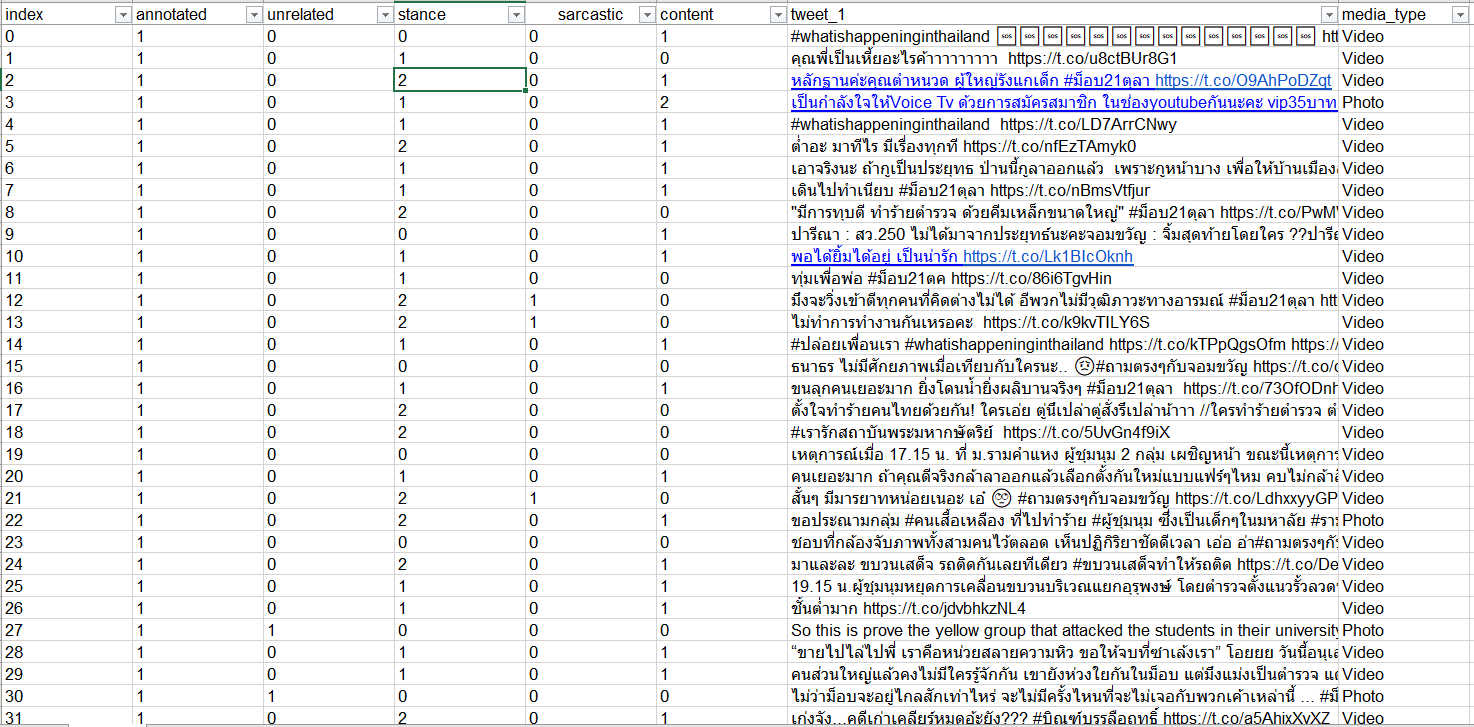
**Auto-Labeling Report**

**08.01.2021**



Dataset ที่ใช้ในการ Auto-Labeling คือ Data ใน Task2 ที่มีการ Label Data แล้วว่าเกี่ยวข้องกับการเมืองหรือไม่ เกี่ยวข้องกับฝ่ายใด(ฝ่ายรัฐบาล/ฝ่ายค้าน) ฯลฯ มีขั้นตอนในการทำดังนี้

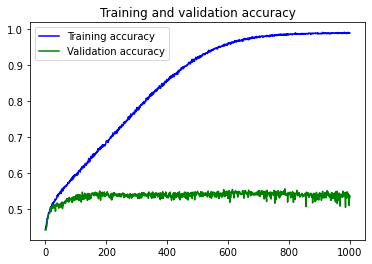
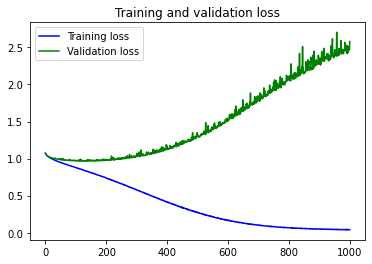
1. แบ่งข้อมูลออกเป็น 2 ส่วน คือ Training Set และ Validation Set เพื่อใช้ในการทดสอบประสิทธิภาพของ Model (ในเคสนี้แบ่งเป็น 80% Training Set : 20% Validation Set)

2. Image Processing: Extract รูปภาพให้เป็นข้อมูลที่สามารถนำมาวิเคราะห์ได้ (X: pixel of images, y: labeled ซึ่งจะลอง stance ก่อนเพื่อเป็น case study)

3. สร้างโมเดลเพื่อทำ Pattern Recognition (Convolutional Neural Network)

4. นำโมเดลมาประเมินผล (Model Evaluation)

ผลที่ได้ (Prototype)

ผลที่ได้คือ โมเดลมีประสิทธิภาพใน Training Set (High Accuracy, Low Loss) แต่ยังไม่มีประสิทธิภาพใน Validation Set (Steadily Accuracy, High Loss) แปลว่าโมเดลยังเกิด Overfitting

สิ่งที่จะทำต่อไป

* แก้ไขปัญหา Overfitting
* เพิ่มประสิทธิภาพของโมเดล (Optimization)
* ทดลองกับ Data จริง เมื่อมีข้อมูลมากพอ

**Auto-Labeling Report**

**11.01.2021 – 12.01.2021**

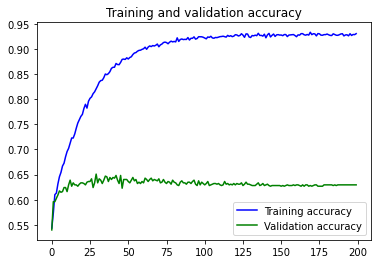
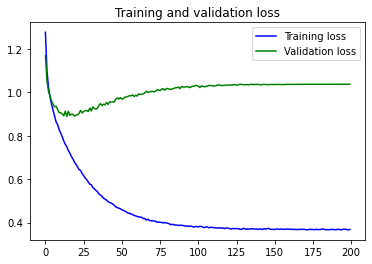
สิ่งที่ทำเพิ่ม

1. พยายามแก้ไขปัญหา Overfitting : Weight decay, Regularization, Dropout, เพิ่ม/แก้ไข Sample

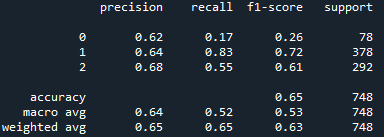
2. ออกแบบหน้าตาผลลัพธ์ที่ได้จากการทำนาย

ผลที่ได้

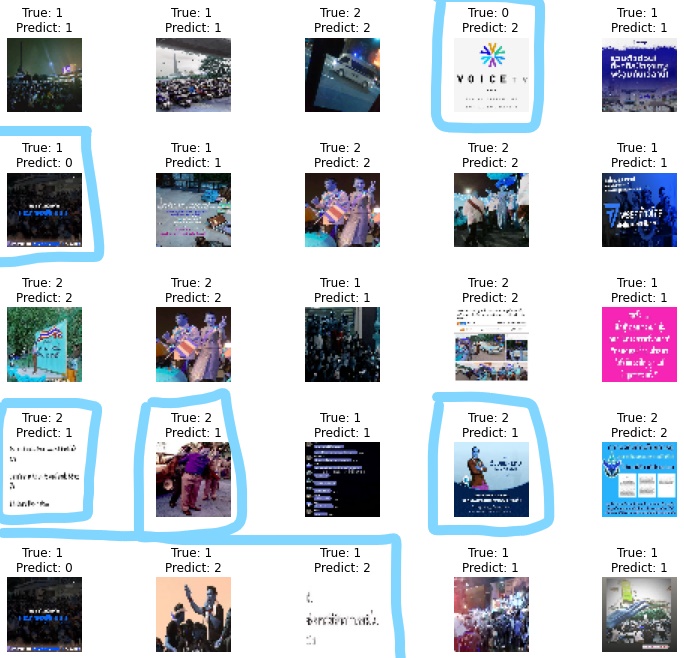
* โมเดลมีความแม่นยำ (Accuracy) ที่สูงขึ้น แต่ยังคงเป็น Steadily Accuracy อยู่ (จาก 53% -> 63.3%) ตอนนี้ 65.10%

* มีการแสดงผลลัพธ์ของโมเดลที่อธิบายให้เห็นภาพมากขึ้น

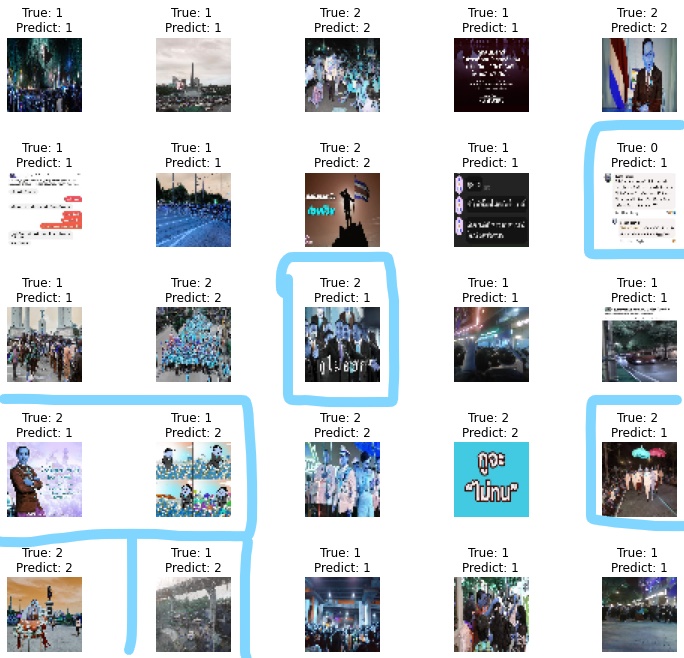






Model 63.3%





Model 65.10%

**Auto-Labeling Report**

**13.01.2021**

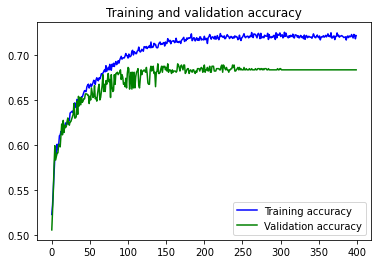
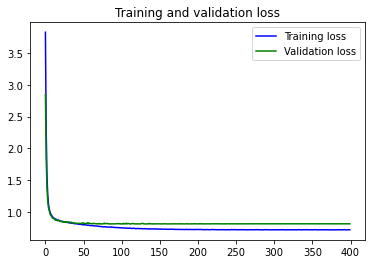
สิ่งที่ทำเพิ่ม

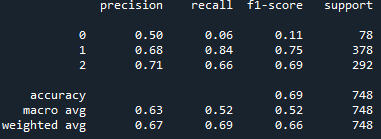
1. หาสาเหตุของปัญหา Overfitting

2. หาวิธีการแก้ปัญหา Overfitting เพิ่ม: Normalization layer, Check Selection bias & Human Error

3. reduce node but increase learning rate / increase layer

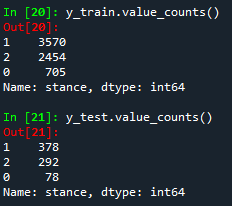
ผลที่ได้



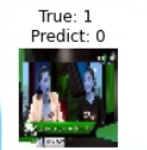
Model มี Accuracy ที่สูงขึ้น (68.98%) และ loss น้อยลง และพบว่าปัญหา Overfitting เกิดจาก Recall ของ label 0 มีค่าต่ำมาก (0.06) นั้นหมายถึงโมเดลที่ได้นั้นไม่มีความสามารถในการทำนาย label 0 ซึ่งอาจจะมีสาเหตุจาก

1. Selection bias: การเลือก Training set/Validation set มีความไม่เหมาะสม เช่น ข้อมูลส่วนใหญ่ที่ใช้ในการ Train จะไม่มีหรือมี label 0 น้อยทำให้โมเดลไม่สามารถทำนาย Pattern ของมันได้ใน Validation set

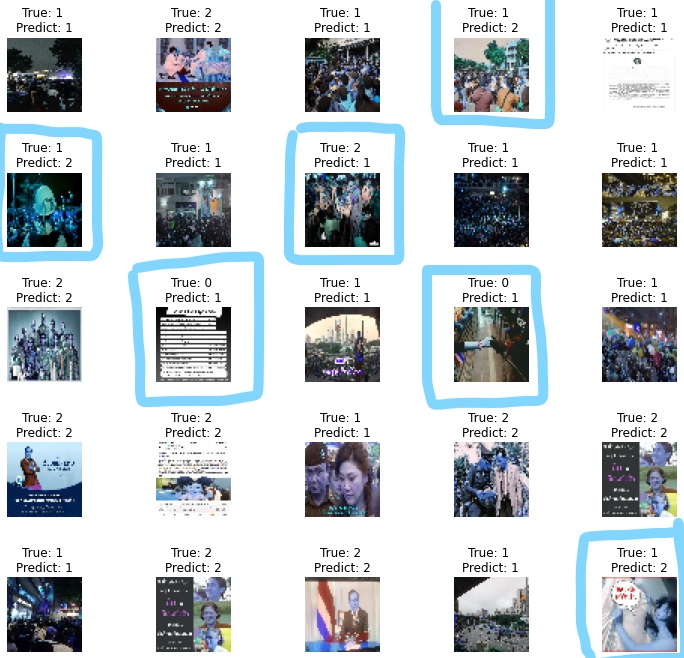


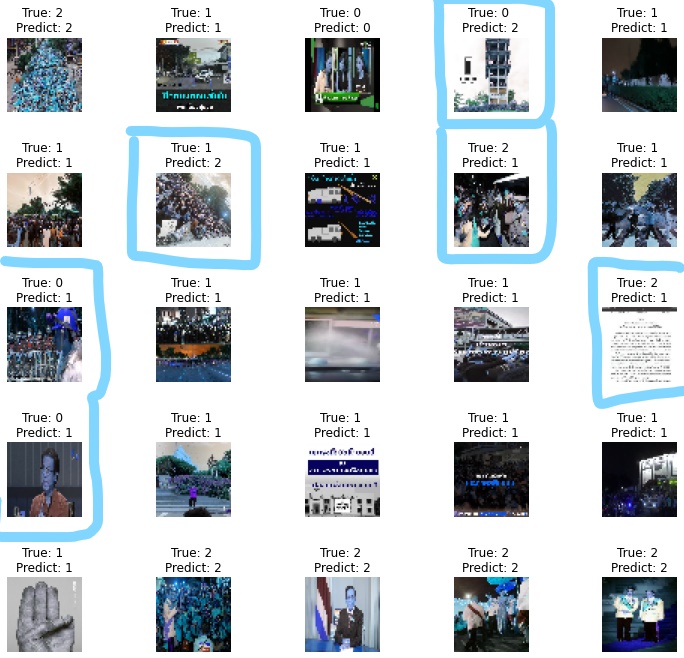
Sample ที่ใช้ใน Training set เทียบกับ Validation set

2. Human Error: เนื่องจาก label 0 หมายถึงกลุ่มที่มีลักษณะคลุมเครือ (เป็นกลาง ขึ้นอยู่กับ text) ทำให้บางครั้งจึงเกิดการ label ที่ผิดพลาดของตัวผู้ทำข้อมูลเอง



ตัวอย่าง Human Error ในกรณีที่รูปซ้ำหรือคล้ายๆกัน แต่ Label ข้อมูลไม่เหมือนกัน ทำให้ Machine เกิดความสับสนในการทำนาย





Model 68.98%