Sentiment Analysis in Chinese Social Media

Course: COMP8755 Individual Computing Project
Supervisor: Lizhen Qu
Student: Nan Xie,u5908569
What is Sentiment Analysis?

- Sometimes known as opinion mining or emotion AI.
- Aims to systematically identify a writer (speaker, or other subject)’s emotion.

For example:

“This restaurant is good !” -> positive
“This restaurant is dreadful !” -> negative
Why Sentiment Analysis?

• A Rational Guide
  – Sentiment Analysis for e-commerce data (review, comment or feedback): Mining about people’s preference for different products to help manufacturer makes better decisions.
  – Sentiment Analysis for social media: Mining people’s opinion about a public figure or a particular event to helps public relations or marketing department doing better work.
Why Chinese Social media?

- Chinese netizen’s population is huge.
- Popularity of social media and e-commerce in China.
- Research about sentiment analysis in Chinese language is rarely.
Characteristics of Chinese Language-PinYin

- Pin Yin is the official romanization system for Chinese.
- Pin Yin input method is one to many map.
  - People may makes typed error.

![Input Method Example]

- la'ji
- 垃圾 (Rubbish, Dreadful)
- 辣鸡 (Spicy chicken)
Characteristics of Chinese Language - PinYin

- **Cyberspeak**
  - Netizen sometimes like replace a word by another word with different meaning but with similar pronounce (similar in PinYin).
  
  For example:

  Following two sentences have same meaning which are “This restaurant is dreadful”

  - 这家/餐馆/真/垃圾 (zhejia/canguan/zhen/laji)
  - 这家/餐馆/真/辣鸡 (zhejia/canguan/zhen/laji)
  - zhèjiā/cānguǎn/zhēn/lājī
d  - zhèjiā/cānguǎn/zhēn/làjī
PinYin Net

• Hypothesis: pin yin is helpful additional information to helping understand writer’s intention.

• Word representation in PinYin Net (concatenated word with it’s pin yin)

| Word embedding | PinYin embedding |
PinYin Net

- Architecture
Result and Discuss

Lazy Pin Yin: without tone
Result and Discuss

![Graph 1: Only word embedding vs word embedding concatenated with pinyin embedding (loss)]

- Blue line: Only word embedding
- Orange line: Word embedding concatenated with pinyin embedding

![Graph 2: Only word embedding vs word embedding concatenated with lazy pinyin embedding (loss)]

- Blue line: Only word embedding
- Orange line: Word embedding concatenated with lazy pinyin embedding
Result and Discuss

![Graph showing comparison between Only word embedding and Only lazy pinyin embedding for accuracy and loss.](image)
Result and Discuss

• PinYin Net didn’t give a significant improvement in accuracy.

• PinYin Net is able to give good performance in very few epoch (Which means easier to train).

• Using lazy pin yin only, still able to give similar result. (Since, one to many map. Vocabulary of lazy pin yin is much less than vocabulary of words, which means train lazy pinyin only is much faster (computational cheaper).)
Thanks!