Text Mining and Search Project, 2023/24 academic year

The Yelp Dataset: Topic Modelling and Topic Classification

Lorenzo Lecce n. matr. 830881 **Andrea Lucini Paioni n. matr. 826578**

ABSTRACT

This project takes inspiration from a part of the Yelp dataset, a dataset which includes 6,990,280 reviews from 150,346 businesses, sourced from the online platform Yelp, famous for sharing opinions on local businesses. The study employed topic modeling and topic classification techniques to unveil the primary themes and content within the reviews, categorizing them into distinct groups. Initial steps involved text preprocessing steps, including normalization, stopwords removal, tokenization, lemmatization... Latent Dirichlet Allocation (LDA) was chosen for topic modeling, while two text representations (TF-IDF and Doc2Vec) were evaluated for the topic classification task. The findings revealed the identification of 10 food-related topics through topic modeling. In terms of classification, few model were developed, with the best one achieving an accuracy of 87%, a recall score of 96% and an F1 score of 90%, in addressing a classification problem with the aim of predicting the score of the reviews based on the text content of the reviews.

CONTENTS

INTRODUCTION

Yelp serves as an online platform where individuals share their perspectives on local businesses, offering a space to review various establishments like restaurants, local businesses, and hotels. Users can explore and evaluate businesses across different categories. Moreover, businesses have the opportunity to create and manage their profiles on Yelp, showcasing information about their offerings, along with photos and contact details. The platform facilitates user reviews, which can be filtered based on criteria such as rating, date, and location.

Yelp extends tools to businesses, aiding them in managing their online presence and monitoring the performance of their listings. Widely embraced, Yelp is considered a valuable resource for both consumers and businesses. The complex and huge dataset provided by the platform is composed by different datasets:

- “*yelp_academic_dataset_review*”: a dataset containing 6,990,280 reviews and information on the businesses on the platform.

- "*yelp_academic_dataset_business*": a dataset containing info about the businesses on the platform.
- "*yelp_academic_dataset_user*": a dataset containing info about the users who write reviews on the platform.
- "*yelp_academic_dataset_tip*": a dataset containing info about some particular reviews about the businesses on the platform.
- "*yelp_academic_dataset_checkin*": a dataset containing info about the time of the reviews about the businesses on the platform.

In this project, the focus is on employing topic modeling to extract the primary themes and content from reviews. Additionally, classification techniques are applied to categorize texts of the reviews. With these tasks in mind, we thought that only the "*yelp_academic_dataset_review*" and the "*yelp_academic_dataset_business*" dataset were needed for our purposes.

OBJECTIVES

The research questions for this project are:

- Evaluate the performances of different text representations considering the classification task.
- perform topic modelling techniques, to find some of the most discussed topics in the Yelp reviews.
- predicting the review stars considering the text of the reviews.

DATA PREPARATION

The datasets considered ("*yelp_academic_dataset_review*" and "*yelp_academic_dataset_business*") were huge, so we had to consider a sample, and we reduced also the number of variables, because some of them weren't useful for our purposes.

In fact, the reviews dataset had 6990820 instances (with 9 variables), and the businesses dataset had 150345 observations (with 14 variables). We decided to consider only 7 variables from the businesses dataset, and then we merged the two datasets, bringing the total to 15 variables:

- *review_id:* to univocally identify the id;

- *user_id*: to univocally identify the user who wrote the review;

- *business_id*: to univocally identify the business;

- *review_stars*: the stars assigned to the review by the user;

- *useful*: how many users found the review useful;

- *funny*: how many users found the review funny;

- *cool*: how many users found the review cool;

- *text*: the full text of the review;

- *date*: date and time of the review;

- *name*: the name of the business;

- *city*: the city of the business;

- *stars*: the mean stars of the business reviews, in categories between 1 and 5 stars (every 0.5 stars);

- *review_count*: how many reviews the business have;

- *attributes*: multiple attributes associated with the business, in a list format;

- *categories*: a list of tags associated with the business;

However, we still had the problem of computational power needed to process a dataset made of six million instances; in order to solve that, we decided to sample the dataset, keeping only the reviews of businesses with the category "*Italian Restaurant*", bringing the total down to 439358 instances.

TEXT PREPROCESSING

The text processing phase was crucial in our project, because it ensured the following analysis to be more accurate and significant. In particular, all the following phases were performed on the variable *text,* which contains all the text from the customers reviews.

In the preprocessing phase of our dataset, several essential steps were undertaken to ensure its readiness for analysis. Initially, all text within the reviews was converted to lowercase to maintain uniformity. Following that, numerical values were removed, ensuring that the dataset focused solely on textual content.

Tabs, empty lines (e.g., \\n), and links or URLs embedded in the reviews were systematically eliminated to enhance the cleanliness of the text. Furthermore, white spaces, emojis, and repeated characters were systematically stripped away to refine the data. The removal of punctuation, tokenization, and subsequent elimination of stopwords contributed to the streamlining of the text for subsequent analyses.

Finally, lemmatization was performed, aiming to reduce words to their base or root form, thereby facilitating a more cohesive and standardized representation of the textual information. These meticulous preprocessing steps collectively played a crucial role in preparing the dataset for robust and meaningful text mining endeavors.

EXPLORATORY ANALYSIS

In our dataset analysis, we conducted three distinct examinations to capture comprehensive insights (all the following analysis were conducted on a copy of the business dataset only, filtered to consider only Italian restaurants). The initial analysis involved generating a bar plot to illustrate the distribution of the stars from the ratings.

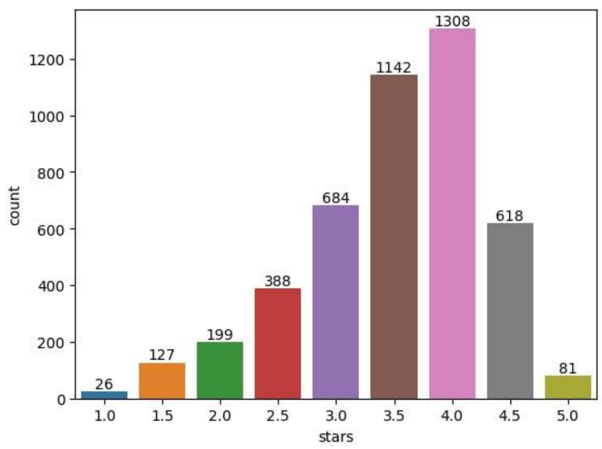

Figure 1: barplot variable "stars"

We noticed that the ratings of 3.5 and 4 were the most prevalent categories, meaning there was a general concentration of positive sentiments within the reviews.

The second analysis was focused on the review count across different categories, with a plot providing information about the number of reviews for each business activity considered in our dataset.

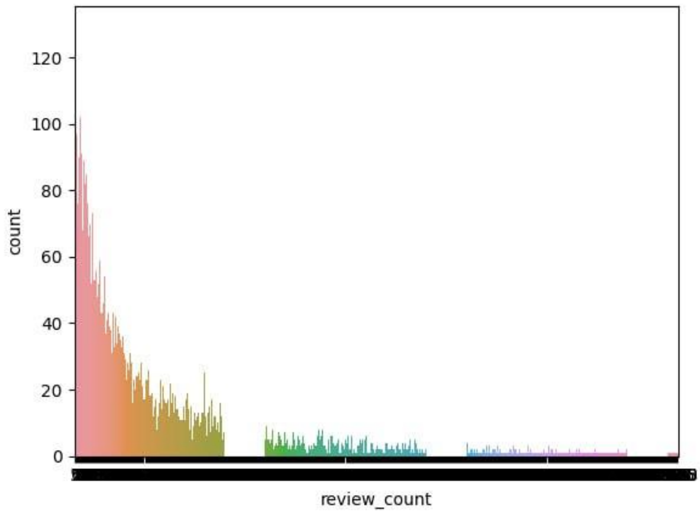

Figure 2: barplot variable "review_count"

Lastly, we developed a boxplot to present the distribution of the variable "stars" (the same of Figure 1) within the dataset. Together, these analyses helped us with interesting insights about significant variables from the dataset, giving useful info on both the sentiment distribution and review frequency patterns.

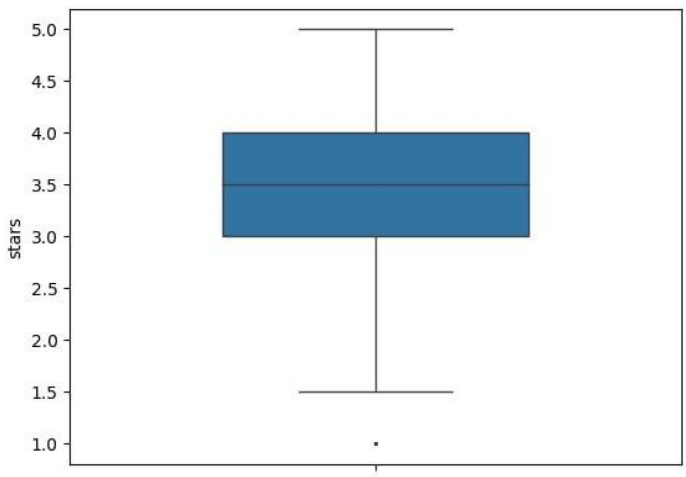

Figure 3: boxplot variable "stars"

To ensure we had more precise information about both the variable “*stars*” and the variable “*review_counts*”, we also developed some statistics about the distribution of the two variables:

var	*mean*	*min*	*max*	*std*	*25%*	*50%*	*75%*
Stars	3.51	1.0	5.0	0.78	3.0	3.5	4.0
Review count	92.39	5	4250	156.77	18	44	108

Table 1: statistics about the variables "stars" and "review_count"

We confirmed the information we had obtained with the plots: the distribution of “*stars*” is asymmetric towards higher values (all the values between 25% and 75% are between 3.0 and 4.0), with a mean of 3.51; and for the “*review_count*” variable we have the opposite situation, with a mean of 92.39 and a distribution towards lower values (75% of the business have less than 108 reviews).

TOPIC MODELING

To ensure the correctness of the task we decided to develop, we decided to further sample the dataset: some task on both topic modeling and topic classification required a lot of time and computational power.

So, we proceeded with another sample of the dataset, considering only the Italian restaurants situated in the city of Indianapolis, providing us with a dataset of 14347 observations and 17 variables (from the original 15, we added “*final_text*”, in the form of a list of token, and “*final_text_string*”, in the form of a string obtained from the list of token from the variable “*final_text*”; both the variables were obtained as a result of the text preprocessing phase).

The goal of the topic modeling phase was to identify a series of topics included in the text from the reviews of the businesses.

We first developed a Wordcloud, to have a general idea of the most common words inside the reviews.

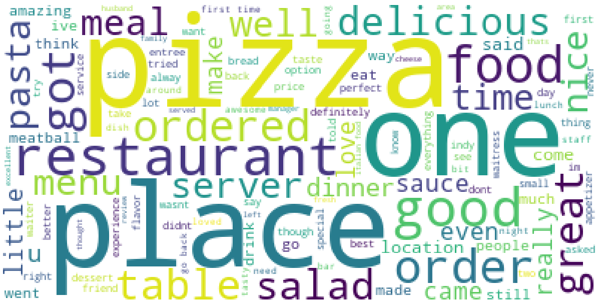

Figure 4: first general Wordcloud

As we can see, some of the most common words used are expected, and related in particular with Italian food: “pizza”, “pasta”, and some comments probably referred to the quality of Italian food, such as “good”, “great”, “well”, “nice”, “delicious”…

We then developed an LDA - or Latent Dirichlet Allocation - model to estimate the topics lying under the text. It is a useful topic modeling method that aim to find topics a document belong to.

It uses a Bayesian inference system to estimate the most common topics in the text collection, with also a weight associated with each topic to show their importance.

The model makes the assumption that each document can be resumed as a bag of words, in a way that only the number of occurrences of each word matters, and the position and the grammar role of each word are not considered in the model (for this reason, we proceeded with the text preprocessing phase, deleting all the stopwords, probably the most common words of the model otherwise, to ensure that only meaningful words would be considered in topic modeling).

An important aspect of the LDA model is the fact that we need to choose an appropriate number of topics: we decided to consider 10 topics. We obtained a model with 10 different topics.

Given the fact that the categories of restaurants were very similar, it’s understandable that the topics were very similar with each other: in general, “pizza” and “good food” were the most prominent topics found, with “great time”, “come back” and “good service” as other common topics.

The 10 topics described can be seen in the Wordcloud reported below:

Figure 5: second Wordcloud, with 10 topics after Topic Modeling

TOPIC CLASSIFICATION

The topic classification task is a mix of supervised methods aimed at classifying a characteristic of the text data; it can be both a multi-class classification and a binary outcome one.

As for the topic modeling, the same sampled dataset was used for topic classification, only considering reviews of Italian restaurants from the city of Indianapolis (a dataset made of 14347 observations and 17 variables).

For the classification purposes, we considered the variable “*stars*”, that indicates the number of stars, between 1 and 5, given by the customer to each business in association with the text review. We created another variable, called “*good_bad*”, with a binary outcome: “*1*” if the review was considered good (only 4 or 5 stars), “*0*” if the review was considered bad (either 1, 2 or 3 stars).

Two sort of representation of data were made for the purpose of binary classification: a *TFIDF* representation, and a *countvectorizer* representation (from the library *sklearn*). For each representation, we then defined three classification methods (a Decision Tree classifier, a Support Vector Machine classifier and a Random Forest classifier), to be able to compare the results among the different representations and methods used.

We divided the dataset in train set (90% of the data) and test set (10% of the data), for the purpose of classification.

Classification with *countvectorizer()*

These below are the metrics obtained as a result of the classification methods performed on data represented with *countvectorizer()* representation:

Class method	*Accuracy Score*	*F1 Score*	*Recall Score*
DT	0.77	0.82	0.83
SVM	0.88	0.92	0.93
RF	0.86	0.90	0.96

Table 2: results classification with countvectorizer() representation for Decision Tree, Support Vector Machine and Random Forest classifier

Among the methods used, we noticed that the highest score on the diagnostic scores were obtained with the Support Vector Machine classification; the “worst” method was achieved with the Decision Tree classifier, although the scores weren’t awful.

Classification with TFIDF

These below are the metrics obtained as a result of the classification methods performed on data represented with *TFIDF()* representation:

Class method	*Accuracy Score*	*F1 Score*	*Recall Score*
DT	0.77	0.84	0.84
SVM	0.90	0.92	0.95
RF	0.87	0.91	0.96

Table 3: results classification with TFIDF() representation for Decision Tree, Support Vector Machine and Random Forest classifier

The situation, with TFIDF representation, was a little better than the scores obtained with the countvectorizer() representation: in fact, alle the scores were tied or higher than the previous one.

In particular, among all methods and representations, the best one, according to the scores obtained, was the classification with Support Vector Machine and TFIDF representation.

RESULTS AND CONCLUSIONS

Our project focuses on two task, topic modeling and topic classification, performed on the Yelp dataset.

For the topic modeling task, we expected to see some words related to Italian food as the most common topics, and our expectations were confirmed: "pizza" and "pasta" were among the most common words, and in particular "pasta" was among the most important words in each of the 10 topics obtained; other positive words were among the most common in the topics obtained, such as "good", "great", "come back", "great service" and "good time".

For the purposes of topic classification, we considered three different classification methods (a Decision Tree classifier, a Support Vector Machine classifier and a Random Forest classifier) and two text representations (a *TFIDF* representation, and a *countvectorizer* representation): we expected better results for the TFIDF representation, as it focuses both on the frequency of words in the corpus and it also provides the importance of words. Our expectations were confirmed.

Among the different classification methods, the Support Vector Machine (with TFIDF, only method able to get all scores above 0.90 on the metrics) classifier achieved the best results metric-wise, while the Decision Tree obtained the worst results among the methods chosen.

TABLE INDEX

FIGURES INDEX

REFERENCES

- Text Mining and Search notes
- https://www.yelp.com/dataset/
- https://scikit-learn.org/stable/
- https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2
- https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf