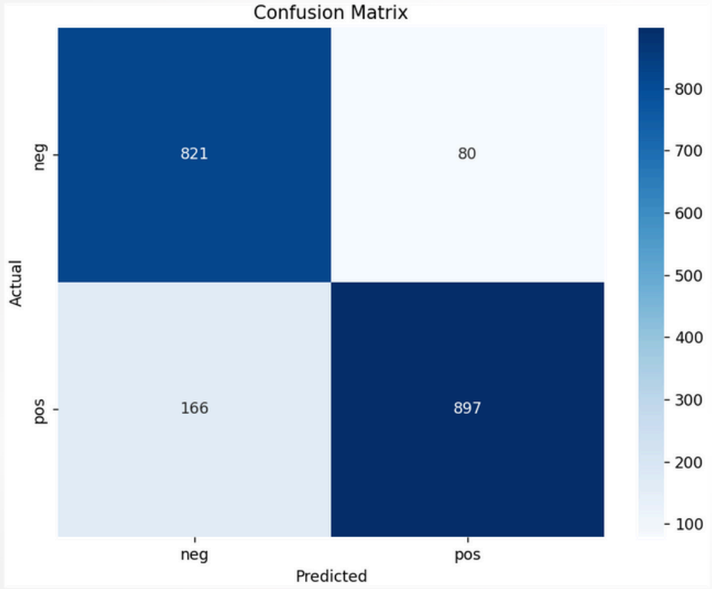
LR



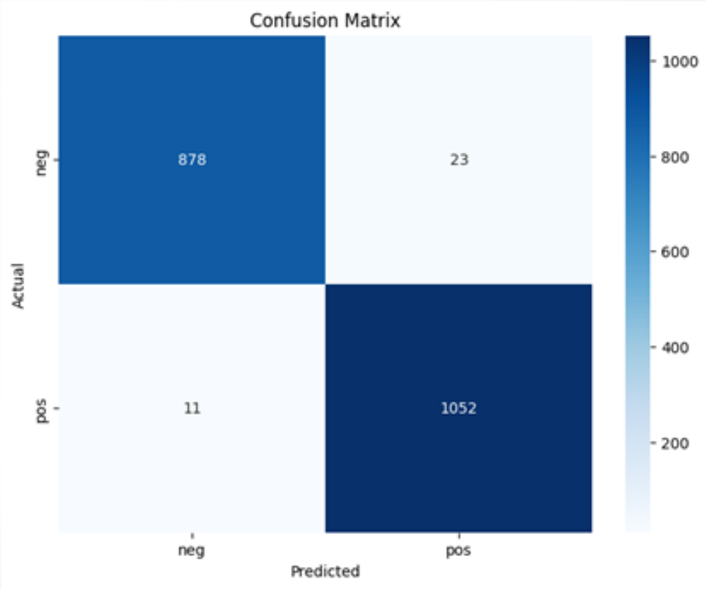
SVM



NB



WangchanBERTa



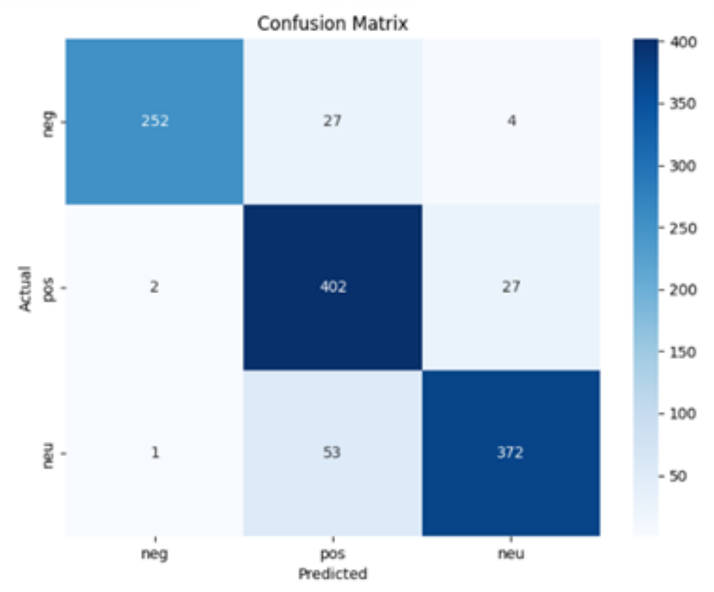
ประเภทชุดข้อมูล : ชุดที่ 4 **รูปแบบคำผสม :** เดี่ยว **จำนวนคลาส :** 3

รายละเอียด : ประโยคที่แสดงความรู้สึกเดียว มีข้อความเชิงบวก เชิงลบ เชิงกลาง

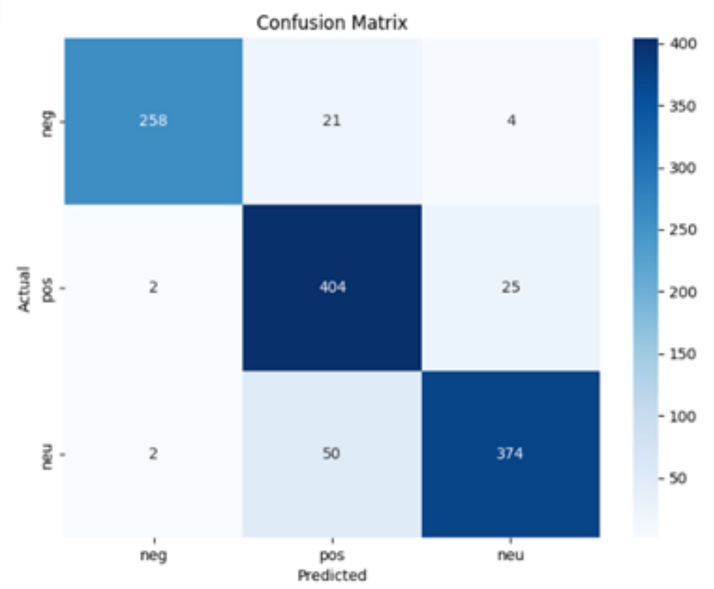
Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.9	0.99	0.89	0.94
			pos		0.83	0.93	0.88
			neu		0.92	0.87	0.9
		SVM	neg	0.91	0.98	0.91	0.95
			pos		0.85	0.94	0.89
			neu		0.93	0.88	0.9
		NB	neg	0.85	0.87	0.85	0.86
			pos		0.84	0.88	0.86
			neu		0.84	0.82	0.83
Group 2	-	Wangcha nBERTa	neg	0.94	0.98	0.96	0.97
			pos		0.93	0.93	0.93
			neu		0.93	0.94	0.94

ตัวอย่างทั้งหมด
2407 ตัวอย่าง
แบ่งเป็น
neg = 283
pos = 431
neu = 426

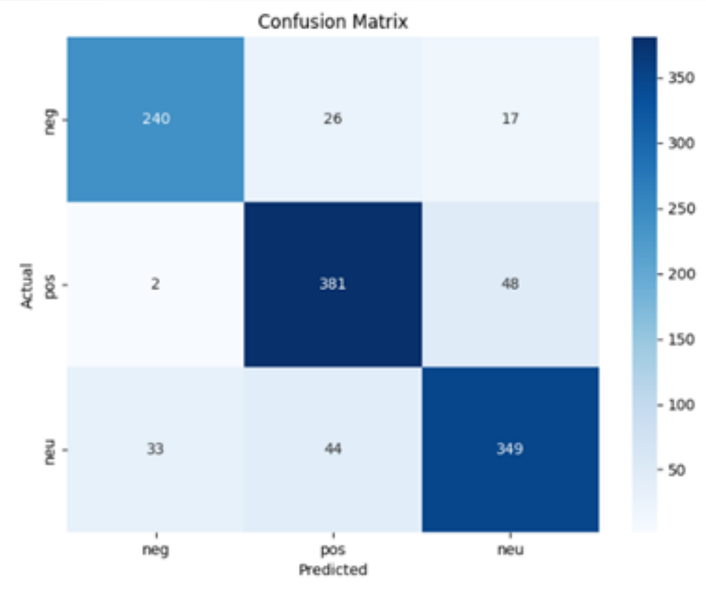
LR



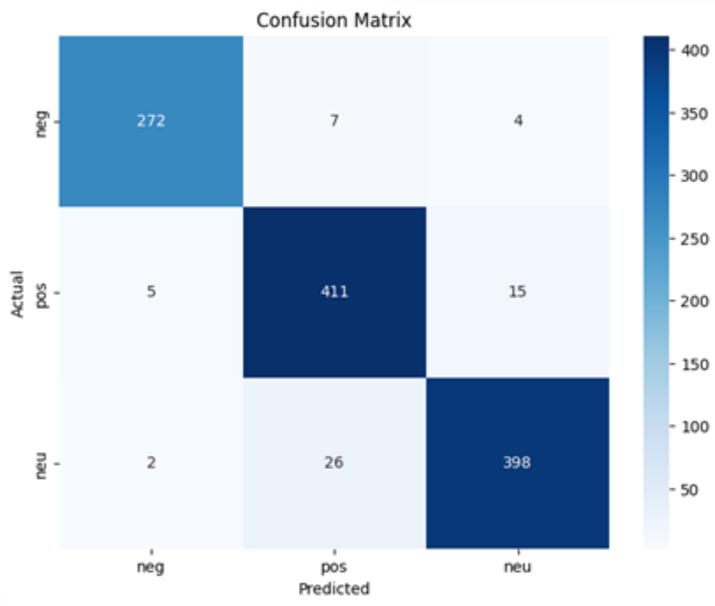
SVM



NB



WangchanBERTa



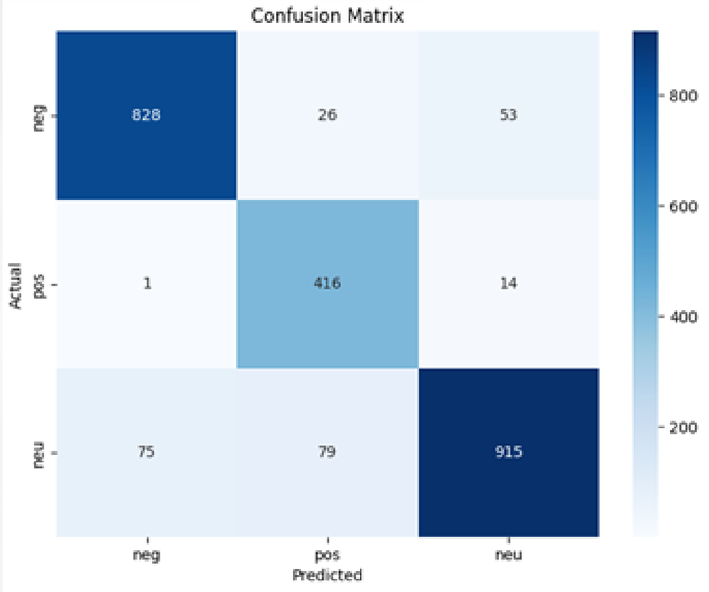
ประเภทชุดข้อมูล : ชุดที่ 5 **รูปแบบคำผสม :** ผสม (การผสมคำ 6 รูปแบบ) **จำนวนคลาส :** 3

รายละเอียด : ข้อความที่แสดงความรู้สึกผสม ข้อมูลปนกันในชุดข้อมูล

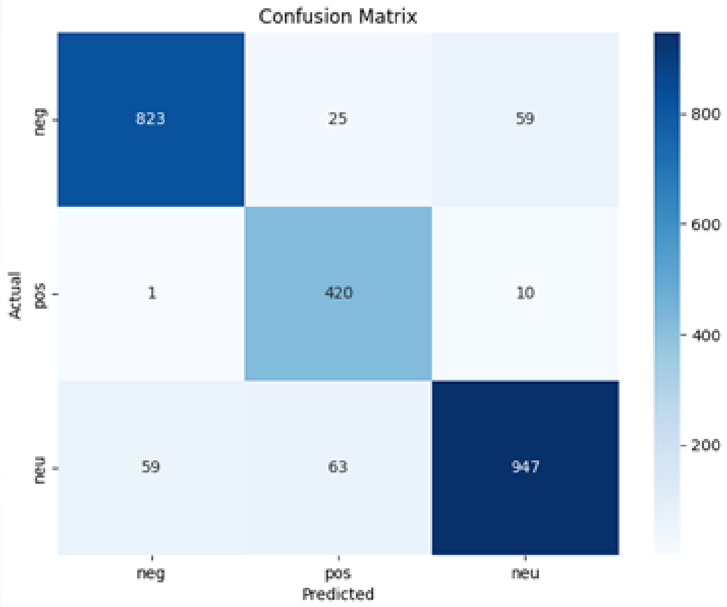
Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.9	0.92	0.91	0.91
			pos		0.8	0.97	0.87
			neu		0.93	0.98	0.89
		SVM	neg	0.91	0.93	0.91	0.92
			pos		0.83	0.97	0.89
			neu		0.93	0.89	0.91
		NB	neg	0.82	0.79	0.9	0.84
			pos		0.78	0.92	0.84
			neu		0.89	0.72	0.8
Group 2	-	Wangchan BERTa	neg	0.96	0.98	0.98	0.98
			pos		0.94	0.9	0.92
			neu		0.96	0.97	0.97

ตัวอย่างทั้งหมด
3261 ตัวอย่าง
แบ่งเป็น
neg = 1761
pos = 431
neu = 1069

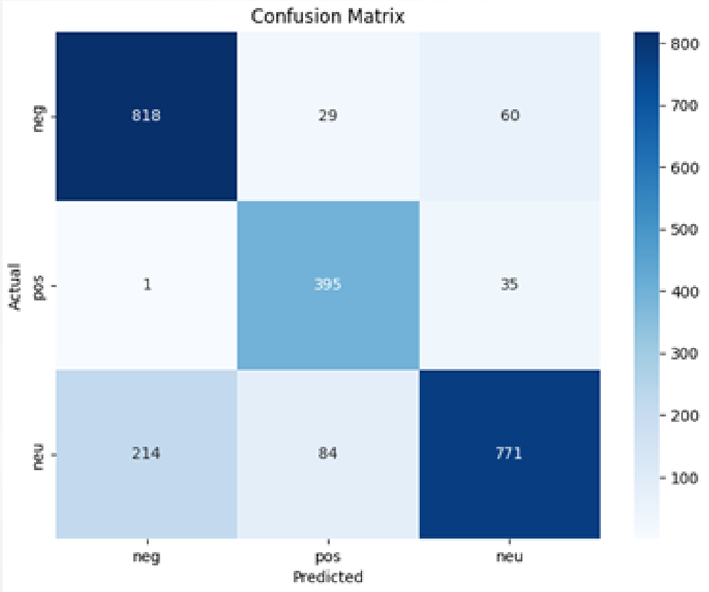
LR



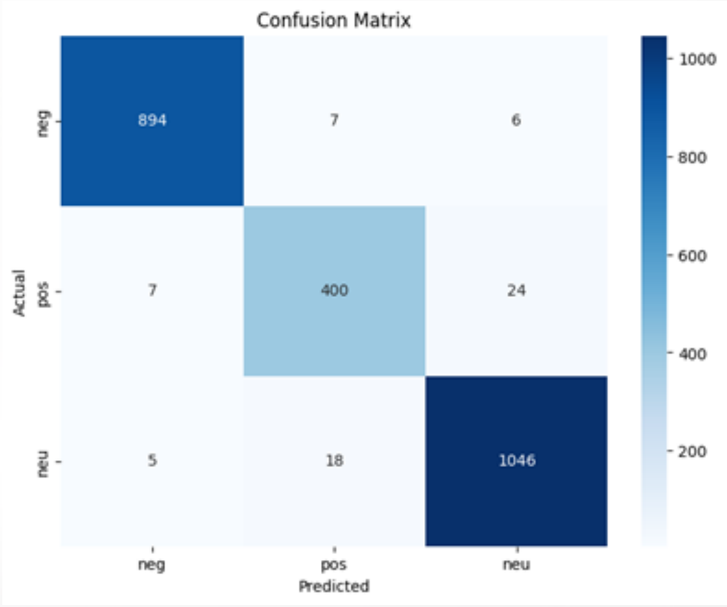
SVM



NB



WangchanBERTa



Q & A

THANK YOU