Introduction

Natural language processing (NLP) or computational linguistics is one of the most important technologies of the information age. Applications of NLP are everywhere due to the fact that now humans communicate almost everything in language: web search, ,emails, language translation customer service, , virtual agents, medical reports advertising, etc.

Cite

[Aman's AI Journal • CS224n: Natural Language Processing with Deep Learning](https://aman.ai/cs224n/)

Steps taken (code)

Libraries imported:

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

import warnings

warnings.filterwarnings('ignore')

from nltk.corpus import stopwords

from sklearn.feature\_extraction.text import TfidfTransformer, CountVectorizer

#from sklearn.metrics import classification\_repo

#rt, confusion\_matrix

from sklearn.model\_selection import train\_test\_split

import string, nltk

from nltk import word\_tokenize

from nltk.stem import PorterStemmer

from nltk.stem import WordNetLemmatizer

https://regenerativetoday.com/exploratory-data-analysis-of-text-data-including-visualization-and-sentiment-analysis/

**Exploratory data analysis**

The chosen dataset was sourced form outscraper.com . It consists of x data columns and x data rows or google maps reviews using the queries with ‘Dublin’ as a location and ‘restaurant’ as a filter. The csv data was read into a panda’s dataframe. The dataset contain’s 8 variables with qualitative data such as:

* ‘reviews’,
* ‘rating’
* ‘author id’
* ‘owner\_answer\_timestamp’
* ‘review\_rating’
* ‘review\_timestamp’
* ’ review\_likes’
* ‘review\_id’,

and 17 variables with quantitative information such as:

‘query’

‘names’

‘google\_id’

* ‘place\_id’
* ‘location link’
* ‘review\_per\_score’,
* ‘review\_id’
* ‘author\_link’
* ‘author\_title’
* ‘author\_image’
* ‘review\_text’
* ‘review\_img\_url’
* ‘review\_img\_urls’
* ‘owner\_answer’
* ’ owner\_answer\_timestamp\_datetime\_utc’
* ‘review\_link ‘
* ’ review\_datetime\_utc’

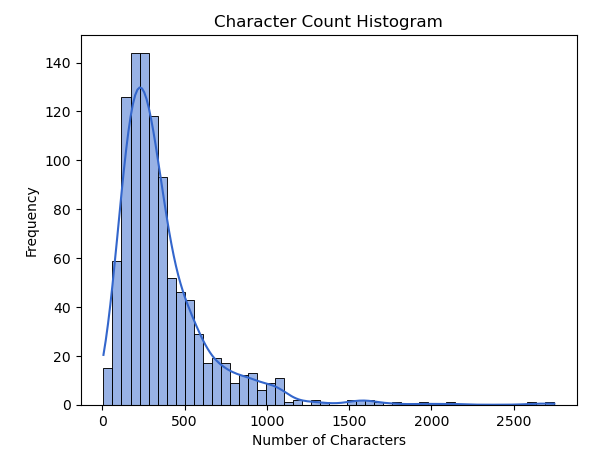
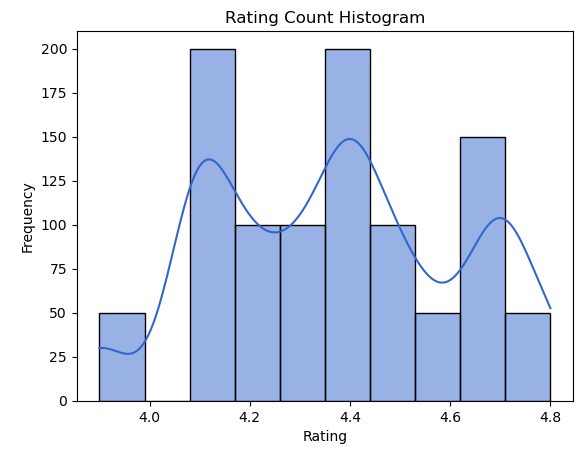
As part of the data reduction the columns 'query', 'google\_id', 'location\_link', 'reviews\_link','reviews\_per\_score', 'review\_id', 'review\_img\_urls','author\_image', 'owner\_answer\_timestamp', 'owner\_answer\_timestamp\_datetime\_utc', ‘review\_datetime\_utc', 'author\_title','review\_img\_url','author\_link','review\_timestamp' were deleted to maintain the privacy of the reviewers in the dataset and to focus on the following 6 categorical variables: ‘name’ ( business name), ‘place\_id’, ‘rating’, ‘review\_id’ , ‘owner\_answer’, ‘review\_text’ and ’ review\_likes’ as part of the NLP processing for this report. The name of the restaurant, the place id and the reviewer id, are the independent variable, while the rating, the owner answer, the number of reviews likes and the actual review text are dependent variables.

Exploratory data analysis showed an initial dataset of x rows and x columns. Summary statistics of the dataset were obtained using the python function ‘.describe’ such as count, mean , standard deviation, minimum and maximum values and the quantiles. ‘Review rating‘ showed a range of between 1 and 5 stars, with a mean value of 4.3. The range of ‘review likes’ spread from 1.5 to 17 like showing a level of engagement among reviewers.

A diagram of data processing process

Description automatically generated

Visualizations of review length, the distribution of the word count, and the sentiment polarity, stop word distribution, and character count distribution were generated to judge if the dataset is skewed in any way.

Applying the vadar compound Word clouds provided visualizations of the frequency of word occurrences in the dataset.

A black background with words

Description automatically generated

**Data cleaning consisted of:**

The text cleaning was completed as part of the preparation for the application of NLP and classification. The number of stop words, the frequency and type of occurring punctuation symbols as well as the syntactic and lexical category quantities in each review adds valuable information and contributes to its correct classification.

Other pre-processing was necessary to remove certain characters such as hashtags, emoticons, non-word characters to ensure that the analysis could be performed effectively and automatically.

[How to use NLTK for POS tagging in Pandas (practicaldatascience.co.uk)](https://practicaldatascience.co.uk/data-science/how-to-use-nltk-for-pos-tagging-in-pandas)

This included tokenizing completed with Averaged Perceptron Tagger as part of the Textblob package. To analyse and structure each individual word, tokenizing removed the spaces and added a split character between each word. The pos\_tags list was generated, which contains tuples of words and their corresponding part-of-speech tags.

Each unit can now be fed into model as input. Chunking was also performed with this package, which segments and labels multi-token sequences to ensure noun phrases such as longer restaurant names ‘Fire Steakhouse & Bar’ are being recognised. This preprocessing permits a model to learn relationships between words as it is now represented as a vector. FreqDist from nltk.probability was used to calculate the frequency distribution of words and identify common or rare words.

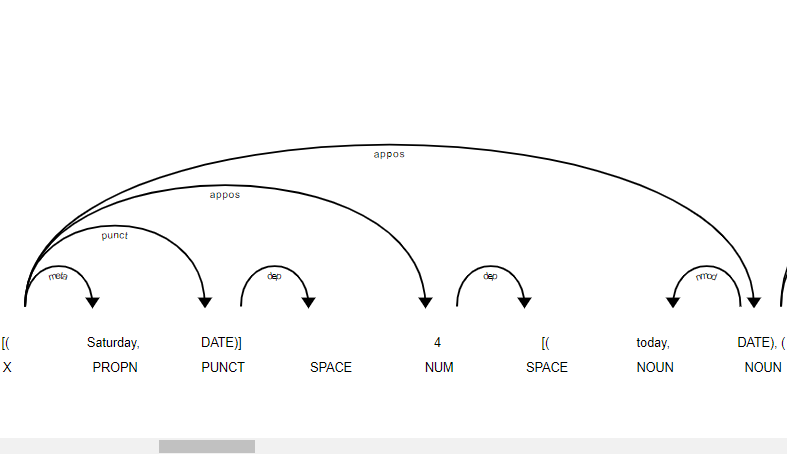
N-grams were used to extract certain features from the dataset such as the frequency of certain words for positive and negative data.

The spaCY package was used to classify the column ‘review\_text’ into named entities and their labels such as locations, cardinal numbers, noun’s and dates. The displacy visualizer allows a sentence to be broken up and it’s dependencies to be examined. This package was useful is analyzing the structure of a sentence and to check for specific details such as dates and quantities. Inauthentic reviews will typically lack any information that it uniquely relevant to the service and will contain generic comments that suit any restaurant such as ‘Great service’ or ‘Good place for an occasion’. Each named entity such as GPE (location), cardinal, quantity, date and time were stored in an additional column ‘named entities’.

A screenshot of a computer

Description automatically generated

The dependency parsing DEP was also generated inside the jupyter notebook. This is particularly useful in semantic role labeling (SRL) and information extraction. The arc label describes the type of syntactic relation which connects a dependent child word to a head variable, such as punctuation, meta data or appositional modifier of a noun.



Data enrichment was executed to extract more information from the review text column as part of the NLP process.

The counter class imported from ‘collections’ module was used to process and store punction types and counts for analyzing frequency and generate visualizations of punctuation types. The punctuation marks of each line were accumulated in a dictionary and each element was transposed into data frame columns for a greater overview. Inauthentic reviews typically have typos, either an excessive amount or complete lack of punctuation in relation to their word count and poor grammar. Three additional columns were added: punctuation count, punctuation list and accumulated punctuation dictionary.

Sentiment Analysis was completed with Vadar, (Valence Aware Dictionary and sentiment reasoner) and Sentiment Intensity Analyzer from the nltk package. Since this tool is a lexicon and operates on rule-based sentiments, it is particularly suited to social media language, which is appropriate for this dataset. This added 2 additional variables to the dataset; vadar compound which is a numerical variable indicating sentiment between 0 and 1 and a second vadar sentiment is a categorical variable; positive, neutral and negative depending on the individual review. Thresholds were set to +/-0.5.

A graph with blue squares

Description automatically generated

Vadar compound shows a mean sentiment of 0.78, a min of -0.97 and a max of 0.99.

Analyzing the vadar sentiment results showed 930 positive reviews, 62 negative reviews and 8 neutral reviews indicating most reviewers were happy with the business.

**Data processing consisted of:**

**Rule based classification:**

8 rule-based classifiers were developed based on the research completed in the literature review using predefined linguistic rules and patterns. These rules aid in categorizing the reviews into specific groups based on the preprocessed data. The extracted categorical and quantitative data will be used to determine if a review is authentic or not, then the result will be compared to machine learning models. None of the rules are individually exhaustive, rather each of them is an indicator that the review may need to be flagged and contains some of the common warning characteristics of an inauthentic review. If a check for inauthenticity succeeds, the review is marked with a binary output of 1, if it fails 0. The sum of the characteristics was chosen as the decisive factor for determining a final pass or fail.

* ‘Check 1’ examines if the dataset contains multiple reviews for the same restaurant from the same person using if-else statements. Is the number of places reviewed, less than the review count for an individual author id? Leaving multiple reviews is an indicator that the reviewer has not visited the business and either has a financial incentive or a personal agenda against the business.
* ‘Check 2’ examines if the author has submitted more than 1 review in the dataset. Serial reviewers maybe be looking for free gifts from a business or working for a specific platform
* ‘Check 3’ uses if/else statements to determine whether a reviewer leaves reviews that are extremely positive or negative based on their average vadar compound result from the sentiment analysis. If their average is less that -0.6 or greater than 0.99, all their reviews were flagged.
* ‘Check 4’ uses string punctuation to count the number of punctuation marks per review and flags if the count was greater than 10
* ‘Check 5’ reads the rows of the review to see if an owner has replied to the review. If the owner has acknowledged and engaged with the review, it was taken as a sign that the review was genuine
* ‘Check 6’ counts the number of characters in a review length. This is a significant point to distinguish spam reviews. If the review substance is excessively short, we can assume the reviewer did not consider the restaurants experience fully. Threshold was set to 150 characters
* ‘Check 7’ uses the preprocessing completed with the spaCY package to asses the level of detail in each review. That is whether the reviewer has left specific details such as names, locations, dates, times and percentages. Separate functions were written to count each type of detail per row of review. If this count was less than 10, the review was flagged.