News Text Classification

Text Classification

Nowadays on the Internet there are a lot of sources that generate immense amounts of daily news. In addition, the demand for information by users has been growing continuously, so it is crucial that the news is classified to allow users to access the information of interest quickly and effectively. This way, the machine learning model for automated news classification could be used to identify topics of untracked news and/or make individual suggestions based on the user's prior interests. Thus, our aim is to build models that take as input news headline and short description and output news category.

DATA SOURCE

Articles in this dataset (thaisum.csv) was collected from several Thai news websites namely Thairath, ThaiPBS, Prachatai and The Standard sharing in the Github text classification repository.

ไทยรัฐ ออนไลน์ THAIRATH ONLINE

Thai PBS NEWS

ประชาไท นิวส์ออนไลน์

THE STANDARD

collection

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 358868 entries, 0 to 358867
Data columns (total 6 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   title    358868 non-null  object
 1   body     358868 non-null  object
 2   summary  358868 non-null  object
 3   type     269183 non-null  object
 4   tags     348273 non-null  object
 5   url      358868 non-null  object
dtypes: object(6)
memory usage: 16.4+ MB
```

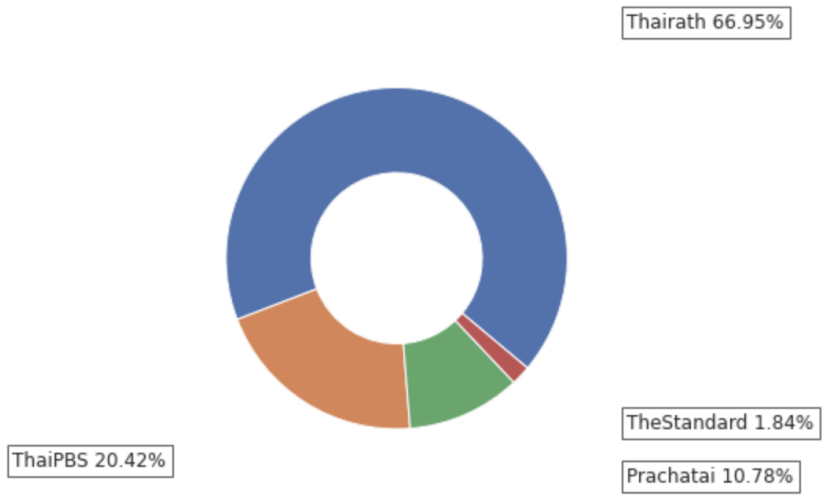
```
ThaSum dataset contains 358868 documents.
```

120889

title	รมว.ยุติธรรม ห่วงเหยื่อกราดยิงโคราช เล็งพิจารณ...
body	จากเหตุการณ์คนร้ายใช้อาวุธปืนกราดยิง เป็นเหตุใ...
summary	รมว.ยุติธรรม และนายกรัฐมนตรี ห่วง เหยื่อมือปืน...
type	ข่าว,อาชญากรรม
tags	วิสามัญ คนร้ายกราดยิงโคราช,วิสามัญ จักรพันธ์ ถ...
url	https://www.thairath.co.th/news/crime/1767448

DATA

Dataset contains over 358,000 articles.




```
There are  240256  articles from Thairath covering up Thairath 66.95%
There are  73290  articles from ThaiPBS covering up ThaiPBS 20.42%
There are  38703  articles from Prachathai covering up Prachatai 10.78%
There are  6619  articles from The Standard covering up TheStandard 1.84%
```

Topic Distribution

	index	sum
0	การเมือง	48980
1	สังคม	29288
2	กีฬา	22072
3	ต่างประเทศ	21927
4	เศรษฐกิจ	18070
5	อาชญากรรม	17344
6	สิทธิมนุษยชน	12528
7	บันเทิง	12180
8	สิ่งแวดล้อม	10673
9	คุณภาพชีวิต	9154
10	ไลฟ์สไตล์	7381
11	วิทยาศาสตร์เทคโนโลยี	6020
12	ผู้หญิง	3894
13	วัฒนธรรม	3737
14	แรงงาน	2638
15	การศึกษา	2267
16	ความมั่นคง	2127

สิ่งแวดล้อม: ภัยพิบัติ + สิ่งแวดล้อม

Word Cloud

สังคม

การเมือง

กีฬา

วิทยาศาสตร์เทคโนโลยี

เศรษฐกิจ

อาชญากรรม

ต่างประเทศ

สิทธิมนุษยชน

การศึกษา

	index	sum
0	การเมือง	48980
1	สังคม	29288
2	กีฬา	22072
3	ต่างประเทศ	21927
4	เศรษฐกิจ	18070
5	อาชญากรรม	17344
6	สิทธิมนุษยชน	12528
7	บันเทิง	12180
8	สิ่งแวดล้อม	10673
9	คุณภาพชีวิต	9154
10	ไลฟ์สไตล์	7381
11	วิทยาศาสตร์เทคโนโลยี	6020
12	ผู้หญิง	3894
13	วัฒนธรรม	3737
14	แรงงาน	2638
15	การศึกษา	2267
16	ความมั่นคง	2127

Methodology

Data Preparation

Type Grouping → Data Cleaning → One-hot encoder → Tokenization

Deep Learning

No Pre-train embedding → LSTM

No Pre-train embedding → Bi-LSTM

Glove embedding → LSTM

Glove embedding → Bi-LSTM

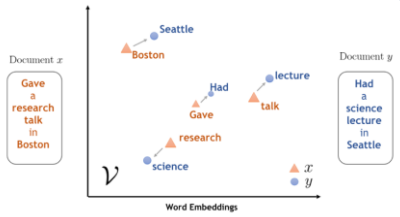
Machine Learning

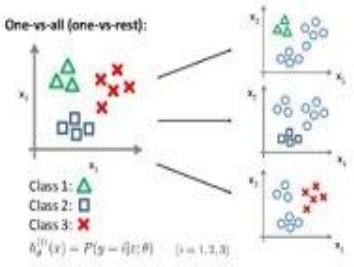
TFIDF embedding → One VS Rest (SVM)

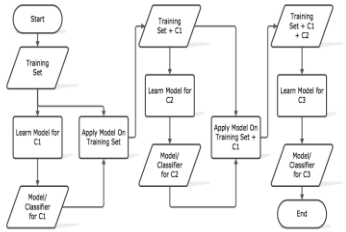
TFIDF embedding → One VS Rest (Naive Bayes)

TFIDF embedding → Multi KNN

TFIDF embedding → Chain







Result

ข้อความ	Actual	DL1 (LSTM No-Pretain)	DL2 (BILSTM + No-Pretain)	DL3 (LSTM + GLOVE)	DL4 (BILSTM + GLOVE)	ML1 SVM	ML2 Naive Bay
ลูกอยู่ไม่นิ่งเดี๋ยวปีนโซฟาบ้างปีนเก้าอี้...	ผู้หญิง,ไลฟ์สไตล์	ผู้หญิง,ไลฟ์สไตล์	ผู้หญิง,ไลฟ์สไตล์	ไลฟ์สไตล์	ไลฟ์สไตล์	ไลฟ์สไตล์	N/A
ปัจจุบันเป็นเอกฉันท์ว่าจะขอประณามการกระทำของนายมงคล กิตติ์ สุขสินธารานนท์...	การเมือง	สังคม	การเมือง	อาชญากรรม	อาชญากรรม	สังคม	การเมือง

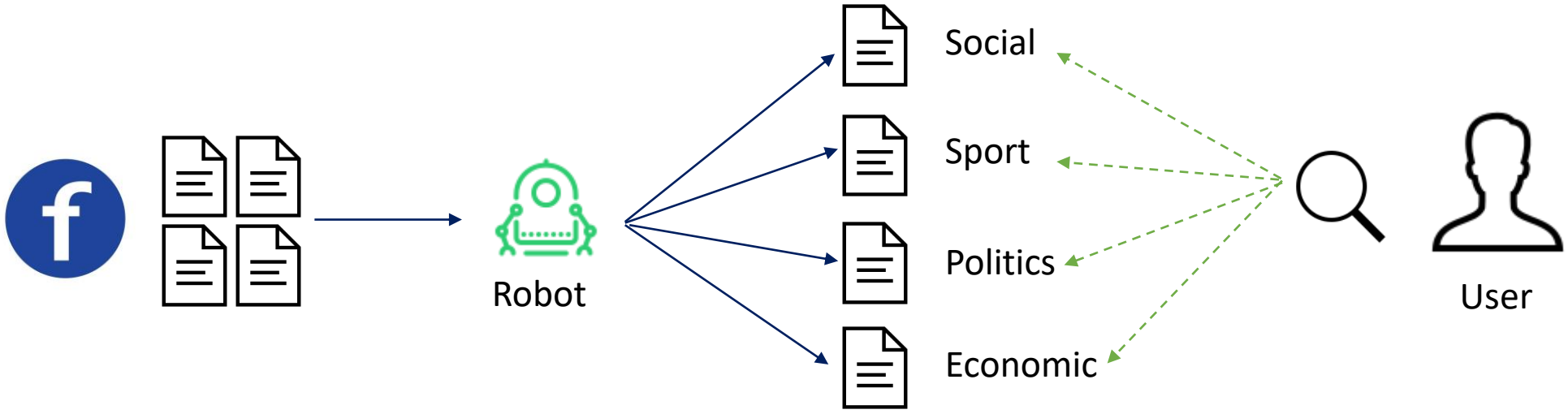
No pretrain + LSTM	No pre-train + BI LSTM	Glove + LSTM	Glove + BI LSTM	OneVSRest SVM	OneVSRest Naive Bayes
66.8%	66.4%	67.13%	67.13%	68.41%	36.5%

Show Case

Problem & Further work

Problem:
1. Limitation of computer resource even use the colab pro.
2. Hard to handle Thai language such as abbreviation.
3. Use human to define categories

Future work:
1. Automate tagging information on the social network
2. Content filtering



Thank you