



ระบบวิเคราะห์ระดับภาษาด้วยเหมืองข้อความ

THE THAI LANGUAGE LEVEL ANALYSIS  
SYSTEM WITH TEXT MINING

# ผู้จัดทำโครงการ

นายสาริน สงครินทร์ 613021008-4

นายอภิสิทธิ์ ไกรยะโส 613021008-4

อาจารย์ที่ปรึกษา : อ.ดร.วรัญญา วรรณศรี

# วัตถุประสงค์

1. เพื่อสร้างชุดข้อมูล สำหรับนำไปวิเคราะห์กับโมเดล
2. เพื่อสร้างโมเดล สำหรับวิเคราะห์ประโยชน์ว่าเป็นทางการหรือไม่เป็นทางการ
3. เพื่อเปรียบเทียบประสิทธิภาพโมเดล สำหรับโมเดลที่เหมาะสมกับระบบที่สุด
4. เพื่อสร้างระบบวิเคราะห์ระดับของภาษาไทย

# ขอบเขตงานวิจัย

1. ระบบจะให้ผู้ใช้ทำการใส่ข้อความหรือประโยคที่ต้องการจะตรวจสอบ และแสดงผลลัพธ์ในรูปแบบข้อความ
2. ระบบมีการแนะนำส่วนที่ต้องแก้ไขในประโยคหรือข้อความ

## ข้อจำกัด

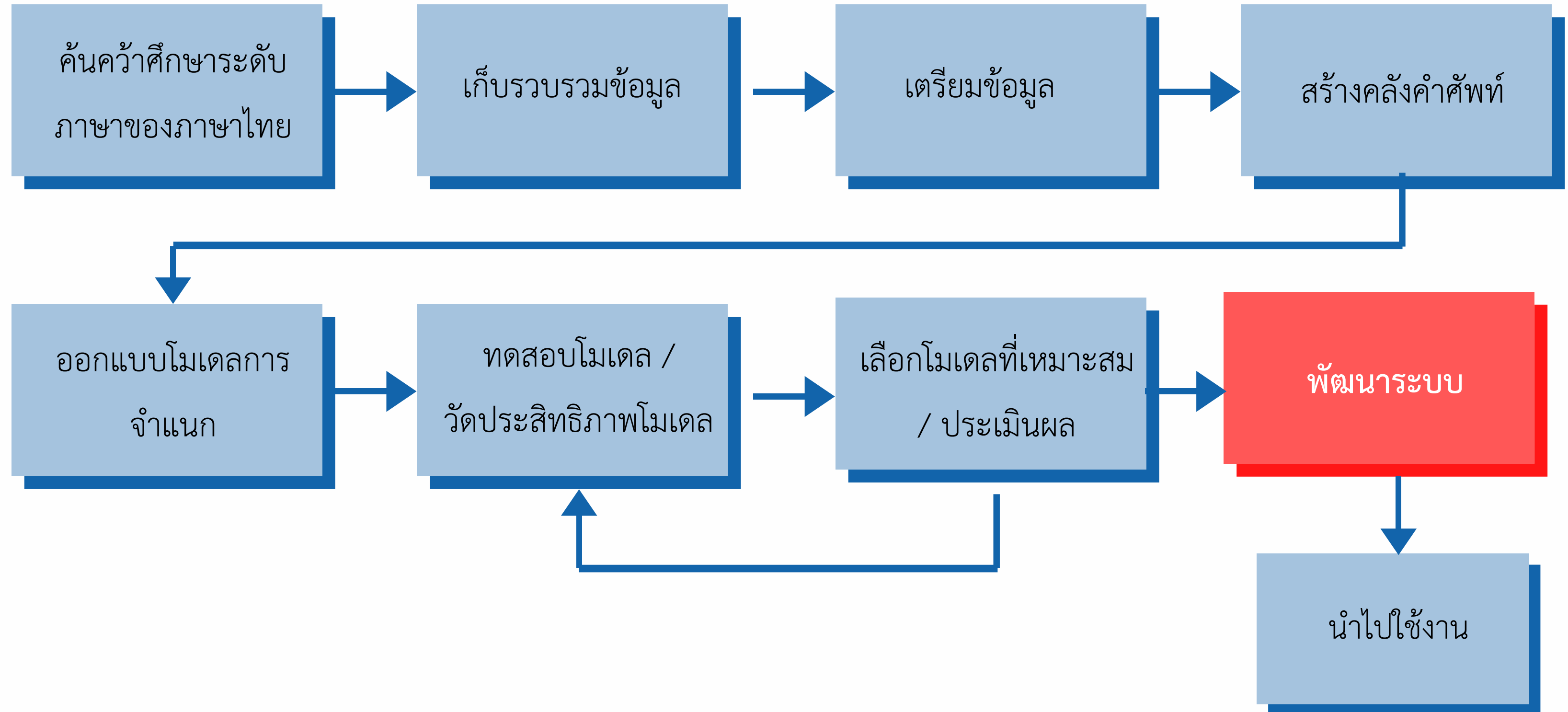
1. จำกัดเฉพาะภาษาไทย
2. ตรวจสอบเฉพาะภาษาที่ใช้ในรายงาน
3. ระบบจะรับข้อมูลเข้าในรูปแบบข้อความเท่านั้น

# ความก้าวหน้าจาก Final

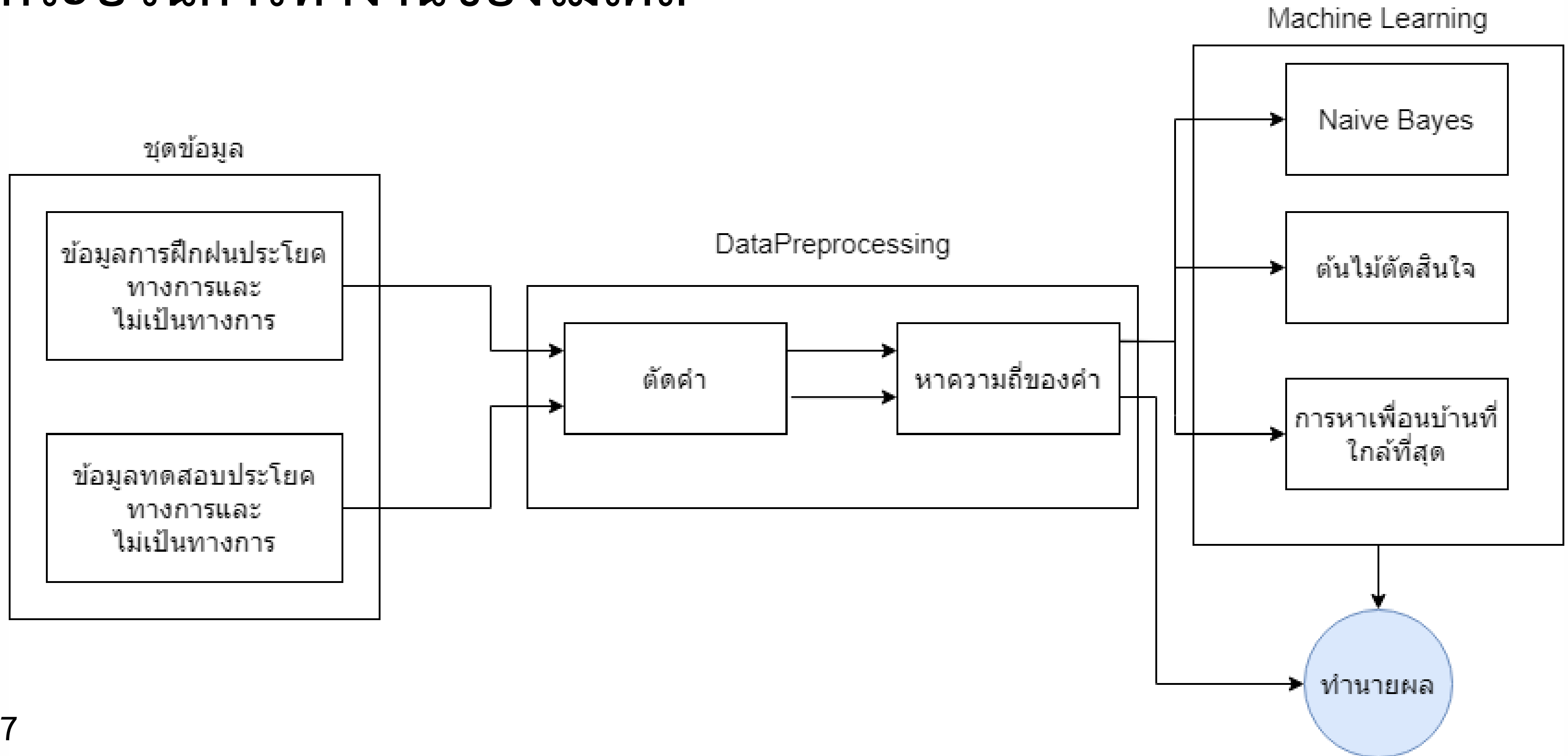
1. พัฒนาโมเดล

2. พัฒนาเว็บแอปพลิเคชัน

# ขั้นตอนการดำเนินงาน



# กระบวนการทำงานของโมเดล



# ชุดข้อมูล

ชุดข้อมูลทั้งหมด 2,000 ประโยคที่จะนำไปใช้ในโมเดลจะประกอบไปด้วย

	ประโยค
ประโยคทางการ	1,000
ประโยคไม่เป็นทางการ	1,000

- ตัวอย่าง - ชาวนาในขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก
- เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่สถานที่รับซื้อต่าง ๆ

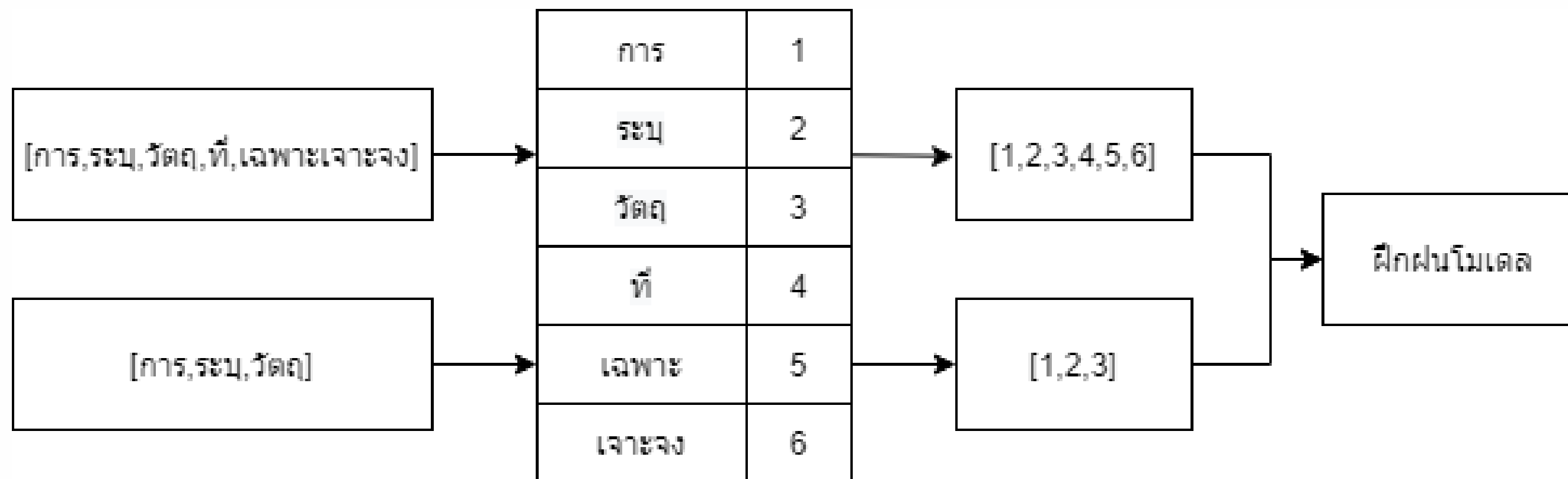


## ตัวอย่างชุดข้อมูล

Sent	Result
<u>ชาวสวน</u> ในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก	ไม่เป็นทางการ
เกษตรกรในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก	ทางการ
เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่ <u>โรงรับซื้อ</u> ต่าง ๆ	ไม่เป็นทางการ
เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่ <u>สถานที่</u> รับซื้อต่าง ๆ	ทางการ

# ตัวอย่างการเตรียมข้อมูล

การระบุวัตถุที่เฉพาะเจาะจง  $\longrightarrow$  การ | ระบุ | วัตถุ | ที่ | เฉพาะเจาะจง



# โมเดลที่นำมาทดสอบ

- Naive Bayes
- การหาเพื่อนบ้านที่ใกล้ที่สุด
- ต้นไม้ตัดสินใจ



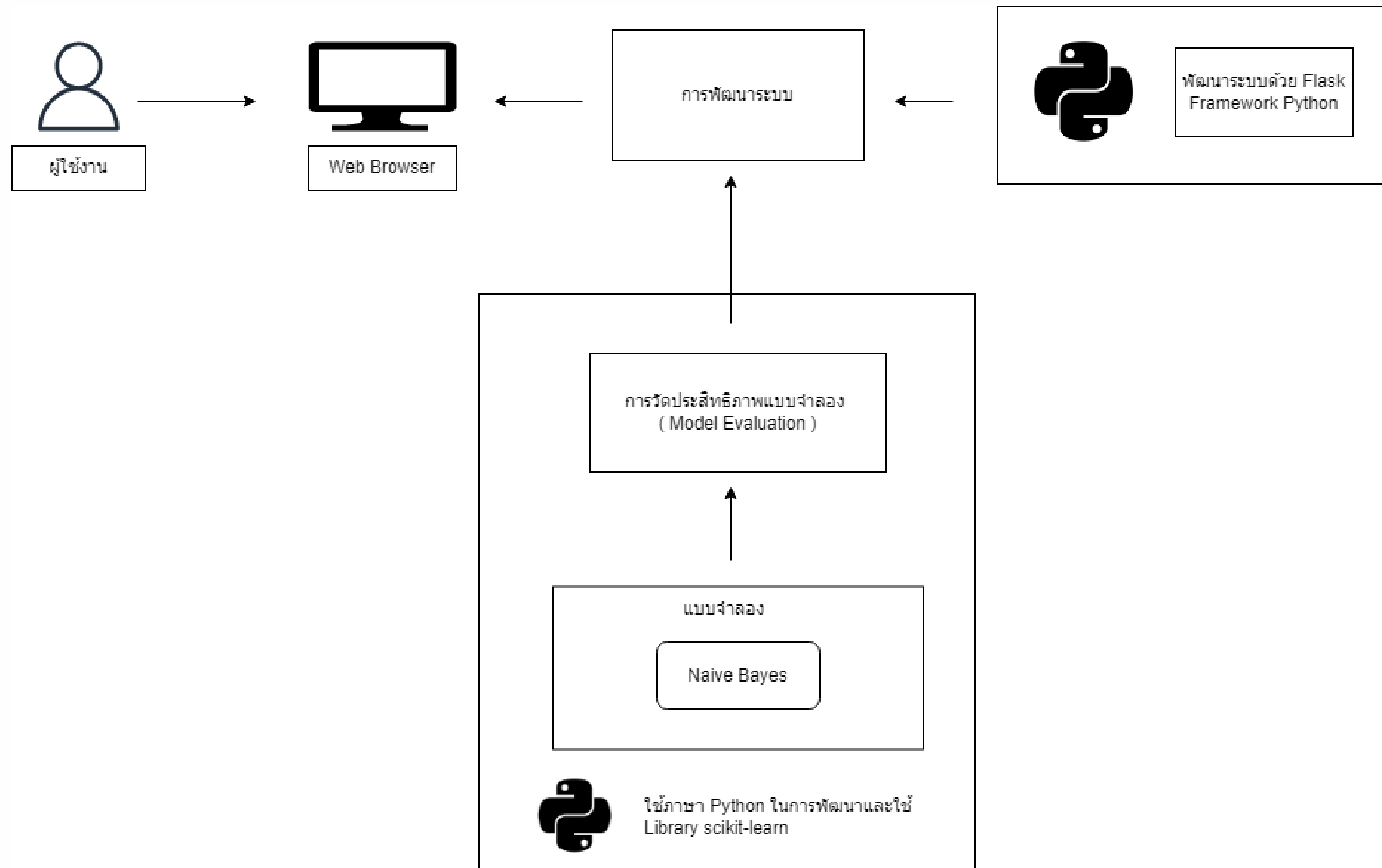
ผลการเปรียบเทียบวัดประสิทธิภาพของระบบด้วย

10-fold Cross Validation

จากการทดสอบ 3 โมเดลสรุปได้ว่า โมเดล Naive Bayes มีประสิทธิภาพสูงที่สุดอยู่ที่ 10-fold ซึ่งมีค่าความแม่นยำอยู่ที่ 75.63%

Model	Accuracy
Naive Bayes	75.63%
ต้นไม้ตัดสินใจ	63.37%
เพื่อนบ้านที่ใกล้ที่สุด ( $k = 5$ )	58.37%

# สถาปัตยกรรมของระบบ



# ความสามารถของแอปพลิเคชัน

1. ฟังก์ชันการตรวจสอบประโยคที่ให้ผลลัพธ์เป็นประโยคภาษาระดับทางการหรือประโยคภาษาระดับไม่เป็นทางการ
2. ฟังก์ชันการแนะนำคำที่เป็นทางการ
3. ฟังก์ชันให้ผู้ใช้งานแนะนำหากเว็บไซต์ตรวจสอบประโยคผิดพลาดหรือแนะนำคำที่ไม่เป็นทางการให้ผู้วิจัยได้ทราบ

# ตัวอย่างเว็บไซต์

THAI LANGUAGE LEVEL

GIVE FEEDBACK

กรอกข้อความ

ตรวจสอบ

คะแนนระดับภาษาของประโยค



# ตัวอย่างเว็บไซต์

## THAI LANGUAGE LEVEL

GIVE FEEDBACK

**ชาวสวน** ในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมากเกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่ **โรงรับซื้อ** ต่าง ๆ

ตรวจสอบ

### คะแนนระดับภาษาของประโยค

ประโยคไม่เป็นทางการ 53 %

ประโยคเป็นทางการ 47 %

จากทั้งหมด 23 คำ มีประโยคทางการ 21 คำ ประโยคไม่เป็นทางการ 2 คำ

# ตัวอย่างเว็บไซต์

**THAI LANGUAGE LEVEL**

GIVE FEEDBACK

**ชาวสวน** ในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมากเกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่ **โรงรับซื้อ** ต่าง ๆ

เกษตรกร

ตรวจสอบ

**คะแนนระดับภาษาของประโยค**

ประโยคไม่เป็นทางการ 53 %

ประโยคเป็นทางการ 47 %

จากทั้งหมด 23 คำ มีประโยคทางการ 21 คำ ประโยคไม่เป็นทางการ 2 คำ

# ตัวอย่างเว็บไซต์

THAI LANGUAGE LEVEL

ชาวสวนในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมากเกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่ **โรงรับซื้อ** ต่าง ๆ

ตรวจสอบ

คะแนนระดับภาษาของประโยค

ประโยคไม่เป็นทางการ 53 %

ประโยคเป็นทางการ 47 %

จากทั้งหมด 23 คำ มีประโยคทางการ 21 คำ ประโยคไม่เป็นทางการ 2 คำ

×

ข้อเสนอแนะ

SEND

## ส่วนที่พัฒนาเสร็จเรียบร้อยแล้ว

1. โมเดลทั้ง 3 แบบ คือ Naive Bayes, ต้นไม้ตัดสินใจ, การหาเพื่อนบ้านที่ใกล้ที่สุด
2. วัดประสิทธิภาพของระบบด้วย K-fold Cross Validation
3. เว็บไซต์แอปพลิเคชันบางส่วน

## การพัฒนาต่อไป

1. พัฒนาหน้าเว็บแอปพลิเคชันให้สมบูรณ์
2. ทดสอบเว็บไซต์แอปพลิเคชัน



80%



# DEMO

Q/A

- [1] ราชวิทย์ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์สมประเสริฐศรี. (2556). การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ. คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
- [2] Avinash Navlani. (2564). KNN Classification using Scikit-learn, คำนวันที่ 30 กันยายน 2564 จาก <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>
- [3] รักษ์พงศ์ ธรรมพสุณา. ระดับภาษาและการใช้ภาษาที่ถูกต้อง. คำนวันที่ 12 กุมภาพันธ์ 2563 จาก <http://gened.siam.edu/wp-content/uploads/2018/07/thaic-handout-03.pdf>
- [4] เอกรัฐ บุญเชียง. การแบ่งกลุ่มข้อมูลและการจำแนกประเภทข้อมูล. เชียงใหม่ : ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2561.
- [5] อุมารินทร์ นอกตะแบก. ระดับภาษาและคำราชาศัพท์. คำนวันที่ 20 กันยายน 2563. จาก <https://sites.google.com/site/018schoolnet/bth-thi1-radab-phasalaea-kharachasaphth/1>
- [6] Moshe Koppel, Jonathan Schler, Kfir Zigdon. Determining an Author's Native Language by Mining a Text for Errors. Computer Science Department Bar-Ilan University. 624-628.
- [7] Hanumanthappa, Narayana Swamy. (2015). Indian Language Text Mining. Department of computer Applications Bangalore University Bangalore.
- [8] Ekaterina Shutova, Patricia Lichtenstein. (2016). Psychologically Motivated Text Mining. Computer Laboratory University of Cambridge, Dept. of Cognitive and Information Sciences University of California, Merced.
- [9] Fadi ABU SHEIKHA, Diana INKPEN. Automatic Classification of Documents by Formality. University of Ottawa, SITE University of Ottawa, SITE 800 King Edward, Ottawa, ON, Canada
- [10] ราชวิทย์ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์สมประเสริฐศรี. (2556). การจำแนกกลุ่มคำถามอัตโนมัติบนเครือข่ายสังคมออนไลน์. คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม