Text Classification and Topic Modeling over Yelp reviews

Bui Xuan Julia Lan[1], Pasinato Alessio[1]

Abstract
For this project a dataset of 6,990,280 reviews and 150,346 businesses was used. It was obtained from Yelp, an online platform for sharing experiences and opinions about local businesses. The project employs topic modeling and classification techniques to extract the main topics and contents of the reviews, as well as classify them into different categories. Preprocessing steps such as normalization, stopwords removal, tokenization and lemmatization were performed on the text. Latent Dirichlet Allocation (LDA) was used for topic modeling, and two different text representations (TF-IDF and Doc2Vec) were tested for the classification task. The results showed that topic modeling identified 8 food-related topics, and the classification model achieved a precision of 93% and a recall of 74% on a multilabel multiclass problem with 44 different classes.

Keywords
Natural Language Processing — Topic Modeling — Text Classification — Latent Dirichlet Allocation

[1] *Department of Informatics, System and Communication , University of Milan Bicocca, Italy*

Contents

1. Introduction

Yelp is an online platform that is used by individuals to share their experiences and opinions about local businesses. The platform allows users to search for and review businesses in a variety of categories, such as restaurants, hotels, and retail stores. Businesses can also create and manage their own profiles on the platform, which can include information about their products or services, as well as photos and contact details. Reviews are also provided by the users, which can be filtered by various criteria such as rating, date, and location. Additionally, Yelp also provides a range of tools for businesses to help them manage their online presence and track the performance of their listing. The platform is widely used and is considered a valuable resource for both consumers and businesses alike. A dataset consisting of 6,990,280 reviews and 150,346 businesses information has been released by the platform, which has been utilized for this text mining project. For this project, topic modeling is employed for the extraction of the main topics and contents of the reviews. Classification techniques instead, are utilized to classify the texts; it is a multi-label multi-class task since an instance can belong to multiple classes and respectively mutliple labels.

2. Objectives

The research questions for this project are:

- development of a model capable of labelling reviews
- test different text representations and compare their performances over the classification task
- perform topic modelling to detect the most discussed topics in the reviews

3. Dataset sampling

Since much computational power is needed to process six million texts, it was decided to sample the dataset and to keep only the reviews of businesses who had been reviewed at least one thousand times; this choice was also made in order to have, for each class, enough data for the classifier to be trained. The result consisted in a dataset cointaining 570,438 reviews. Since the ”Restaurants” label was present in 94% of the reviews, hence it was very an unbalanced class, it was removed.

4. Exploratory Data Analysis

The sampled dataset contains 570,438 reviews and 200 different classes, with "Restaurants" being the most frequent one, occurring 538,450 times, and "Hot Dogs" being the least frequent one, occurring 1,045 times. In the graph below [Fig. 1] the top 50 classes are considered: the frequency distribution of reviews is displayed.

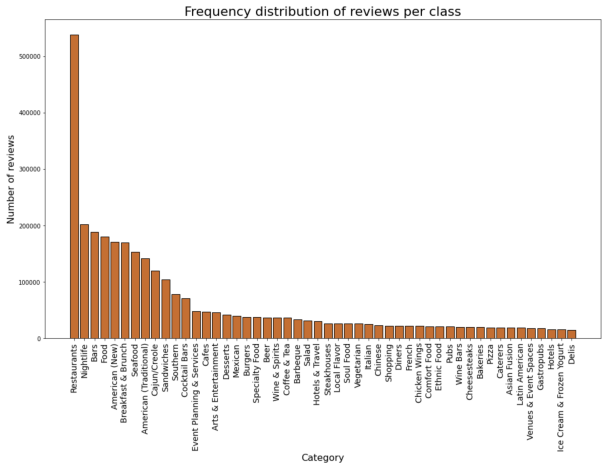

Figure 1. Top 50 - Frequency distribution of reviews per class

As shown, the class "Restaurant" is highly unbalanced with respect to the other ones; it can be considered as a very generic category to which the majority of the businesses can be assigned to, hence, it was decided to be removed.

5. Text Pre-processing

Text pre-processing refers to the cleaning and preparation of raw text data for further analysis of natural language processing. Its steps and their orders depend on the specific task.
For the purpose of this project, the first step of pre-proccesing consisted in lower case folding; it is the process of converting all the characters in a document into lower case in order to normalize words. Furthermore, all the occurrences of newlines, tabs, emojis, extra white-spaces, numbers and URls were removed. More precisely, specific methods were applied in order to remove URLs even if written in different forms (i.e. those written with the entire URL and those written with only ".com" at the end). Additionally, SpaCy library [1] was used in order to implement tokenization, with the aim of splitting the text document into semantically meaningful units, called "tokens". Stop words and punctuation were also removed, followed by the application of lemmatization to the resulting tokens. Lemmatization was chosen over stemming, since the latter one just removes (or "stems") the last few characters of a word, often leading to incorrect meanings and spelling. Lemmatization, instead, considers the context and converts the word to its base form, which is called "lemma". Lemmas were created with both SpaCy and WordNet libraries. Finally, in order to improve the pre-processing, repeated characters removal (reducing repetitions to two characters for alphabets) and spelling correction were applied. However, spelling correction used with the "autocorrect" library [2], seemed not to perform well (i.e correcting "i'm glade" into "i'm grade", instead of "i'm glad"). Therefore, it was decided not to consider this last step in the final definition of the dataset.

6. Classification

Due to the large amount of data present even in the reduced dataset that would have implied days for the models to be trained, another sampling procedure took place. Specifically, the reviews related to 15 randomly selected businesses were taken into account, resulting in a dataset containing 27512 reviews and 44 different classes. The most frequent category was "Bars" with 14383 occurrences, while the least frequent category was "Donuts" with 1074 reviews. To encode the labels, the MultiLabelBinarizer module from the scikit-learn library [3] was utilized. Subsequently, a stratified sampling technique was employed to divide the dataset into training and test data, comprising respectively 70% and 30% of the dataset.

6.1 TF-IDF

In the first classification phase it was decided to adopt tf-idf representation for representing texts, and TfidfVectorizer [4] was used for this task. Only words that appeared in more than 5 documents were initially taken into account and subsequently only the most 500 frequent ones were kept. The multiclass multilabel task was addressed as 44 independent binary classification problems and OneVsRestClassifier from skicit-learn [5] was implemented. As regards to the testing phase, for each text input, the classifier provided the probability of the text belonging to each class; after having tested different probability thresholds, such as 65%, 70%, 75% and 80%, it was decided to use 65% as a threshold because it provided better predictions. The performance of both WordNet and SpaCy lemmatizers were compared by testing texts processed with each method.

	Precision	Recall	f1-score
micro avg	0.93	0.75	0.83
macro avg	0.94	0.73	0.82

Table 1. Scores WordNet lemmatizer

	Precision	Recall	f1-score
micro avg	0.93	0.75	0.83
macro avg	0.95	0.75	0.83

Table 2. Scores SpaCy lemmatizer

Table 1 and table 2 display the precision, recall and f1-score for the two lemmatization methods. The micro average and macro average scores for both tables are very similar, with a slight difference in the macro average for precision in the

SpaCy's table, indicating that there is a negligible difference in performance between the two lemmatization methods.

6.2 Doc2Vec

To test the TF-IDF representation's performance, it was decided to compare it with the Doc2Vec one. To train the model it was decided to remove words that appear in less than 5 documents and the vector size was set to 50; then, the model was trained for 100 epochs.

	Precision	Recall	f1-score
micro avg	0.83	0.55	0.66
macro avg	0.81	0.53	0.62

Table 3. Scores Doc2Vec

As shown in table 3, Doc2Vec performances are significantly lower than the ones obtained by the previous methods; to try to improve them, it was decided to increase the vector size to 100. However, due to the significant increase in computational cost, it was determined that this would result in an excessively prolonged training phase for the classifier and the modification was not implemented.

7. Topic Modeling

Topic modeling is a method used in natural language processing to identify the underlying themes or topics present in a collection of text documents. One of the common algorithms for topic modelling is Latent Dirichlet Allocation (LDA).

7.1 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model, assuming that each document is a mixture of a small number of topics, and that each word in the document is associated with one of the topics. It is based on Bayesian inference to estimate the latent topics and their corresponding mixture weights. The output is a list of topics with associated clusters of words (and their probabilities) [6].
LDA treats documents as bags of words; therefore term frequency (TF) representation, which measures the importance of a specific term in a document by counting the number its occurrences, was used for this topic modelling task.

However, the LDA model does not provide the optimal number of topics for the text itself. Hence, it needs to be determined by the user.
Finally, the model is evaluated in terms of:

- Perplexity metric: it's a measure of uncertainty, indicating the degree of surprise of a model when presented with new data. A lower perplexity score indicates that the model is more certain of the data and therefore, it is considered to be a better performing model.
- Coherence metric: measuring the degree of semantic similarity between high scoring words in the topic. More precisely, "CV" measure is used, whose values lay between 0 and 1 range [7].
- Human evaluation: observation-based (eg. observing the top 'n' words in a topic) and interpretation-based (eg. 'word intrusion' and 'topic intrusion' to identify the words or topics that "don't belong" in a topic or document).

7.2 Results

Initially, a number of k = 20 topics was randomly selected, reaching a value of perplexity equal to -7.38 and a value of coherence equal to 0.391. Among the 20 topics considered, few of them presented coherent characterizing words. Results were found to be inadequate, since many words seemed not to belong to the topic assigned. For some of the clusters, it was possible to extract topics dealing with "excellent restaurant with good service", "seafood at the beach", "breakfast" and "bars for celebrations". In order to improve the interpretation of the results, the optimal number of topics was found, based on coherence measure, as shown in Figure 2.

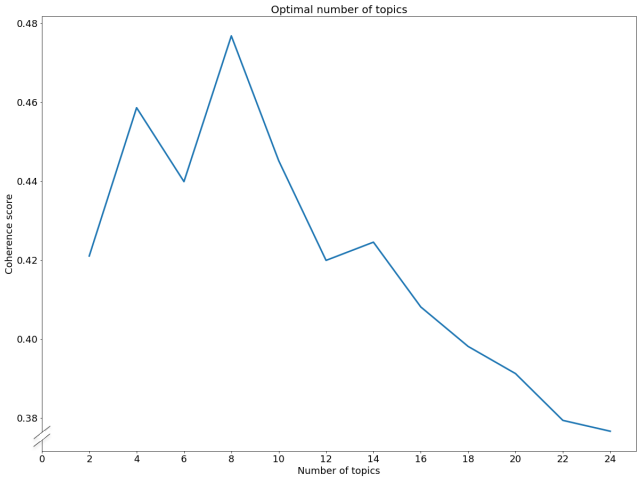

Figure 2. Optimal number of topics. Truncated Y axis. Coherence scores against number of topics

The highest value of coherence equals to 0.477 and it is given by the number of topics k = 8. Therefore, LDA was implemented again, with k = 8. The metric of perplexity reached a value of -7.34, similar to the one obtained with k = 20. It is important to notice that, using a lower number of topics, allowed to gain more understandable interpretations of the results.

The top words can be considered representative of the considered topics and there are no words that seem not to belong to the topic assigned. The topics extracted can be labelled in the following way:

1. Seafood
2. Hotel
3. Dessert
4. Breakfast / Brunch
5. Bar
6. Sandwiches

7. Positive food experiences
8. Orders / Waiting time

The wordclouds in Figure 3 represent the most characterizing words of the extracted topics.

Figure 3. LDA with k = 8. Wordclouds per each topic.

8. Conclusions

The goals of the project consisted in creating a model capable of classifying reviews and perform topic modelling, to understand the main discussed arguments of the documents. In order to do that, it was necessary to perform preprocessing steps such as lowercasing, stopwords removal, tokenization and lemmatization. To infer topics, LDA was performed on the preprocessed text; for the classification task, instead, two different representation of the documents were tried, TF-IDF and Doc2Vec, and the corresponding classifiers were trained and compared. Topic modelling allowed the characterization of 8 food-related topics. On the other hand, classification allowed the creation of a model with metrics reaching a value of 93% for precision and 74% for recall (computed as the average value between micro and macro measures).

References

[1] Matthew Honnibal and Ines Montani. spacy - industrial-strength natural language processing. https://spacy.io/.

[2] Filip Sondej Jonas McCallum. Autocorrect - spelling corrector. https://pypi.org/project/autocorrect/.

[3] sklearn. Multilabelbinarizer. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html.

[4] skicit learn. Tfidfvectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[5] skicit learn. Onevsrestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html.

[6] Wang Y. Yuan C. et al. Jelodar, H. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimed Tools Appl 78, 15169–15211 (2019).*

[7] Topic coherence. https://radimrehurek.com/gensim/models/coherencemodel.html.