

# Aspect-Based Sentimental Analysis On Customer Reviews

Group 12

Nguyen Hai Dang   Le Thi Trang   Le Tran Ngoc Minh

John von Neumann Institute 2018

July 3, 2019

# Table of Contents

- 1 Introduction
- 2 Solution
- 3 Implementation
  - Prepossessing
  - Aspect Analysis
  - Sentimental Analysis
- 4 Experimental Result
- 5 Conclusion

# Outline

## 1 Introduction

## 2 Solution

## 3 Implementation

- Prepossessing
- Aspect Analysis
- Sentimental Analysis

## 4 Experimental Result

## 5 Conclusion

# Introduction & Data Description

- Objective: Doing aspect and sentiment analysis for customer reviews
  - ① Identify the topic that is mentioned in each customer feedback
  - ② Identify the polarity of each comments (whether it is positive or negative)
- Data: data is customer reviews, crawled from the website **foody.vn**, including two data sets  
(source: <https://forum.machinelearningcoban.com>)
  - ① The first one contains 30,000 reviews with label *positive* or *negative*, each are written in a separated text file
  - ② The second one contains 174,437 reviews from different shops without labels

# Introduction & Data description

	review_content
40	Nghe danh toco toco khá nhiều, nay uống thử. Menu rất nhiều món để lựa chọn. Quán ở đây làm thức uống có vẻ nhạt sao ah. Mặc dù mình đã đề 70% đườ...
41	NGHE BẠN BÈ GIỚI THIỆU SUSHI TEI ĐỒNG DU MỚI MỞ NÊN MÌNH VỚI MẸ ĐƯA BẠN VÀO ĂN THỬ HIHI\NHÀ HÀNG RẤT LÀ ĐẸP LUÔN NHA MẸY BẠN, THÍCH HỢP CHUP HIN...
42	Mình tới mua mang về vào lúc trưa 1h53, phần của mình ord 1 kimbab và 1 nudu kimbab và cái mình thất vọng ở đây là quán làm quá lâu trong khi mình...
43	Nhân viên nhà hàng nhiệt tình, vui vẻ 🍻🍻🍻🍻🍻🍻. Món ăn tươi ngon, thực đơn đa dạng và giá cả hợp lý. Món ăn phục vụ nhanh , và món ăn yêu thích của ...
44	Bánh Đức Lá Dứa Hôm Nay Sĩ & Lẻ\Giao Hàng Tận Nơi MIỄN PHÍ Giao Hàng \nĐảm bảo GIÒN DAI trong bánh KHÔNG SỬ DỤNG CHẤT ĐỘC HẠI như : HÂN THE, PHÂM...
45	Bánh Đức Lá Dứa Hôm Nay Sĩ & Lẻ\Giao Hàng Tận Nơi MIỄN PHÍ Giao Hàng \nĐảm bảo GIÒN DAI trong bánh KHÔNG SỬ DỤNG CHẤT ĐỘC HẠI như : HÂN THE, PHÂM...
46	Lướt foldy vô tình thấy quán này. Thấy hình ảnh hấp dẫn quá nên mình đặt thử. Đặt 1 tré trộn cóc non và xoài non xốc muối Thái. Tré khá ngon, có c...
47	Không gian đẹp quán mới mở nhưng khá đông. Đồ uống vừa phải không quá ngọt thơm vị trà đặc biệt ở đây có milk farm sầu riêng ngon ngay thơm mùi sả...
48	Done 🍷🍷🍷
49	Trưa trưa thèm trà sữa nên mấy ce order uống, mình chọn trà sữa trân châu truyền thống.. ngon.. chai đẹp còn có thêm túi zip xinh xinh nữa mà có 2...

Figure 1: Original unlabeled data

# Outline

1 Introduction

2 Solution

3 Implementation

- Prepossessing
- Aspect Analysis
- Sentimental Analysis

4 Experimental Result

5 Conclusion

## 2. Solution

Firstly, the data is preprocessed. Since a customer feedback usually contains many sentences. We assume that each sentence represents an aspect.

Secondly, we separate our problem into two tasks:

- Aspect analysis: Identify the aspect which are mentioned in each sentence in unsupervised approach.
- Sentiment analysis: Identify the polarity of each sentence (whether it is positive or negative) in supervised approach.



Figure 2: System workflow

# Outline

- 1 Introduction
- 2 Solution
- 3 Implementation**
  - Prepossessing
  - Aspect Analysis
  - Sentimental Analysis
- 4 Experimental Result
- 5 Conclusion



# Outline

- 1 Introduction
- 2 Solution
- 3 Implementation**
  - Prepossessing
  - Aspect Analysis
  - Sentimental Analysis
- 4 Experimental Result
- 5 Conclusion

# Reviews on Foody

```
Đi ăn rồi ghé qua đây uống trà sữa luôn hehe . Không gian đẹp , giá cả hợp lí . Ở đây còn có bingsu , lần sau ghé sẽ  
=====  
Món dễ cơm ngon , thơm tuyệt vời nhất mà mình dc ăn  
=====  
Hom trc có dan bạn gái đi ăn , bạn mình khoái món này lắm . Thay dòng dòng ghé vào thì mình thay chất lượng k như mọi  
=====  
Ăn cũng tạm tạm , không ngon như các bạn miêu tả . Không gian hơi bí , chỗ ngồi cũng chật nữa . Trc m đi vs 4 đứa bạn  
=====  
Hồi mới mở quán trông mấy bộ bàn ghế nhỏ xinh đáng yêu , các bạn nhân viên gọn gàng và cũng là quán phong cách hiện đ  
Nhưng đồ ăn uống thực sự chưa ấn tượng , giá hơi mắc : (
```

Figure 3: Reviews on Foody

'Đà Nẵng dạo này mở nhiều quán kem, xinh xắn đẹp mắt nhưng ăn ở đây từ thời scoop kem còn 7k, 14k. Dù ít ghé nhưng vẫn nhớ mãi cái quán kem màu tím xinh xinh, vị kem cũng ngon. Và chiều bất chấp ghé quán lúc chiều. Má ơi, quán vắng tanh nhưng thất vọng trước hết là bàn ghế đã quá cũ kĩ, nhìn nhợt nhúa, nhếch nhác. Quán cũng ko buồn sửa lại nữa. Màu cũ kĩ bao trùm cả quán. Nhưng ráng vì kỉ niệm cũng bước vào. Scoop kem bây giờ đã 25k. Mình chọn 1 scoop 25k và một Suri frie nd 55k. Order trước rồi lên tầng 2 ngồi ( vì tầng 1 nâng chiếu vào và nhếch nhác tồi tàn lắm). \nLên tầng 2, ko gian trên cao thôi, còn lại ko khá hơn. Có khi tệ hơn. Mùi cũ kĩ, ẩm mốc, bàn g hế đầy bụi. Lựa chọn chỗ ngồi, vẫn nóng giữa tiết trời đáng lý ra phải chui vào điều hoà mát lạ nh. Thấy một dãy quạt đứng, nhưng ko có cái nào dùng đc cả. Đợi nhân viên mang kem lên nhờ cho nguồn điện giùm chứ nóng và ko dùng quạt đc. Nhân viên nam mặt trời ra, ko đáp cũng ko trả lời. \nKo sao, chắc có lẽ ảnh nặng tai. Nói lại lần nữa, vẫn thái độ ko ưa đc. Chưa kể mình xin 2 ly nc lọc bỏ đá vào. Vẫn còn mặt trời ra. Nói lại lần nữa thì gục gục rồi đi xuống và mất tâm. Còn mình thì ngồi đợi quạt và nước đến héo người. Bực mình dễ sợ. Có hư hại hay gì thì ít chi cũng nên nói lại cho ng ta bik, ai đời có cái thái độ này. Kem cũng như xưa, ko đổi gì mấy nhưng chắ ng thấy ngon vì tan nhanh và chắc ko nuốt nổi cái ko gian nhợt nhúa, cũ, nóng, dơ dơ. \nChờ đợi hoài ko thấy phản hồi, mình xún góp ý cho quán. Nhân viên nữ thu ngân ko khá hơn. Cười cười cho qua ko bik nói lại gì, về mặt lại kênh thương hại. Nhân viên nam lúc này nghe mình góp ý cũng y chang bộ mặt đó. \nHaizzz, thật đáng buồn. Quán đi xuống là điều tất yếu bởi các yếu tố này. Hè n gì. Thật tội nghiệp và cặn lời. \nKo gian ko thêm sửa sang, nhân viên ko được gì, đến quạt và nc uống xin thêm đá còn ko cho. Mình nghĩ quán nên có cái đầu quản lý tinh táo hơn hoặc là cù l ẫn chấp nhận chỉ là kỉ niệm hoài ức của người Đà Nẵng xưa. \nMọi ng ko nên ghé vào, buồn lắm đấ y. Một thời của mình. \nVà mình chắc chắn, ko bao giờ quay lại với dịch vụ ko văn minh như vậy r òi. Kỉ niệm cũng chịu. Bik bao nhiêu quán đang chờ ta, đâu có đáng, giá cũng có rẻ đâu. \nNg ta bỏ tiền đi ăn chứ có đi xin đâu? Ha'

Figure 4: Negative Review

# Challenging

- Do not have standards stop words
- Icons
- Wrong Spelling.
- Wrong Decode.
- No spelling mark or bi-language sentences.

# Stop Words Selecting

- Checking vn-stopwords.
- Retain importance's advert words on sentimental analysis domain.
- Removing high occurrences but unimportant words

# Words Occurrences Count

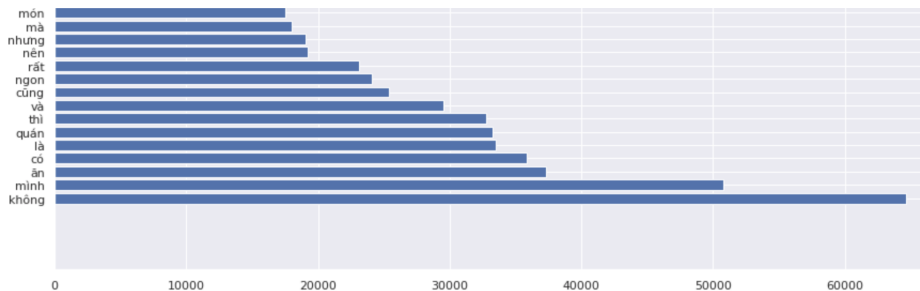


Figure 5: Word Occurrences

# Stop Words Selecting

```
stop_words = 'ngoài lấy nha giờ con chả ngay cứ có_thể đâu nói_chung thôi mấy lắm cái ăn cứ vs  
những chẳng chẳng anh chị bạn sao bên bảo toàn nhé kiểu vậy chắc xong chứ lúc nữa được nhận rằn  
g quá riêng gì muốn rồi số thấy hay lên lần nào qua bằng điều biết lớn khác vừa nếu thời gian h  
ọ từng đây năm tháng trước chính cả việc chưa do nói ra nên đều đi tới tôi có_thể cùng vì làm l  
ại mới ngày đó vẫn mình chỉ thì đang còn bị mà nhất hơn sau phải như trên tại theo khi nhưng và  
o đến từ sẽ ở cũng về để này những một các cho được với có trong đã là và của thực_sự ở_trên tấ  
t_cả dưới hầu_hết luôn giữa bất_kỳ hỏi thường ai cảm_ơn'  
stop_words = stop_words.split()
```

Figure 6: Stop words

# Outline

1 Introduction

2 Solution

3 **Implementation**

- Prepossessing
- **Aspect Analysis**
- Sentimental Analysis

4 Experimental Result

5 Conclusion



# Word Embedding: Word2Vec

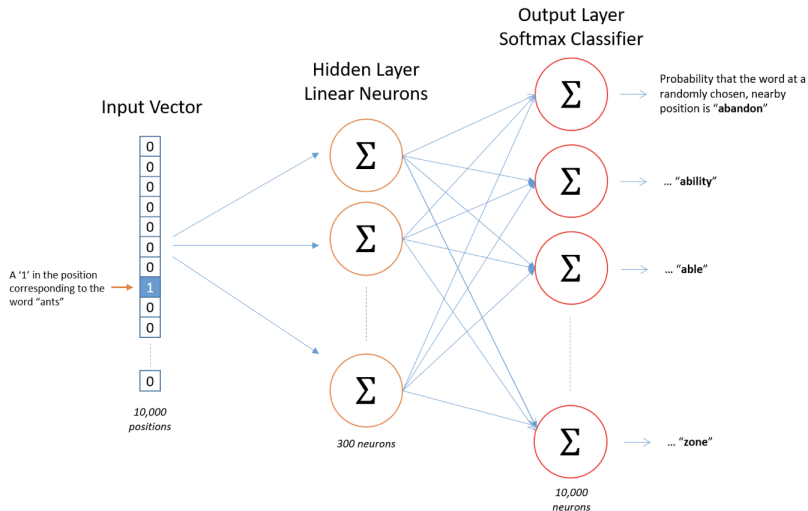


Figure 7: Word2Vec Models

# Hyper Parameters

- **Hidden Layer Nodes:** Number of dimensions that our vector will present the word. In practical, we finding that the neuron works best in 300 hidden nodes.
- **Min-Count:** The threshold of word's occurrences. We want to eliminate the unimportant word, which could lead us into false clustering topics. We set it to 500.
- **Windows:** The distances around words to get the cor-occurrences probability. We set it to 2.

# Similar Vectors

```
# ,negative = ['không ngon'],topn = 100)
w2v_model.most_similar(positive = ['ngon'],topn=100)

/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:2: DeprecationWarning: Call to deprecated 'most_similar' (Method will be removed in 4.0.0, use self.wv.most_similar() instead).

[('vừa_miệng', 0.7536715269088745),
 ('dậm_đà', 0.7421907782554626),
 ('lạ_miệng', 0.7152743935585022),
 ('thơm', 0.7108073830604553),
 ('dậm_vị', 0.6667570471763611),
 ('đặc_biệt', 0.6557033658027649),
 ('ngọt', 0.6503500938415527),
 ('mềm', 0.6491641402244568),
 ('mặn', 0.633951723575592),
 ('hợp_khẩu_vị', 0.6339314579963684),
 ('béo', 0.6318671703338623),
 ('vừa_phải', 0.6309441328048706),
 ('dậm', 0.6251562833786011),
 ('ngán', 0.578175961971283),
 ('giòn', 0.5757705569267273),
 ('dẻo', 0.5748008489608765),
```

Figure 8: Similar Vectors

# Visualization



# Clustering

```
Không Gian :      không_gian , đẹp , chỗ_ngồi , bàn_ghế , rộn  
Chụp Hình :      sang , nhìn , hoa , menu , tên , hấp_dẫn , t  
Trivial ? :      liền , trc , thì_phải , hồi , người , khách  
Bánh :           bánh , mua , dùng , voucher , trái_cây , bán , pi  
Dịch Vụ :        dịch_vụ , tốt , phong_cách , đầu_tiên , thực_s  
Món ăn (quán) :  món , cơm , sốt , chua , lẩu , bún , tươ  
Phục Vụ :        nhân_viên , thân_thiện , phục_vụ , nhiệt_tình  
Uống :           uống , trà_sữa , nước , trà_đào , kem , dâu , sữa  
Đánh giá 1 :     giá , tuy , cao , hơi , mắc , nhỏ , thức_ăn  
Phô mai :        thơm , béo , giòn , ngọt , cục , phô_mai , khô  
Dọn dẹp, order : gọi , đem_ra , thêm , đem , dọn , nguyê  
Hành động đi ăn : địa_điểm , lần_sau , ghé , thử , các_b  
Đánh giá 2 :     chọn , ở_đây , tuyệt_vời , chung , thích ,
```

Figure 10: 12 Clusters by Gaussian Mixture Models

```
('Thử cho biết chữ mình thấy mix tùm lum ăn không ngon, được cái tàu hủ với '
'nước dừa béo, quán nhỏ thôi, được 6,7 cái bàn, nhà mình đi 5 người mà xài hết '
'3 cái rồi, ai đi đông nên cân nhắc nha, vị trí dễ tìm, cử công viên thờ '
'trắng rồi thấy bánh sầu riêng durio là đi vô thôi.')
['Đồ uống', 'Phục vụ', 'Đồ ăn', 'Không gian']
=====
('Udon gà nướng tiêu đen với udon xé xiu khô, hai món đều ngon, bạn mình khen '
'riú rít, phục vụ nhanh lẹ chắc mình đi buổi trưa nên k có ai :))) Cho kèm 2 '
'chén súp ok lắm, sẽ ghé quán thường xuyên luôn 🍜🍜🍜')
['Phục vụ', 'Đồ ăn', 'Không gian']
=====
('Thời gian làm mọi thứ bằng qua cuộc đời ngắn ngủi của chúng ta, chỉ có những '
'khoảng khắc giữ chúng ta lại tại một vùng trời gọi là quá khứ.')
['Đồ ăn']
=====
'Quán có cfe mạn đặc biệt .'
['Đồ ăn', 'Phục vụ', 'Không gian']
=====
('Đọc review thấy mọi người khen cũng có chê cũng có nên quyết định đi ăn thử '
'cho biết mặc dù rất ít khi ăn chay.')
['Đồ uống', 'Phục vụ', 'Đồ ăn']
=====
'Năm ở Nguyễn Thị Thập q7\NQuán ăn có 2 tầng , điều hoà , wifi ok .'
['Phục vụ', 'Đồ ăn', 'Không gian']
=====
'Lần đầu tiên mình duoc ăn loại tóp mỡ đặc biệt này to giòn phồng, mập.'
['Đồ ăn']
=====
'Mình đặt 1 bông lan chà bông sốt bơ, 1 bông lan phô mai và 1 flan gatoux (?).'
['Đồ ăn']
=====
('Sau hơn một năm nhượng lại cho 101 tea n coffee, cuối cùng chago cũng trở '
'lại hehe.')
['Đồ uống']
```

Figure 11: Result

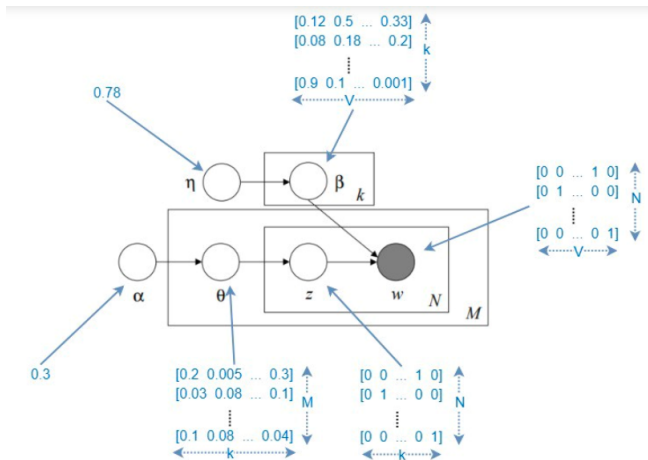


Figure 12: Graphical model of the LDA

With the notations as follow

- $k$ : Number of topics a document belongs to (a fixed number)
- $V$ : Size of the vocabulary
- $M$ : Number of documents
- $N$ : Number of words in each document
- $w$ : A word in a document. This is represented as a one hot encoded vector of size  $V$  (i.e.  $V$ : vocabulary size)
- $\mathbf{w}$  (bold  $w$ ): represents a document (i.e. vector of “ $w$ ”s) of  $N$  words
- $D$ : Corpus, a collection of  $M$  documents
- $z$ : A topic from a set of  $k$  topics. A topic is a distribution words. For example it might be



- 1 Split cleaned reviews into separated sentences based on punctuation (.,!?)
- 2 Tokenize each review into separated words and remove stop words
- 3 Build dictionary from the tokenized words in each review
- 4 Create Term Frequency-Inverse Document Frequency(TF-IDF) matrix from the corpus.
- 5 Feed the dictionary and TF-IDF matrix into the LDA model (use gensim package in Python) and try with different number of topics
- 6 Choose the best number of topics and apply the result to each sentence to get back the topic with the highest probability.

## Best number of topics: 6

Topic: 0 Word: 0.019\*"trà\_sữa" + 0.019\*"trà" + 0.015\*"uống" + 0.012\*"vị" + 0.010\*"ngon" + 0.010\*"ngot" + 0.010\*"trần\_châu" + 0.008\*"thơm" + 0.008\*"kem" + 0.008\*"thích"

Topic: 1 Word: 0.041\*"nhân\_viên" + 0.031\*"phục\_vụ" + 0.025\*"nhiệt\_tình" + 0.016\*"thân\_thiện" + 0.014\*"nhanh" + 0.011\*"khách" + 0.011\*"nhanh\_nhen" + 0.011\*"dễ\_thương" + 0.008\*"lâu" + 0.008\*"quán"

Topic: 2 Word: 0.008\*"quán" + 0.006\*"trà\_sữa" + 0.006\*"đi" + 0.005\*"ở" + 0.005\*"mới" + 0.005\*"đến" + 0.005\*"ăn" + 0.005\*"thử" + 0.004\*"lần" + 0.004\*"gong\_cha"

Topic: 3 Word: 0.020\*"giá" + 0.016\*"giá\_cả" + 0.015\*"uống" + 0.014\*"đồ" + 0.012\*"chất\_lượng" + 0.011\*"rẻ" + 0.011\*"ngon" + 0.010\*"ăn" + 0.008\*"món" + 0.008\*"ổn"

Topic: 4 Word: 0.019\*"quán" + 0.017\*"không\_gian" + 0.012\*"ngồi" + 0.011\*"đẹp" + 0.010\*"tầng" + 0.009\*"đi" + 0.008\*"chỗ" + 0.008\*"đông" + 0.007\*"rộng\_rải" + 0.007\*"ở"

Topic: 5 Word: 0.017\*"ăn" + 0.016\*"ngon" + 0.010\*"gà" + 0.009\*"thịt" + 0.009\*"xôi" + 0.009\*"món" + 0.007\*"bánh" + 0.007\*"nhiều" + 0.007\*"gỏi" + 0.007\*"nướng"

Figure 13: LDA model result

	sentence	aspect
0	quán steak hiếm_hoi mà mình cực_kì ưng_ý từ lâu_nay khi quán bên Nguyễn_siêu_bữa_nay đi thử chi_nhánh bên võ_thị sáu mà ưng quá	ambiance
1	quán nằm trong_hẻm xe_hơi_hướng võ_thị sáu chiều mình đi hai_bà_trung queo xuống hẻm nằm gần quán cafe higland nên dễ_tìm lắm	ambiance
2	quán có bảng_hiệu khá nhỏ nhưng do hẻm cụt nên cũng dễ nhận biết quán	ambiance
3	phục_vụ niềm_nở chào_đón và phục_vụ chu_đáo	staff
4	nói tới phần thức_ăn mình gọi mac cheese béo ngậy phô_mai phần thịt steak tendeloin achentina và striploin g kèm salad phô_mai	food
5	phải nói là rất hài_lòng vì thịt mềm và juicy lắm thêm cái khoản tendeloin nên bò mềm phải_biết nha	food
6	giá steak khá tốt không quá mắc so_với các nhà_hàng steak nổi_tiếng và sang_chảnh nhưng steak về chất_lượng thì mình không bàn_cãi nhiều	price

Figure 14: LDA result for unlabeled data set

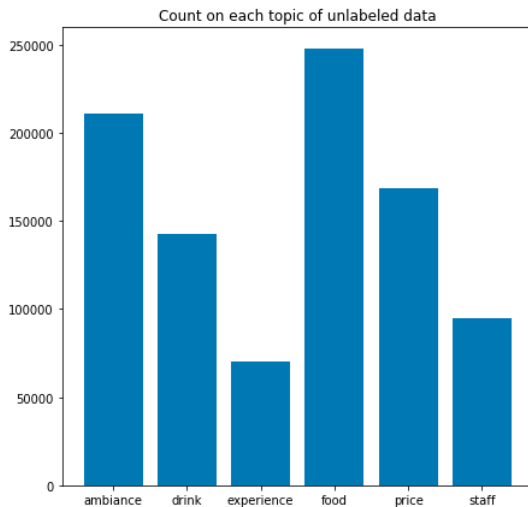


Figure 15: Count on each topic of unlabeled data

# LDA vs Word2Vec Cluster

- Word2Vec generate more accurate cluster than LDA with small numbers of corpus. But LDA is out perform Word2Vec when it comes to larges data set.
- The Word2Vec assigned each word on fixed topics then it will have higher bias than LDA.

Therefore, we choose the LDA.

# Outline

- 1 Introduction
- 2 Solution
- 3 Implementation**
  - Prepossessing
  - Aspect Analysis
  - **Sentimental Analysis**
- 4 Experimental Result
- 5 Conclusion

# Sentiment Analysis

The labeled data set which contains 50,000 reviews was split into 3 subsets:

- 30,000 reviews for training.
- 10,000 reviews for validation.
- 10,000 reviews for testing.

Mình tập\_thể\_dục xong qua làm một phần . 49k mà nhiều quá\_trời luôn .  
Chả cua với thịt ăn vừa\_miệng . Nước\_mắm được hâm\_nóng , ăn hơi lạt .  
Rau xanh\_tươi , sạch . Quán đang mừng khai\_trương nên khuyến\_mãi  
nước\_ngọt nữa . Chén\_đũa để sạch\_sẽ , gọn\_gàng . Nhân\_viên lịch\_sự

Figure 16: A sample review

We use 3 models: Logistic Regression, Neural Network, Bidirectional LSTM.

# Sentiment Analysis

## RNN

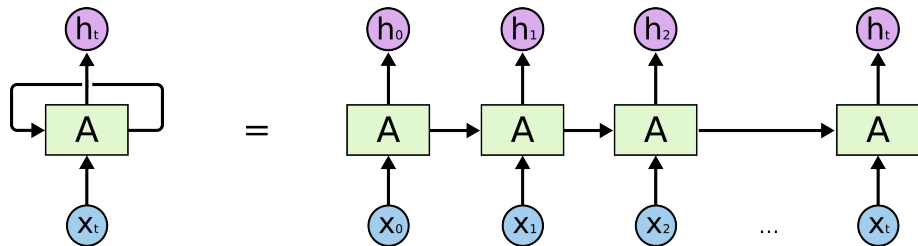


Figure 17: An unrolled recurrent neural network <sup>1</sup>

<sup>1</sup><https://colah.github.io/posts/2015-08-Understanding-LSTMs/>



# Sentiment Analysis

## RNN

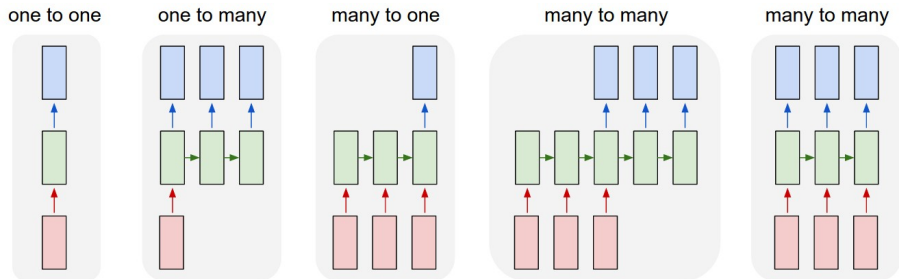


Figure 18: RNN process sequences <sup>2</sup>

<sup>2</sup><http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# Sentiment Analysis

## RNN

We can process a sequence of vectors  $\mathbf{x}$  by applying a **recurrence formula** at every time step:

$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state / old state input vector at some time step

some function with parameters  $W$

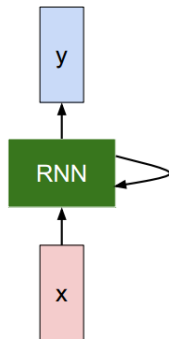


Figure 19: RNN reference fomular <sup>3</sup>

<sup>3</sup><http://cs231n.stanford.edu/>

# Sentiment Analysis

## RNN

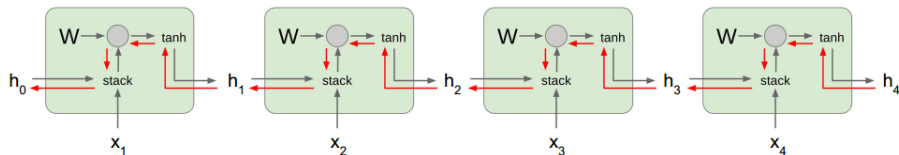


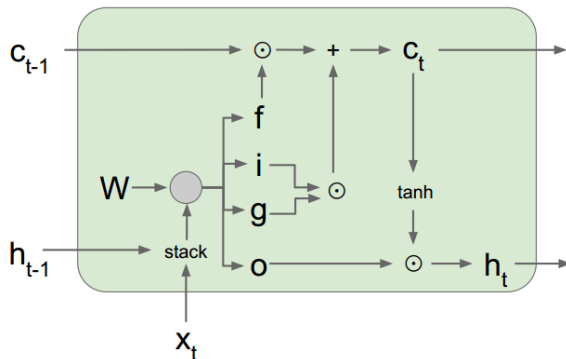
Figure 20: RNN gradient flow <sup>4</sup>

Computing gradient of  $h_0$  involves many factors of  $W$  (and repeated  $\tanh$ ).  
 $\Rightarrow$  Vanishing gradients.

<sup>4</sup><http://cs231n.stanford.edu/>

# Sentiment Analysis

## LSTM



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

Figure 21: LSTM architecture <sup>5</sup>

## Understanding LSTM backpropagation

<sup>5</sup><http://cs231n.stanford.edu/>

# Sentiment Analysis

## Bidirectional LSTM

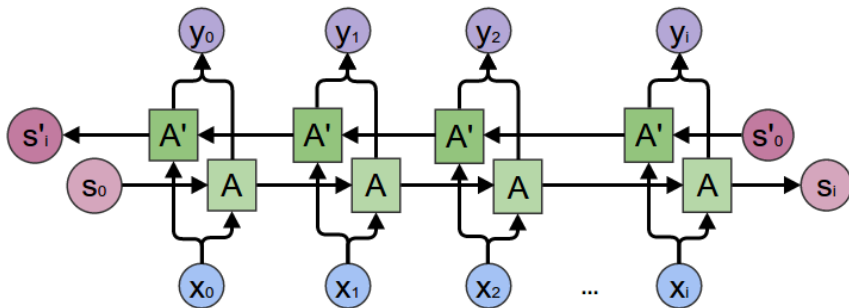


Figure 22: Bidirectional LSTM architecture <sup>6</sup>

<sup>6</sup><http://colah.github.io/posts/2015-09-NN-Types-FP/>

# Sentimental Analysis

We train 7 binary classifiers to compare the performance and select the most appropriate classifier. We then use the selected classifier to predict the sentences retrieved from Aspect Analysis step. Following are our classifiers:

- C1: Logistic Regression using bag-of-word.
- C2: Neural Network using bag-of-word.
- C3: Neural Network with embedding layer.
- C4: Neural Network with pre-trained word embedding in labeled dataset.
- C5: Neural Network with pre-trained word embedding in unlabeled dataset.
- C6: Bidirectional LSTM with pre-trained word embedding in labeled dataset.
- C7: Bidirectional LSTM with pre-trained word embedding in unlabeled dataset.

# Sentiment Analysis

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 10)	535100
dense_4 (Dense)	(None, 1)	11
Total params: 535,111		
Trainable params: 535,111		
Non-trainable params: 0		

Figure 23: C2: Neural Network using bag-of-word.

# Sentiment Analysis

Vocabulary size: 60538

Max length of review: 1149

Building model...

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 1149, 100)	6053800
global_max_pooling1d_1 (Glob	(None, 100)	0
dense_2 (Dense)	(None, 10)	1010
dense_3 (Dense)	(None, 1)	11

Total params: 6,054,821

Trainable params: 6,054,821

Non-trainable params: 0

Figure 24: C3: Neural Network with embedding layer.



# Sentiment Analysis

```
Vocabulary size: 60538
Max length of review: 1149
Building embedding matrix...
Building model...
Trainable: False
```

Layer (type)	Output Shape	Param #
embedding_4 (Embedding)	(None, 1149, 100)	6053800
global_max_pooling1d_4 (Glob	(None, 100)	0
dense_7 (Dense)	(None, 10)	1010
dense_8 (Dense)	(None, 1)	11

```
Total params: 6,054,821
Trainable params: 1,021
Non-trainable params: 6,053,800
```

Figure 25: C4: Neural Network with pre-trained word embedding in labeled dataset.

# Sentiment Analysis

```
Vocabulary size: 170104
Max length of review: 1149
Building embedding matrix...
Building model...
Trainable: True
```

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, 1149, 100)	17010400
global_max_pooling1d_6 (Glob	(None, 100)	0
dense_11 (Dense)	(None, 10)	1010
dense_12 (Dense)	(None, 1)	11

```
Total params: 17,011,421
Trainable params: 17,011,421
Non-trainable params: 0
```

Figure 26: C5: Neural Network with pre-trained word embedding in unlabeled dataset.

# Sentiment Analysis

```
Vocabulary size: 60538
Max length of review: 1149
Building embedding matrix...
Building model...
Trainable: False
```

Layer (type)	Output Shape	Param #
embedding_13 (Embedding)	(None, 1149, 100)	6053800
bidirectional_7 (Bidirection	(None, 128)	84992
dense_19 (Dense)	(None, 1)	129

```
Total params: 6,138,921
Trainable params: 85,121
Non-trainable params: 6,053,800
```

Figure 27: C6: Bidirectional LSTM with pre-trained word embedding in labeled dataset.

# Sentiment Analysis

```
Vocabulary size: 170104
Max length of review: 1149
Building embedding matrix...
Building model...
Trainable: False
```

Layer (type)	Output Shape	Param #
embedding_11 (Embedding)	(None, 1149, 100)	17010400
bidirectional_5 (Bidirection	(None, 128)	84992
dense_17 (Dense)	(None, 1)	129

```
Total params: 17,095,521
Trainable params: 85,121
Non-trainable params: 17,010,400
```

Figure 28: C7: Bidirectional LSTM with pre-trained word embedding in unlabeled dataset.

# Sentimental Analysis

Model	Train accuracy	Validation accuracy	Test accuracy	Epochs
C1	0.909	0.875	0.879	
C2	0.972	0.853	0.853	10
C3	0.999	0.864	0.867	10
C4	0.800	0.787	0.791	30
C5	0.956	0.843	0.844	50
<b>C6</b>	<b>0.930</b>	<b>0.879</b>	<b>0.886</b>	10
<b>C7</b>	<b>0.929</b>	<b>0.873</b>	<b>0.885</b>	10

Table 1: Performance of 7 classifiers in labeled dataset

Related work achieved the accuracy of 87.64% with LSTM model <sup>7</sup>. Since the data we want to predict is in the unlabeled dataset, we choose the classifier **C7** to utilize the pre-trained word embedding in the unlabeled dataset.

<sup>7</sup><https://streetcodevn.com/blog/sav>

# Sentimental Analysis

Sentence: menu nhiều loại salad lạ dị với giá cả k và các loại cuộn k nha

Label: Positive

Sentence: đại\_lộc bánh\_tráng thịt\_heo & mì quảngtrưa nay nhác về nhà ăn mà lại thêm mùi mắm quá thể nên các bạn trong văn\_phòng khao\_nhau order mấy suất bánh\_tráng thịt\_heo lên ăn cho đã thêm

Label: Positive

Sentence: hai\_đứa ăn hết hơn k nhưng như đã nói là rất chất và trc khi đi cungz xác\_định r nên thấy bình thường

Label: Negative

Sentence: quán không có đèn lớn đa phần là đèn mắt cáo nên khá tối nên không thích hợp để học\_bài hay làm\_việc cho lắm

Label: Negative

Sentence: ấn\_tượng nhất với chú chủ quán vui\_tính nhiệt\_tình nên đến quán có cảm\_giác gần\_gũi lắm nên ăn cũng ngon miệng nữa

Label: Positive

Sentence: chắc\_chắn sẽ\_quay lại

Label: Positive

Sentence: hai combo có cả khoai\_tây nữa mà có chút\_xiu ăn lướt lấy đã thấy hết trơn

Label: Negative

Sentence: vị nhìn chung không giống lắm có hơi việt hóa nhưng hợp\_khẩu vị

Label: Negative

# Outline

- 1 Introduction
- 2 Solution
- 3 Implementation
  - Prepossessing
  - Aspect Analysis
  - Sentimental Analysis
- 4 Experimental Result
- 5 Conclusion

# Experimental Result

For each shop, we observe all reviews and analyze the reviews to understand what customers mostly talk about.

brand_name	most_common_topic	common_topic_percentage	common_topic_pos_percentage	common_topic_neg_percentage
Hanuri - Quán Ăn Hàn Quốc - Sư Vạn Hạnh	???	24.44	34.59	65.41
Mì Ý Double B	Món ăn	28.44	49.54	50.46
Hebe - Tea & Coffee	Thức uống	42.66	40.51	59.49
IZZIBING Snow Dessert Coffee - Bingsu	Thức uống	31.55	56.22	43.78
Koi Thé Café - Ngõ Đức Kế	Thức uống	39.99	46.99	53.01
Papa's Chicken - Phú Mỹ Hưng	Món ăn	37.64	42.96	57.04
Phúc Long Coffee & Tea House - Lý Tự Trọng	Thức uống	34.17	47.51	52.49
Texas Chicken - Phan Xích Long	Món ăn	28.08	53.39	46.61
Trà Sữa R&B Tea - Ngõ Đức Kế	Thức uống	47.21	36.16	63.84
Xôi Chè Bùi Thị Xuân - Bùi Thị Xuân	Món ăn	26.24	36.93	63.07

Figure 30: An analysis in top 10 most reviewed shops



# Experimental Result

Shop: Xôi Chè Bui Thị Xuân - Bui Thị Xuân

Sentence: xôi chè bùi thị xuân bùi thị xuân quán xôi này hình như lâu đời lắm rồi nên lúc nào cũng khách ra vô liên tục

Topic: Món ăn, Sentiment: Negative

Sentence: quán cũng khá rộng nhưng không gian hơi nóng và bí

Topic: Không gian, Sentiment: Negative

Sentence: phục vụ nhanh chóng nhưng nhân viên không vui vẻ thân thiện cho lắm style đơn mạch

Topic: Phục vụ, Sentiment: Possitive

Sentence: chỗ để xe nhỏ hẹp bảo vệ giữ xe cũng thái độ lồi lõm như các nhân viên phục vụ

Topic: Phục vụ, Sentiment: Negative

Sentence: menu phong phú đa dạng từ xôi đến chè

Topic: Giá - chất lượng, Sentiment: Possitive

Sentence: thấy quán này nổi tiếng lắm xôi dẻo ngon nhưng nước sốt thịt chả lụa không quá đặc sắc vị bình thường

Topic: Thức uống, Sentiment: Negative

Sentence: chè thì ngọt trời ơi đất hỡi luôn

Topic: Thức uống, Sentiment: Negative

Figure 31: An detail analysis in Bui Thi Xuan restaurant

# Experimental Result

Shop: Koi Thé Café - Ngõ Đức Kế

Sentence: gud machiato uống nhiều lần r mà h mới viết rì viu nè rất thích uống black machiato ở koi kem cheese béo béo thơm thơm trà ko quá ngọt rất dịu size nhỏ nhìn nhỏ nhưng chất lm ko hề ít uống size lớn thì xđ là căng bụng luôn ahihi

Topic: Thức uống, Sentiment: Possitive

Sentence: món uống ở đây thì ai cũng biết rồi khỏi chê ngon từng giọt ngọt từng hạt trân châu

Topic: Thức uống, Sentiment: Possitive

Sentence: mình có kinh nghiệm xương máu là uống coi nên chọn đá ít thôi vì chọn đá nhiều quá vị sẽ bị loãng không còn ngon nữa được cái foam ở đây con tuyệt cú mèo

Topic: Thức uống, Sentiment: Possitive

Sentence: koi thé ☺☺☺trưa nay mình ghé koi ngõ đức kế mua trà sữa trân châu hoàng kim size m %sugar

Topic: Thức uống, Sentiment: Possitive

Sentence: ☺☺☺mình rất nghiện trà sữa koi hương vị ko lẫn vào đâu được trân châu bùi bùi dai dai 🍡🍡nhân viên nhiệt tình mình đi tắm h chiều hơi mưa nên quán ko đông khách lắm ko phải đợi lâu hehe ☺☺☺nói chung uống koi quen rồi nên uống trà sữa nơi khác ko quen hihi

Topic: Phục vụ, Sentiment: Possitive

Sentence: hi vọng dịp khác sẽ nán lại ở quán lâu hơn nhé 🍡🍡🍡

Topic: Phục vụ, Sentiment: Possitive

Figure 32: An detail analysis in KOI Cafe

# Experimental Result

Shop: Phúc Long Coffee & Tea House - Lý Tự Trọng

Sentence: không có gì đặc biệt mình chẳng hiểu sao nhiều bạn lại ghiền phúc long đến thế chứ riêng mình thì không thấy có gì đặc sắc ở loại trà này cả chắc có lẽ chọn chưa đúng món ngon nhưng khuyến cáo mấy bạn đừng nên gọi món phúc long cocktail tea không ngon mà còn phải chờ lâu kinh khủng

Topic: Thức uống, Sentiment: Negative

Sentence: best ♥ trà sữa phúc long ngon cực

Topic: Thức uống, Sentiment: Possitive

Sentence: trà xanh đá xay thì toẹt với hơn

Topic: Thức uống, Sentiment: Possitive

Sentence: cream ngon ngon béo béo

Topic: Thức uống, Sentiment: Possitive

Sentence: nhất định sẽ quay lạiiiii ♥♥♥

Topic: Phục vụ, Sentiment: Possitive

Sentence: wtf mình đang uống cái quái gì v

Topic: Giá - chất lượng, Sentiment: Negative

Sentence: ngó qua ngó lại xem phải phúc long không

Topic: Giá - chất lượng, Sentiment: Negative

Figure 33: An detail analysis in Phuc Long Coffee and Tea House

# Outline

- 1 Introduction
- 2 Solution
- 3 Implementation
  - Prepossessing
  - Aspect Analysis
  - Sentimental Analysis
- 4 Experimental Result
- 5 Conclusion

# Conclusion

- We achieve the objective of this project in terms of doing aspect-based sentiment analysis.
- For aspect analysis, LDA generate more accurate cluster than Word2Vec with large number of corpus.
- For sentiment analysis, we have not tuned hyperparameters for the neural networks.
- Customer reviews on e-commerce platforms have unusual characteristics: false marks, no marks, icon,... that need to be handled carefully.

# THANK YOU FOR LISTENING!