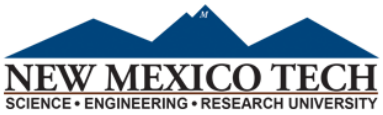



# Class Project: Topic modelling and sentiment prediction on customer review data

*Course Title*
*(CSE-489-05-Machine Learning)*

Md Ismail Hossain

May 10, 2022

# Contents

**1 Introduction** **3**

**2 Objective and Data** **4**

**3 Methodology** **5**

3.1 Text prepossessing . . . . . 5

3.2 LDA Topic Modelling . . . . . 6

3.3 Classification Models Architecture . . . . . 7

**4 Results** **9**

**5 Conclusion** **10**

# List of Figures


1 Language analytics few examples. . . . . 3
2 Topic modelling. . . . . 6
3 Topic modelling Architecture. . . . . 7
4 Architecture of LSTM model. . . . . 7
5 Architecture of CNN model. . . . . 8
6 Architecture of Hybrid CNN+LSTM model. . . . . 8
7 Architecture of BERT model. . . . . 8
8 Architecture of Fasttext model. . . . . 9
9 Topics found analysing 50k reviews. . . . . 9
10 Topics found analysing 50k reviews. . . . . 10

# 1 Introduction

Natural Language processing (NLP) is one of the most promising and fast growing branch of machine learning and in AI. Because now a days the interaction between machine and human increasing day by day, like we are using Siri, Google Assistant, Alex in our day to day works. Also there are lots of other aspects where NLP is vastly used like product recommendation, improving the service or products, extracting information from vast amount of text and so on. So, in this project we have used to use Topic modelling and customer review classification on Yelp data set and finally recommend usable model in production level after analysing the outputs of several classification models.

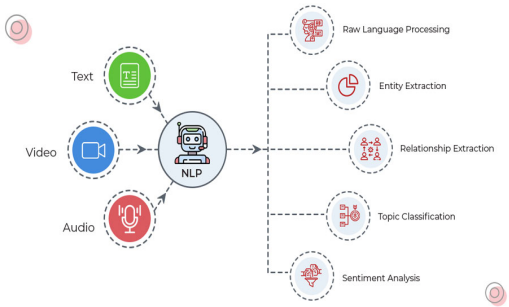



Figure 1: Language analytics few examples.

# 2 Objective and Data

1. Carry out Exploratory data analysis (EDA) and **Topic modelling** using LDA (latent dirichlet allocation) [1] 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) [3].

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

**Data analysis language and environment:** Python, Google Colab.

# 3 Methodology

## 3.1 Text prepossessing

The texts are converted into word level after doing several processing steps, like remove unnecessary characters from sentences, doing lemitization, tokenization and so on. Then we have to convert the text into numbers using following ways:

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

**Unique numbers and One hot encoding** .

Label Encoding

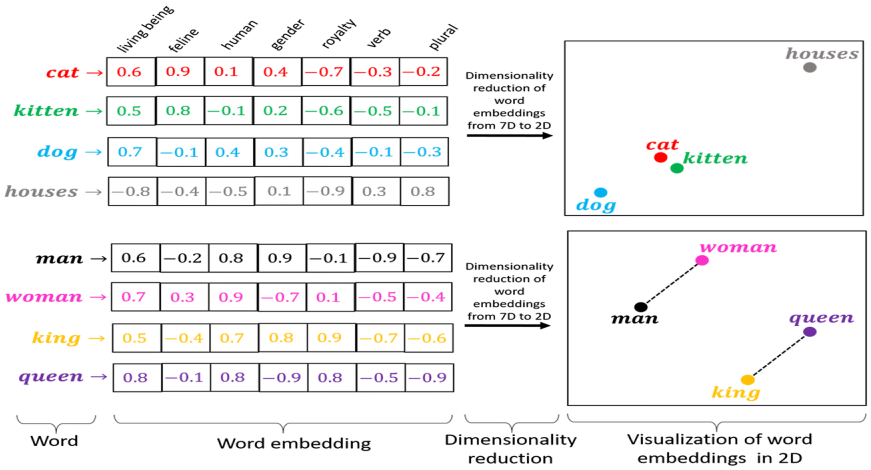
| Food Name | Categorical # | Calories |
|---|---|---|
| Apple | 1 | 95 |
| Chicken | 2 | 231 |
| Broccoli | 3 | 50 |

→

One Hot Encoding

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

**Word Embedding** ( TF-IDF, Word2Vec) .



## 3.2 LDA Topic Modelling

LDA is a probabilistic topic modelling method, used Gibbs sampling in distributing the topics over reviews. There are whole bunch of theory behind the algorithm. Interested people could go through this paper [2] for better understanding. In training the algorithm we need to feed the data and number of topic that we want to consider.

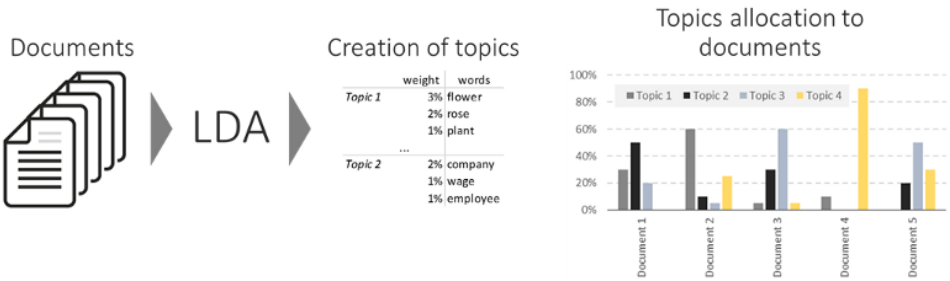



Figure 2: Topic modelling.

.

Key assumptions for Topic modelling:

- 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.

$$p(\beta,\theta,z,w)=\left(\prod_{i=1}^{K}p(\beta_i|\eta)\right)\left(\prod_{d=1}^{D}p(\theta_d|\alpha)\prod_{n=1}^{N}p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)$$

$$(\beta_d|\eta)\sim Dir(\beta) \qquad (\theta_d|\alpha)\sim Dir(\alpha) \qquad Z_{d,n}\sim Multi(\theta_d) \qquad W_{d,n}\sim Multi(\beta_{z_{d,n}})$$

$$p(z_{d,n}|\theta_d)=\theta_{d,z_{d,n}} \qquad p(w_{d,n}|z_{d,n},\beta_{1:K})=\beta_{z_{d,n},w_{d,n}}$$

.

.

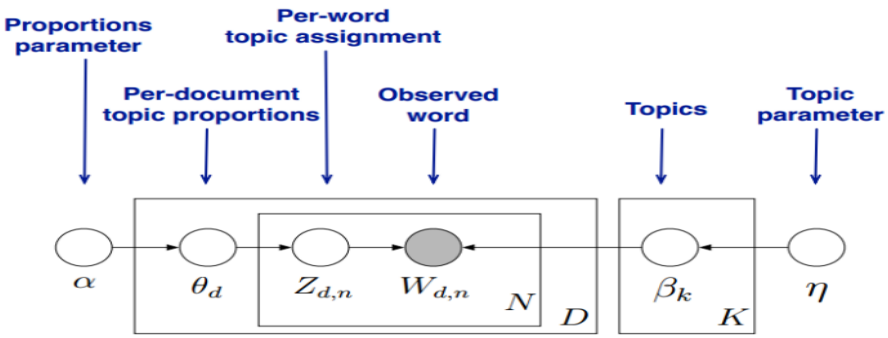



Figure 3: Topic modelling Architecture.

## 3.3 Classification Models Architecture

The architecture of the selected models are presented below:

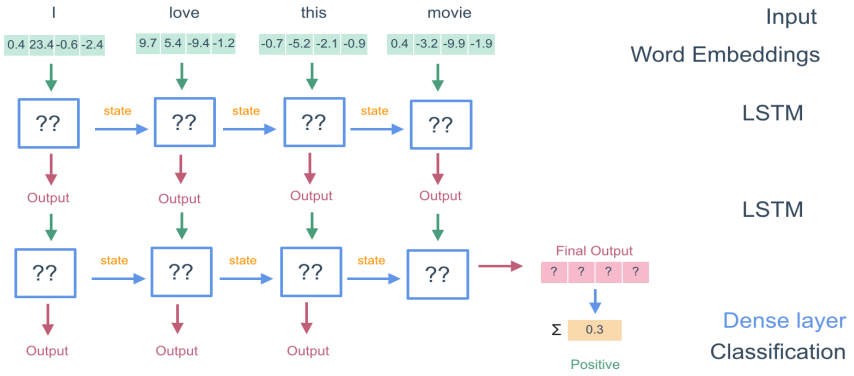



Figure 4: Architecture of LSTM model.

.
.
.
.
.

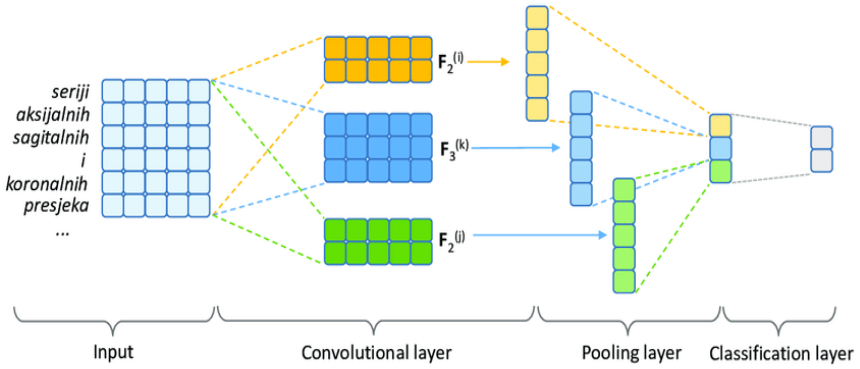



Figure 5: Architecture of CNN model.

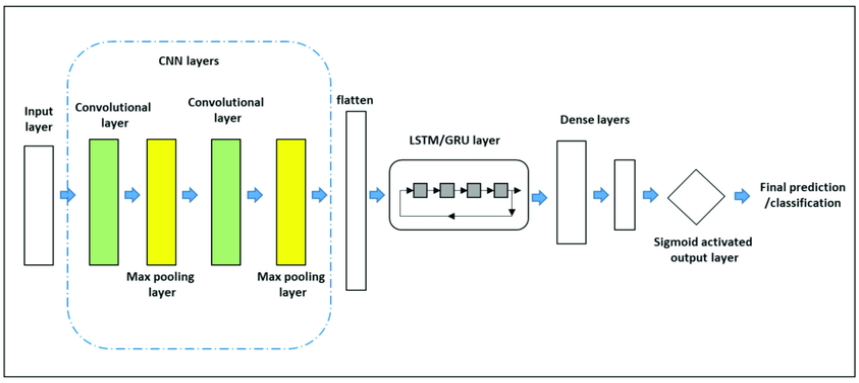



Figure 6: Architecture of Hybrid CNN+LSTM model.

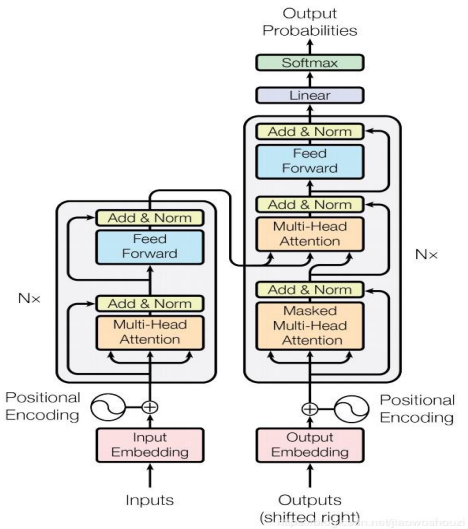



Figure 7: Architecture of BERT model.

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

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

$$\cdots\cdots$$

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

Figure 8: Architecture of Fasttext model.

# 4 Results

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 9: Topics found analysing 50k reviews.

.

The topics have to be verified and interpreted by human and we can try different number of words for the topics.

.

In predicting 5 category LSTM performs best in terms of time and accuracy, then for 3 category prediction the hybird of CNN+LSTM is best and finally for 2 category prediction CNN model achieve highest accuracy with minimal computational time.

| 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% |

Figure 10: Topics found analysing 50k reviews.

# 5 Conclusion

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. 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.

# References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003.

[2] William M. Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. 2011.

[3] Anwar Ur Rehman, Ahmad Kamran Malik, Basit Raza, and Waqar Ali. A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, 78(18):26597–26613, 2019.