Topic modelling and sentiment prediction on customer review data

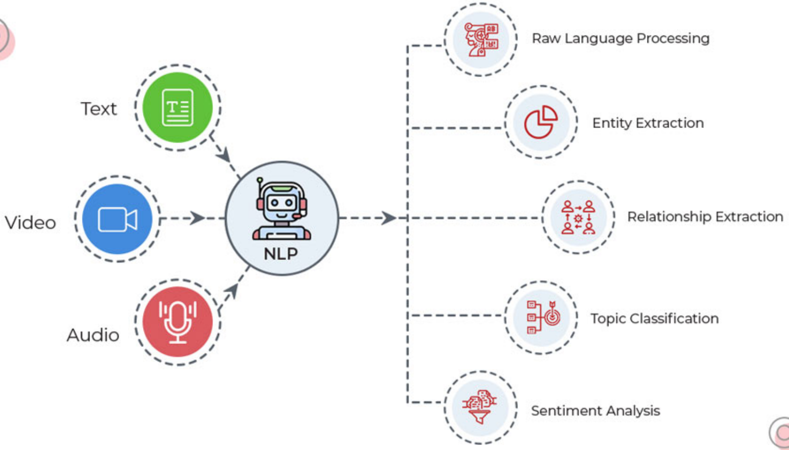
Md Ismail Hossain

Spring 2022: CSE-489 Machine Learning
New Mexico Institute of Mining and Technology

May 9, 2022

Table of Contents

Introduction



Objectives

1. Carry out Exploratory data analysis (EDA) and **Topic modelling** using LDA (latent dirichlet allocation) [Blei et al., 2003] a unsupervised ML techniques to find key topics of the reviews.
2. **Classify** the reviews in terms of sentiments using classiccal and hybrid supervised ML techniques (LSTM, CNN, CNN+LSTM, BERT, Fasttext) [Rehman et al., 2019].

Why Topic modelling and Classification text so important:

- Topic modelling usually used in organizing large block of textual data, information retrieval, feature selection, article recommendation engines.
- Text classification highly important for the field like Social media, marketing, customer experience management, digital media etc.

Data and Tools

Data: Randomly selected 50k Yelp reviews. ▸ Data Source

Data analysis language and environment: Python, Google Colab.

Exploratory Data Analysis (EDA)

Total 50k reviews have been selected from 8.5 million of reviews. Among them 12 states have selected and highest selected from Massachusetts (12,377).

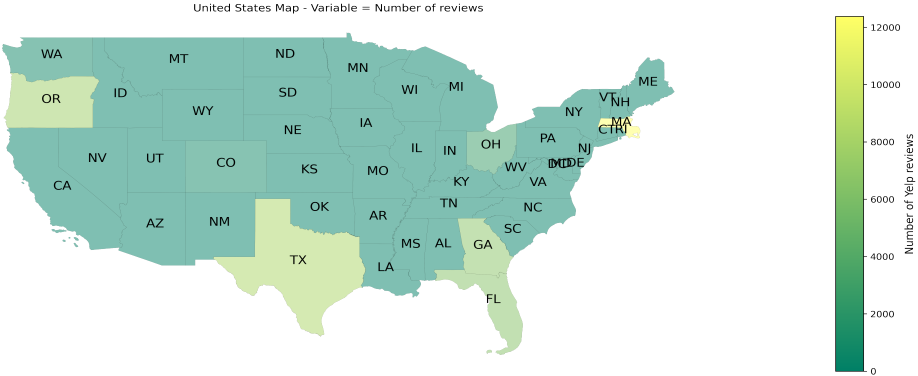


Figure: Selected reviews by states.

Exploratory Data Analysis (EDA)

Reviews by industries:

Figure: Selected reviews by industry.

.

Exploratory Data Analysis (EDA)

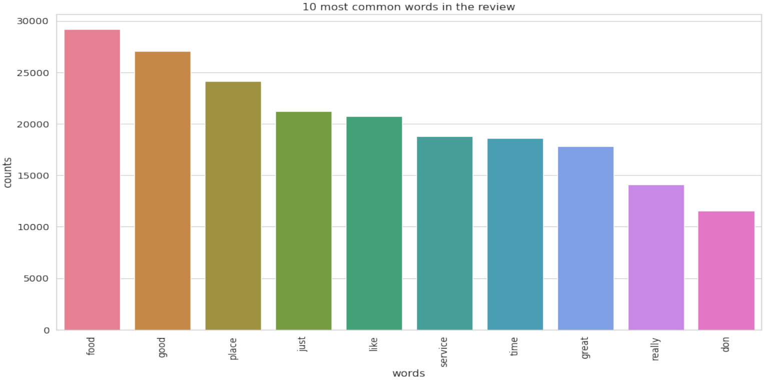


Figure: 10 most common word in the review.

Topic Modelling

- **What** Topic modelling is a unsupervised way to represent text document using several topics.
- **Why** To find unknown or latent topics from unstructured data or texts.
- **How** In this case we used Latent Dirichlet Allocation (LDA) procedure, developed by Prof. David M. Blei in 2003.

LDA: LDA is a probabilistic topic modelling method, used Gibbs sampling in distributing the topics over reviews.

LDA topic modelling try to make:

- Each review as **monochromatic** as possible, meaning representing review by the minimum number of topics.
- Each topic as **monochromatic** as possible, meaning represnting each topic by as small words as possible.

By taking into account these two goals LDA Topic modelling used the Gibbs sampling technique to distribute the words in different topics and distribute the topics in different reviews.

Topic Modelling

Inputs: 'NOUN', 'ADJ', 'VERB', 'ADV' are considered as an input in the algorithm !

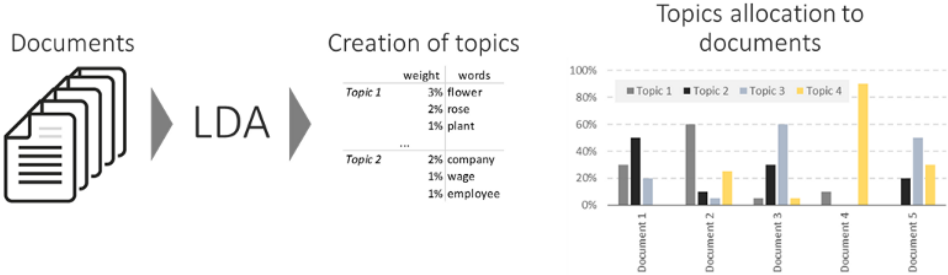


Figure: Topic modelling.

Convert words to numbers for classification

1. Unique numbers
2. One hot encoding
3. Word embedding

Converting text to numbers for classification

Unique numbers and One hot encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

$\rightarrow$

One Hot Encoding

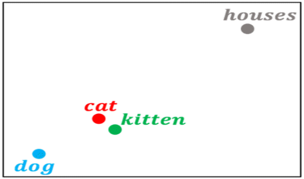
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

.

Word Embedding (TF-IDF, Word2Vec)

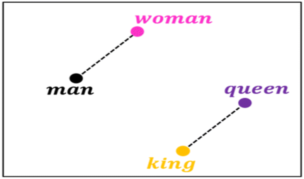
Word	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

Dimensionality reduction of word embeddings from 7D to 2D →



Word	living being	feline	human	gender	royalty	verb	plural
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D →



Word | Word embedding | Dimensionality reduction | Visualization of word embeddings in 2D

Text classification

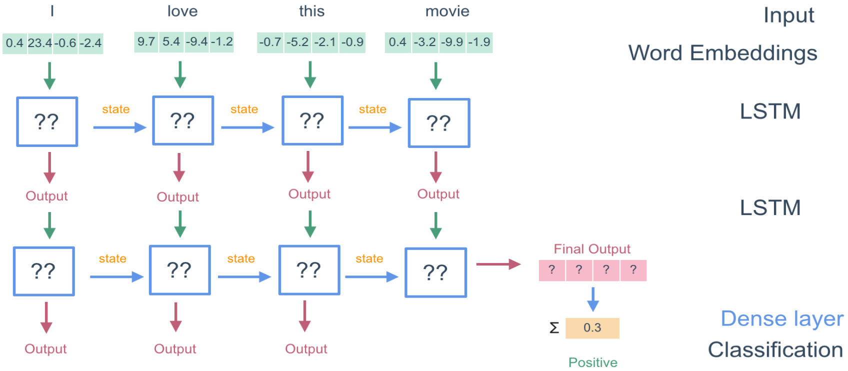


Figure: Architecture of LSTM model.

Text classification

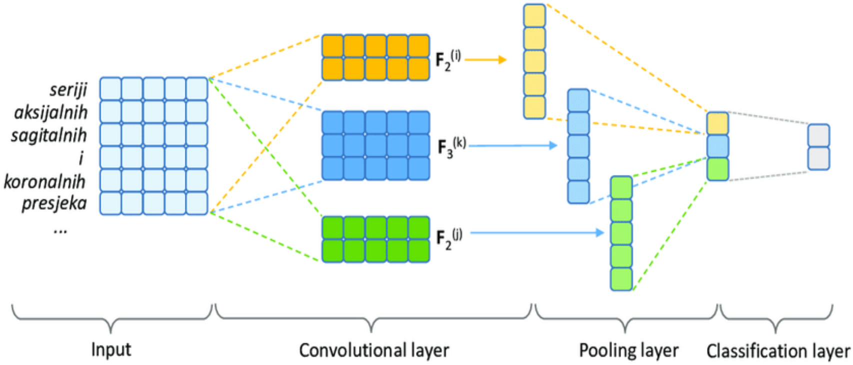


Figure: Architecture of CNN model.

Text classification

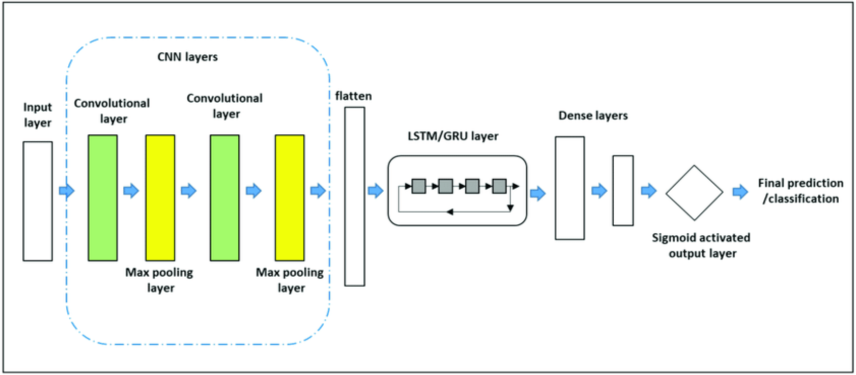


Figure: Architecture of Hybrid CNN+LSTM model.

Text classification

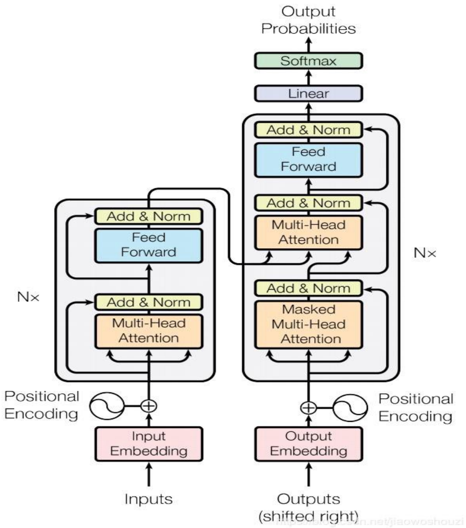


Figure: Architecture of BERT model.

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\beta_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

Figure: Architecture of Fasttext model.

.

Results of Topic Modelling

In our case 3 topic and first 5 words of the topic have considered:

	Word 0	Word 1	Word 2	Word 3	Word 4	Topics
Topic 0	food	order	place	come	service	[food, order, place, come, service]
Topic 1	time	service	say	work	make	[time, service, say, work, make]
Topic 2	place	love	pizza	make	time	[place, love, pizza, make, time]

Figure: Topics found analysing 50k reviews.

.

Results of Classification Models

Model	Number of class	Execuion Time (H:M)	Predictibility
LSTM	5	2:35	68.7%
	3	1:29	73.1%
	2	1:05	90.6%
CNN	5	0:19	55.5%
	3	0:12	69.01%
	2	0:7	91.5%
CNN +LSTM	5	2:29	35.9%
	3	1:06	73.4%
	2	1:31	84.38%
BERT	5	8:07	55.75%
	3	6:17	67.73%
	2	4:31	89.44%
FASTTEXT	5	0:19	52.10%
	3	0:14	66.00%
	2	0:09	84.5%

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Conclusion

In conclusion we can say that there is scope of improvement of the study:

- Topic modelling using LDA provide the result might change if we use more text or reviews. So, observing topics after adding more text could a future research, like how it will change or not ! Also different number of topics and words can also be explored.
- To predict 5 class LSTM is a good choice for this data set, for 3 class CNN+LSTM hybrid or LSTM and for 2 class CNN perform better.

Further Study

- For further research we can use several other data sets like Amazon product review data to observe the model performance and finally make a solid decision on appropriate classification model to classify review in different classes.
- Use of different word embedding techniques (One Hot Encoding, TF-IDF, Word2Vec) in observing the text classification performance by different algorithms.

Thank You

References I

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
J. Mach. Learn. Res., 3(null):993–1022.

Rehman, A. U., Malik, A. K., Raza, B., and Ali, W. (2019).
A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis.
Multimedia Tools and Applications, 78(18):26597–26613.