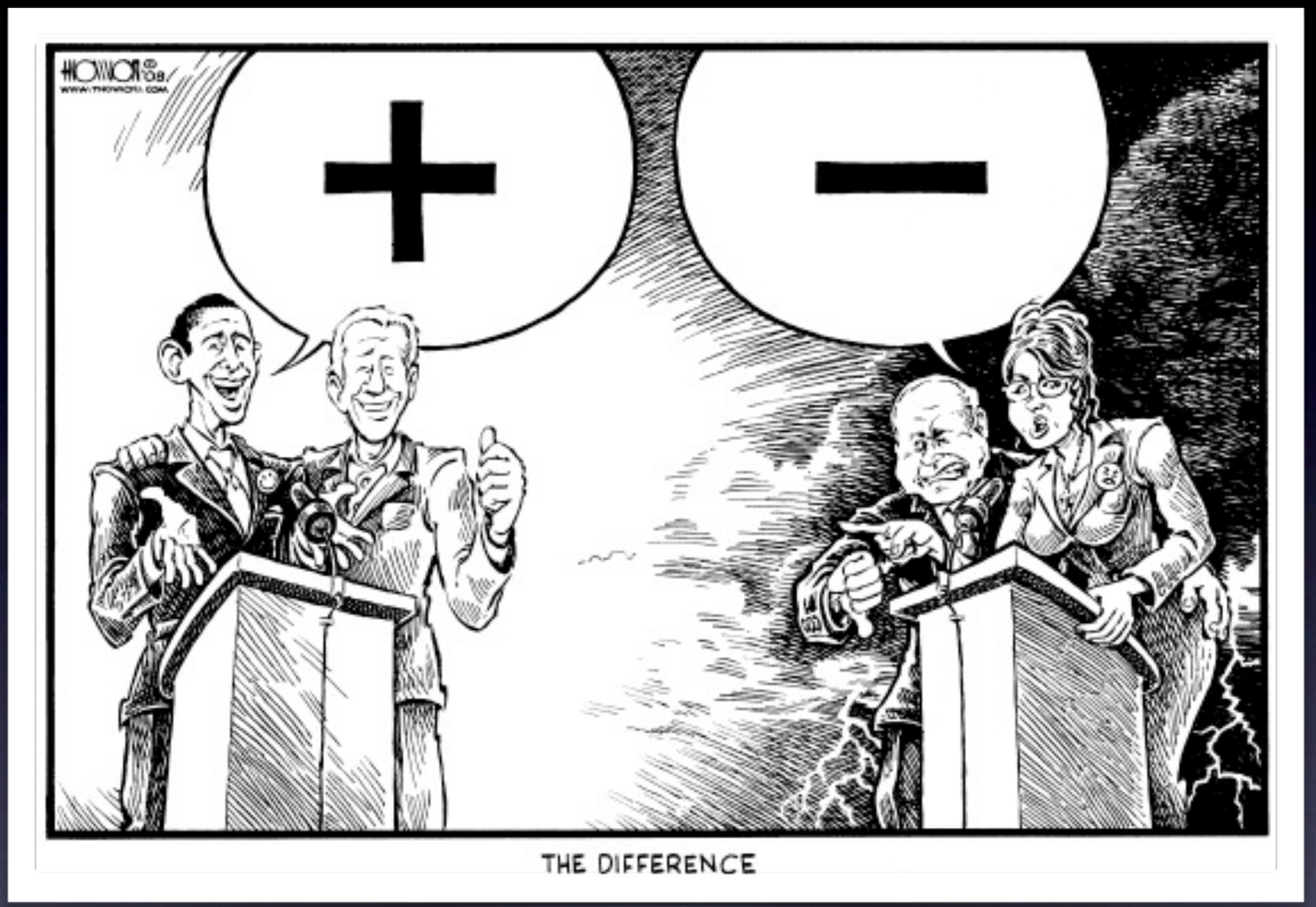
Thai Tweet Sentimental Analysis

นายประยุกต์ เจตสิกทัต | นายพงศกร อุชุปาละ

ทฤษฎีเซตเบื้องต้น
ทฤษฎีเซตและความสำคัญ

Twitter



เชิงบวก VS เชิงลบ

Trend

ขอบเขตของการศึกษา

- จำแนกทวีต “ภาษาไทย”
- จำแนกอารมณ์ “เชิงบวก” หรือ “เชิงลบ” เท่านั้น
- อนุมานอารมณ์ของ Train Data จาก Emoticon

การทดลอง

เครื่องมือ

โครงสร้างของระบบ

- ตาราง Tweets
- ตาราง Tweets หลังทำการ Preprocess
- ตาราง Tweets หลังทำการตัดคำ
- ตารางความน่าจะเป็นของ Class
- ต่างรางความน่าจะเป็นของ Feature (Given Class)

การทำงาน 8 ขั้นตอน

I. ดึงทวีตจาก API

- Query
- http://search.twitter.com/search.json?rpp=100&q=:) OR :(&lang=th

2. Preprocessing (1/3)

- คัดเลือกและจำแนก Train Data
  - Emoticon เชิงบวก
    - =), :), :), :-), (;, (;, (-;, :D, ;D, ^_^, ^^, <3
  - Emoticon เชิงลบ
    - :(, : (, :-(, TT, T_T, - -", - -'

2. Preprocessing (2/3)

- ตัดสัญลักษณ์อื่นๆ ทิ้ง
  - :p, :P, >^<, >_<, >__<, >3<, -3-, :3, = =, -_-, - -a
- ตัวเลข
- สัญลักษณ์พิเศษต่างๆ เช่น #, @, $, %, \r, \n

2. Preprocessing (3/3)

- Tokenization
- แทนที่ @mention ด้วย (:username)
- แทนที่ #hashtag ด้วย (:hashtag)
- แทนที่ URL ด้วย (:url)

3. ตัดคำ

- ใช้โปรแกรม SWATH
  - ตัดทวีตที่ไม่สามารถแปลงเป็น CP874 ทิ้ง
- ตัด | หน้าสุด และหลังสุด (ถ้ามี)
- ตัดช่องว่างทิ้ง

4. คำนวณ P(C)

- P(+) = Positive Tweets / Total Tweets
- P(-) = Negative Tweets / Total Tweets

5. คำนวณ P(f|C)

- ลักษณะ
  - Unigram
  - Bigram
- เลือกเฉพาะ Feature ที่ปรากฏใน Class นั้นๆ 2 ครั้งขึ้นไป

7. เตรียม Test Data

- เลือกทวีตที่ไม่ซ้ำกับ Train Data จำนวน 100 ทวีต
- จำแนกอารมณ์ด้วยคน

8. การทดสอบ

- นำ Test Data มา Preprocess และตัดคำ
- ดึง Feature ทั้งหมดออกมา
- แทน Feature ที่ไม่รู้จักเสมือนว่ามีปรากฏอยู่ในแต่ละ Class 1 ครั้ง
- จำแนกด้วย Naive Bayes Classifier

$$C_{nb} = \underset{c \in \{+,-\}}{\arg max}\ P(c) \prod_i P(f_i|c)$$

Naive Bayes Classifier

ผลการดำเนินงาน

Train Data

- จำนวน 67,047 ทวีต
- เชิงบวก 54,737 ทวีต (78.72%)
- เชิงลบ 14,797 ทวีต (21.2%)

Features

- Unigram 5,068 features
- Bigram 66,291 features
- Unigram + Bigram 71,259 features

Test Data

- จำนวน 100 ทวีต
- เชิงบวก 50 ทวีต (50.00%)
- เชิงลบ 50 ทวีต (50.00%)

Results

ความถูกต้อง	Unigram	Bigram	Unigram + Bigram
โดยรวม	82%	81%	87%
กลุ่มเชิงบวก	78%	80%	78%
กลุ่มเชิงลบ	86%	82%	96%

สรุป

สรุปผลการทดลอง

- การใช้ Unigram + Bigram ให้ผลลัพธ์ดีที่สุด
- สามารถจำแนกข้อมูลเชิงลบได้ถูกต้องมากกว่าเชิงบวก
- Train Data เชิงบวกมากกว่าเชิงลบมาก
- มีการใช้ Emoticon เชิงบวกกันอย่างพร่ำเพรื่อ

ปัญหาที่พบ

- การจำแนกด้วย Emoticon ไม่เหมาะสมกับภาษาไทย
- คำศัพท์แปลกๆ จำนวนมากที่ไม่สามารถตัดได้อย่างถูกต้อง

ข้อเสนอแนะ

- ควรเลือก Train Data ให้ดีขึ้น
  - อาจใช้ Semi-Supervised
- ควรเพิ่ม Class จำแนกอารมณ์เฉยๆ ด้วย
- ต่อยอดโดยนำไปจำแนกข้อความทั่วไป เช่น บทความ

ถาม-ตอบ

ขอบคุณครับ