# Menu Price Prediction using Neural Networks

**Michele Ceru**
Center for Data Science
New York University
mc3784@nyu.edu

**Luisa Quispe Ortiz**
Center for Data Science
New York University
lqo202@nyu.edu

## Abstract

Culture diversity and socioeconomical groups strenghten are marked by the way language is used in each of them. Many studies have shown that there is a relation between how a restaurant describes itself and price a restaurant charge (or the price range it has), based on the previous statement. Therefore, it may be possible to find use menu's description to predict the price of a dish. We aim to perform this task taking advantage of tools such as word embeddings and neural networks, that we believe will aide and facilitate the objective.

## 1 Introduction

New York is known as being a city that has influence of several cultures so it's logical that its cuisine reflects that too. Across all the city it is possible to find chinese, italian, mexican and many more kitchens of the world having a myriad of plates and prices.

Many factors influence the price, from external sources such as inflation, availability, critics reviews and even customer opinions (Jurafsky, 2014).

Each restaurant knows their target customer, hence all the business is focused on it including the menus. This implies a relation between socioeconomic class and language (Freedman and Jurafsky, 2011), which implies also Bordieu's distinction (Jurafsky et al., 2016). Therefore, there is evidence of a relation between how a restaurant describes its plates and how much it charges. Would it be possible to predict the price of a dish given its description?

There are some previous works that are related to the mentioned task, some of them from a economics, culinary approach or a linguistics point of view. The objective varies according to the approach taken. For example, in economics it is important to see the impacts of reviews on sells, while in culinary and hospitality the objective is to analyze language used in the menus to mainly make recommendations for writing (Chahuneau et al., 2012).

Although the variety, some handful insights were found such as in (Chahuneau et al., 2012; Jurafsky, 2014; Jurafsky et al., 2016): 5-stars restaurants tend to use more fancy words and borrow expressions from other languages making the description larger but succinct, while cheaper restaurants have a wider variety of food and tends to use just a small phrase to describe its plates and focus on adjectives and filler words.

None of the previous literature, however, used distributed word representations and neural networks to predict the menu's price per item using just text. Using word representations will aide to understand the underlying semantic relation between words (Mikolov et al., 2013a) in the culinary aspect. And training using neural networks will bring us the possibility of capturing non-linear relationships (Beale and Jackson, 1990) between the embeddings and the target variable price , even more if we consider the dynamics of word order as in recurrent neural networks (LSTM or GRU). We believe

that it's possible to confirm some conclusion other studies have had, besides reaching an improvement of the evaluation metric due to the use of the mentioned deep learning tools.

## 2 Prior Literature

As stated in (Chahuneau et al., 2012), most of the previous works are in the linguistics, hospitality research and economics fields. Each of them gives a slightly different approach, a few focusing in predicting price range of a restaurant and others in the analysis of the linguistic implications of the menu. However, it's good to have a starting point.

Socieconomic groups are defined by the way they use the language, this is widely known and used by politicians. In (Freedman and Jurafsky, 2011), the authors claim that the previous phrase is true, even more, they assert that food is a "robust marker of group identity" too. They show that prices affect the language in food advertising, specifically in potato chip bags. A similar study, this time experimental was done in (McCall and Lynn, 2008) but to see implications of text complexity on perceptions of quality and purchase intentions. The view was more psychological, and they proved that longer, more descriptive menu lines usually made the consumer think it was expensive.

There has been also works in prediction using text, such as the one described in (Archak et al., 2011) where they analyze the impact of reviews on product sales, a marketing and economics point of view. They focus on a couple of products and identify "beliefs" a customer has in a product's feature, not considering a polarized review (positive or negative based on the rating). This has not been the only prediction task using text that can be found, some examples are predicting risk in financial markets or using reviews to predict movies profitability (Joshi et al., 2010), to other several economic tasks.

In the field of culinary and hospitality research, there are many manuals that have recommendations for writing menu descriptions, according to (Chahuneau et al., 2012) and (Jurafsky, 2014), such as (Kasavana et al., 1990) . Actually is interesting to notice that by conducting an experiment in a small cafeteria (Wansink et al., 2005) showed that the menu description does affect the customer's behavior and perception.

The logic that menu's descriptions produce change considering the restaurant's market niche (public they are directed to), and the fact that this follows linguistic socieconomical differentiation, is mentioned and explained in the first chapter of (Jurafsky, 2014), where he cites a some works and their main implications in the relationship between prices and descriptions, besides giving his own perceptions. Jurafsky assures that expensive restaurants' menus tend to use "native" words borrowed from other languages such as french, italian, spanish; they use fancy words in long descriptions and they focus more in the detail of the description and avoid using filler words. On the other hand, common cheaper restaurants have a wider variety of plates, for, they try to sell them by describing in a simple and shorter way using a lot of filler words ("real cream cheese", "home-baked cookies").

Our main source is (Chahuneau et al., 2012), as one its goals is to predict the price and how it is influenced by the use of language. The article uses menu descriptions from main cities across the U.S. as well as some descriptive features extracted from Yelp reviews (flag variables of wi-fi, parking, delivery, etc). Their analysis is mainly linguistical, as they wish to know how each word or change of word affects the price, using regression tools.

In a more recent background, a study (Jurafsky et al., 2016) focuses on studying the reflections of Bourdieu's distinction, which happens to consider that there are deep associations linking food culture with social class and other aspects of identity, using the language of food. They split their study in 4 aspects, that define the relation between food and social culture aiming to predict the price range a restaurant belongs to. Using Yelp reviews they conclude that words used in menu descriptions reflect the many aspects of Bourdieu's distinction, therefore these help determine the price range. In general terms, the insights found are mainly the ones mentioned in the previous paragraphs with a couple of new findings (e.g. adjectives are used mostly in cheaper backgrounds), so the study is kind of confirmatory.

Finally, a more studied related field involves sentiment analysis using reviews, there are plenty of state-of-the-art literature. Some use a more tradi-

tional approach such as in (Jurafsky et al., 2014), while others start using deep learning to get better results (Tang et al., 2015).

## 3 Data

To gather the information required by the project, we considered previous works and decided to crawl AllMenus.com (`www.allmenus.com`). Initially we only got New York City menus, but to ensure consistency in our results by having a larger sample, we decided to include San Francisco. We ended up with 467,669 NY menu descriptions and 200,020 from SF. For us, each of these rows is a observational instance corresponding to a menu item. So, the features captured were: restaurant name, section of the menu, subsection of the menu, dish name, dish price and description, from which we will take the price as our target and the dish name and description as our input text.

Before moving on the modeling task, we made a quick analysis of our information in order to detect any anomaly. The first we could find is that only 99 observations had null price which we set to zero price (and later erased), the rest was encoded in different ways but it was possible to obtain the price value of the plate.

Then, we wanted to see the distributions per city, and whether there was significant difference that pointed us to do separately analysis. The mean price of SF is of $9.77, one more dollar than the mean of NYC, a similar relation was found in the medians as NYC's was of $6.99 against $7.99 in SF. Despite this, by looking at the normalized (standardized) distributions in Figure 1 it was possible to see that both seem alike, highly skewed.

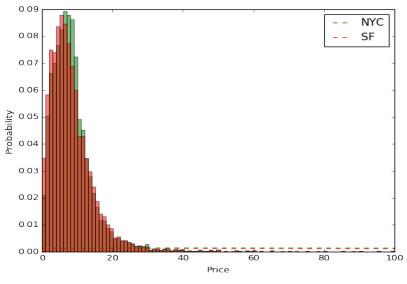


**Figure 1:** Distribution of the prices in NYC and SF

This gave us evidence to keep on analyzing all the dataset together.

In the text, we made a tokenizer that deleted some punctuation signs and that lower case all the input text.

For all our runs, the whole data set of restaurants item's price was split into training(80%), validation (10%) and testing(10%) set randomly, considering that the proportion of observations per city will be the same due to randomness.

## 4 Modeling Framework

### 4.1 Some definitions

Before explaining our protocol, we give some brief definitions about the tools that will be used.

**Word Embedding** Distributed representation of words are widely used nowadays, they are just ubiquitous to most NLP tasks, as they provide semantic relation between words (Hill et al., 2016). A good example is the known neural embedding **Word2Vec** result of an Google's experiment which trains words against other words that neighbor them in the input corpus. Their approach based on two techniques, the first predicts the word that is in the middle based on the 4 surrounding neighbors which is known as Continuous Bag of Words. The second model is actually the reverse process, to predict the words around a specific one known, called Skip-gram (Mikolov et al., 2013b; Mikolov et al., 2013a).

For some experiments, we'll use the pretrained Word2Vec inputs, as they are available online (`https://deeplearning4j.org/word2vec`). There are 3 million vectorized words which have a dimension of 300.

**Neural Networks** Neural networks (deep learning) have started to be used in NLP for a while now. Most of them are used in tasks such as sentiment analysis, POS tagging, and machine translation tasks (Cho, 2015).

One of the most used neural networks is the recurrent neural networks Long Short Term Memory and Gated Recurrent Unit (Cho, 2015) as they don't suffer from vanishing or exploding gradient problems. A huge advantage of GRU and LSTM is that they capture the sentence sequence dynamics, therefore it is possible to get more insights (semantically and

syntactically) of the whole text (Gers et al., 2000; Cho et al., 2014; Cho, 2015).

There's some literature that suggest working schemes for continuous targets such as in (Specht, 1991) which uses a statistical framework. Based on material seen in class and on (Cho, 2015) we find that by changing the loss function, it is possible to focus in a continuous target. We'll use both, categorical an continuous targets in the neural networks to predict price of a restaurant menu's item.

### 4.2 Protocols Proposed

As it has been stated, the proposed work just considers text as input (item's descriptions and names) to predict a target which is the price. We planned to launch some experiments considering the embedding type, the type of neural network and the type of target.

**Table 1:** Experiment setup

| Target type | Embedding Type | Neural Nework |
|---|---|---|
| Class | Self-learned | MLP |
| Class | Self-learned | LSTM |
| Class | W2V fixed | MLP |
| Class | W2V fixed | LSTM |
| Class | W2V initial | MLP |
| Class | W2V initial | LSTM |
| Continuous | Self-learned | MLP |
| Continuous | Self-learned | LSTM |
| Continuous | W2V fixed | MLP |
| Continuous | W2V fixed | LSTM |
| Continuous | W2V initial | MLP |
| Continuous | W2V initial | LSTM |

Regarding the embedding, it was possible to set the initial embedding to random numbers, so that it would be another parameter to estimate. Another option was to make the embedding not trainable so it would be a fixed pre-trained matrix. A third option consisted in a mix of the previous: instead of starting from a random point, the embedding matrix could be trained starting from a pre-trained matrix and learned through the optimization process. We decided to used pre-trained Word2Vec embeddings(W2V) in the cases pre-trained matrices were needed.

Another option in modeling was to take into account the order of the words, which is usually captured by a recurrent neural network: a LSTM or GRU. Since we wanted to know if order mattered, we decided to launch a Multilayer Perceptron (MLP) and a LSTM based model for the experiments.

Finally, we had to decided how was the price target going to be predicted. As most of the literature focuses in range predictions, we choose that path first: priced were group in 10 labels cuts were $5 dollars. This selection was made in order to try capture the tail of the distribution, given that if percentiles were chosen steps between groups may be too short.

To give another point of view and to made our experiments comparable, we decided to use a continuous target by setting the loss function to a squared loss.

Our final setup can be seen in Table 1.

### 4.3 Model Configuration

All the models are believed to have the default configuration showed in Table 2, otherwise the change will be explicitly mentioned.

**Table 2:** Default Configuration

| Configuration Variable | Value |
|---|---|
| BATCH_SIZE | 64 |
| CHECKPOINT_EVERY | 200 |
| DROPOUT_KEEP_PROB | 1 |
| EMBEDDING_DIM | 300 |
| EVALUATE_EVERY | 50 |
| INIT_SCALE | 0.1 |
| L2_REG_LAMBDA | 0 |
| MAX_GRAD_NORM | 5 |
| NUM_EPOCHS | 10 |

The optimizer used depended on the model, for MLP Adam Optimizer was used, with a learning rate of 0.003. For the RNN LSTM a Gradient Descent was used using an initial learning rate of 1, and then a decay of 0.5 if no improvement was found. Some experiments though considered for all the neural networks an Adam Optimizer, again if so it will be said explicitly.

Besides all the mentioned details, specifically for the LSTM cell, we considered a mean of all the output states ($h_t$) similar to what is done in .

### 4.4 Training and Model Selection

We trained some models considering the 10 epochs, but most of them were trained with a 4 step early stopping, similar to what's suggested in (Cho, 2015).

Our metric to evaluate the early stop was based on the loss function. If in 4 evaluation periods the loss was increasing it meant that the model reached a local optima.

All of our runs were done in a HPC server requesting around 32GB of RAM and 1 GPU unit, and depending on the experiment it could take from an hour to 36 hours to finish.

### 4.5 Evaluation

Depending of the whether the model was continuous or class it had different loss functions. Therefore, for the class we used accuracy as the main indicator to see how well the model was performing. The loss in this case was set to use a cross entropy (having behind it a 10 class categorical distribution). For the continuous model our performance metric was the mean squared error and the mean absolute error, using a mean squared L2-norm as stated in (Cho, 2015).

Furthermore, in order to have more insights on what our model was driven by, we decided to look at the resulting embeddings and see how words affect the price prediction. For the primer, using t-SNE (PCA component) to reduce dimensionality to 2D helped us to plot the embedding. Also, for the latter, we adapted a script to predict interactively prices, this will help us to understand the influence of words. The objective here was to see if our models capture the relations explained in all the state-of-the art literature mentioned in part 2: expensive restaurants characterize for using foreign words, making long descriptions but not using filler words nor adjective while cheaper restaurants tend to focus more in describe provenance and use a lot of filler words (”real”, ”over”,etc).

## 5 Results

Results running all the epochs are displayed in . They show that there is a gain in running the 10 epochs, but it may not be worth to do it as the results do not increase drastically.

On the other hand, we had a bunch of models to select from each of protocol we decided to test since randomnes of the batches affected of all them. The best among all our results per category are shown in Table 3.

As part of the evaluation, in order to understand more about the model, we chose top 100 words in frequency of a learned embedding. The chosen model was , whose project embedding is 2D is displayed in .

Finally, we evaluated a couple of models with our interactive prediction section. We tested some phrases found on (Jurafsky, 2014) to see how our models reacted. The results of this evaluation are displayed in table

## 6 Analysis

## 7 Conclusions

### Limitations

**Future Work** We used word embedding as word representation to capture the syntactic relation between words. However, it is possible to use sentence embeddings to see if these capture more and therefore brings a better model ()


## Acknowledgments

We’d like to thank Prof. Bowman for allowing us to the explore relationship between text in food descriptions and prices using neural networks, and for his advice throughout the project.


## Collaboration Statement

## References


[Archak et al.2011] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.

[Beale and Jackson1990] Russell Beale and Tom Jackson. 1990. *Neural Computing-an introduction*. CRC Press.

[Chahuneau et al.2012] Victor Chahuneau, Kevin Gimpel, Bryan R Routledge, Lily Scherlis, and Noah A Smith. 2012. Word salad: Relating food prices and descriptions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1357–1367. Association for Computational Linguistics.

[Cho et al.2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

**Table 3:** Best results by protocol

| Target type | Embedding Type | Neural Network | Loss Type | Training Loss | Dev. Loss | Test Loss | Indicator |
|---|---|---|---|---|---|---|---|
| Class | Self-learned | MLP | Cross-entropy | | | | |
| Class | Self-learned | LSTM | Cross-entropy | | | | |
| Class | W2V fixed | MLP | Cross-entropy | | | | |
| Class | W2V fixed | LSTM | Cross-entropy | | | | |
| Class | W2V initial | MLP | Cross-entropy | | | | |
| Class | W2V initial | LSTM | Cross-entropy | | | | |
| Continuous | Self-learned | MLP | Squared | | | | |
| Continuous | Self-learned | LSTM | Squared | | | | |
| Continuous | W2V fixed | MLP | Squared | | | | |
| Continuous | W2V fixed | LSTM | Squared | | | | |
| Continuous | W2V initial | MLP | Squared | | | | |
| Continuous | W2V initial | LSTM | Squared | | | | |


[Cho2015] Kyunghyun Cho. 2015. Natural language understanding with distributed representation. *arXiv preprint arXiv:1511.07916*.

[Freedman and Jurafsky2011] Joshua Freedman and Dan Jurafsky. 2011. Authenticity in america: Class distinctions in potato chip advertising. *Gastronomica: The Journal of Critical Food Studies*, 11(4):46–54.

[Gers et al.2000] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

[Hill et al.2016] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

[Joshi et al.2010] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics.

[Jurafsky et al.2014] Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

[Jurafsky et al.2016] Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2016. Linguistic markers of status in food culture: Bourdieus distinction in a menu corpus. *Cultural Analysis*.

[Jurafsky2014] Dan Jurafsky. 2014. *The language of food: A linguist reads the menu*. WW Norton & Company.

[Kasavana et al.1990] Michael L Kasavana, Donald I Smith, and Raymond S Schmidgall. 1990. *Menu engineering: a practical guide to menu analysis. Rev*.

[McCall and Lynn2008] Michael McCall and Ann Lynn. 2008. The effects of restaurant menu item descriptions on perceptions of quality, price, and purchase intention. *Journal of Foodservice Business Research*, 11(4):439–445.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Specht1991] Donald F Specht. 1991. A general regression neural network. *IEEE transactions on neural networks*, 2(6):568–576.

[Tang et al.2015] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432.

[Wansink et al.2005] Brian Wansink, Koert Van Ittersum, and James E Painter. 2005. How descriptive food names bias sensory perceptions in restaurants. *Food quality and preference*, 16(5):393–400.