Enhancing the Accuracy of Inauthentic Review Detection using Machine Learning and Sentiment Analysis

Teresa Quain

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

August 2022

September 2023

Supervisor: Vikas Tomer

Contents

[Problem Statement 2](#_Toc144536688)

[Introduction 3](#_Toc144536689)

[Hypothesis and Research Objectives 5](#_Toc144536690)

[Literature Review 6](#_Toc144536691)

[Literature Summary Table 6](#_Toc144536692)

[Methodology 6](#_Toc144536693)

[Dataset description 6](#_Toc144536694)

[Exploratory data analysis 8](#_Toc144536695)

[Data preprocessing 17](#_Toc144536696)

[Rule based classification 22](#_Toc144536697)

[Machine Learning 25](#_Toc144536698)

[Supervised Learning Method 25](#_Toc144536699)

[Unsupervised Learning Method 25](#_Toc144536700)

[Results and Evaluation 29](#_Toc144536701)

[Data pre-processing 29](#_Toc144536702)

[Rule-based classification 29](#_Toc144536703)

[Machine learning 31](#_Toc144536704)

[Supervised 31](#_Toc144536705)

[Unsupervised 33](#_Toc144536706)

[Conclusion 34](#_Toc144536707)

Research title and Topic Area Introduction

Cite

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

# Problem Statement

As a result of this inauthentic’ reviews of products and services has become an epidemic online. Customers depend on genuine reviews to inform them for quality and economic and safety purposes.  These nuggets of wisdom and caution aim to bridge information and issues between buyers and sellers, or buyers and other buyers by providing information that may not be otherwise disclosed. Businesses are paying to have more positive reviews about them online to increase sales and hotel stays to increase their profits. Other reviews have been created by consumers who have been financially incentivised.  Dubious performing businesses can damage the reputation of a platform for other transparent businesses. They tend to also generate revenue for that business which wouldn’t otherwise have been generated which raises ethical and legal concerns.

Inauthentic reviews can lead to financial and safety risks for consumers as they may end up buying low-quality products or services or worse, be at risk of scams or fraud.

In addition to financial and safety concerns, inauthentic reviews also pose ethical and legal issues. Businesses that engage in fake review practices, such as paying for positive reviews or incentivizing consumers to leave reviews, are engaging in dishonest and manipulative behaviour that undermines the credibility of the review system. Furthermore, the presence of inauthentic reviews on a platform can harm the reputation of the platform and legitimate businesses that operate on it. Consumers may lose trust in the platform, leading to reduced traffic and revenue for both the platform and honest businesses. This can lead to a vicious cycle in which legitimate businesses are forced to compete with fake reviews to maintain their visibility, further eroding trust in the review system and the platform as a whole. Such practices also unfairly advantage some businesses over others, leading to an uneven playing field in the marketplace.

# 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 via an online language: web searches, emails, language translation customer service, virtual agents, medical reports advertising and much more.

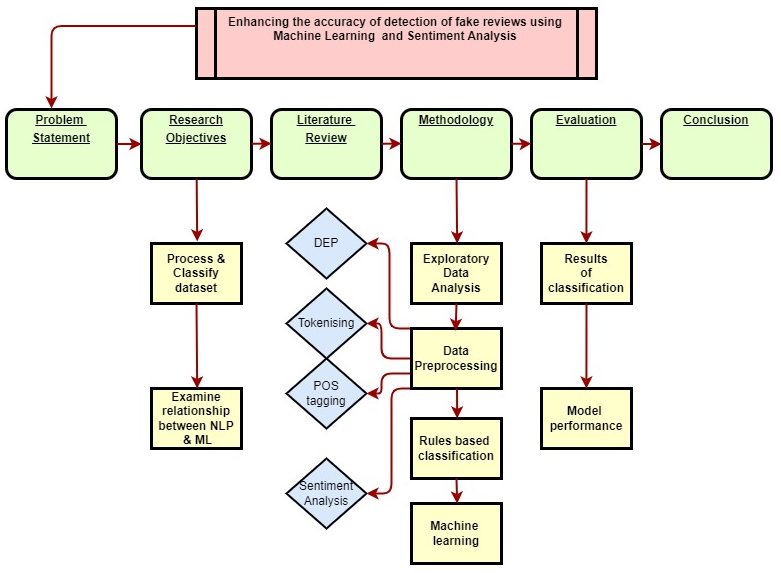


Figure Thesis structure flowchart

The structure of this thesis will follow fig 1. above:

* The problem statement and ramifications of inauthentic reviews online will be outlined
* Three research objectives will be presented to summarize the purposes of the study and organise the thesis into clearly defined components
* The literature review and summary table will summarize the wide range of sources consulted for this paper and identify the gaps in existing knowledge
* The methodology will be based on the code programmed in the Jupyter Notebook. It will consist of Exploratory Data Analysis (EDA) section, preprocessing using NLP techniques such as Dependency Parsing (DEP), Tokenisations, Part-of-Speech Tagging (POS) and sentiment analysis. Rule-based-Classification will be outlined and justified. The Machine Learning (ML) will follow as an alternative to the Rule-based-Classification. This will include both supervised and unsupervised learning methods.
* The Evaluation section will present the core findings of the Rule-based-Classification and ML modelling and describe how the outcomes were obtained and analysed
* The conclusion will restate the original research objectives and present the evidence from the evaluation section supporting the findings. The entire paper will be summarized, and the key ideas discussed

In this thesis, the primary programming language that will be utilized is Python. Python is widely used in the field of data science and machine learning. It has an extensive collection of libraries and frameworks that can be utilized for NLP, natural language processing. Jupyter notebooks will be employed to write the code for the data cleaning as part of the pre-processing, the feature engineering, the sentiment analysis and the model training and critical evaluation. For NLP techniques, libraries such as NLTK (Natural Language Toolkit), scikit and TextBlob will be used for tasks such as tokenization, part-of-speech tagging, and named entity recognition. These libraries will also be used for feature engineering tasks such as extracting bag-of-words and tf-idf features.

Various machine learning algorithms will be utilized in this thesis proposal, including supervised and unsupervised learning algorithms, such as logistic regression, decision trees, and support vector machines (SVM) will be employed. The use of Python, Jupyter Notebooks, NLP techniques, and machine learning algorithms will be crucial in achieving the research objectives of this master's thesis. The combination of these tools and techniques will enable the development of an accurate and efficient algorithm for inauthentic review detection. Machine learning and sentiment analysis techniques can be leveraged to identify patterns of suspicious activity in review data, such as an unusually high number of positive reviews from a single IP address or similar language used across multiple reviews.

The colours of the various graphs, flowcharts and other plots are based on the google colour palette to reflect the dataset which consists of google reviews.

This topic is particularly interesting because many consumers have experienced it themselves while travelling, where a restaurant has a huge number of recently published, highly positive reviews which give a false positive image of the business. Consumers can find it extremely frustrating, and it may push a platform over the edge completely due to a lack of trust and a poor reputation. This manipulation of reviews can be dangerous, as it can mislead customers into making poor decisions that may have serious consequences. There is a good reason why customers are encouraged to be aware of this practice, and it is also important that they take steps to ensure that the reviews that they read are genuine.

**Keywords: Sentiment analysis; machine learning; e-commerce; natural language processing**

# Hypothesis and Research Objectives

The hypothesis of this research is that AI can detect online reviews which are not left by a genuine consumer, by using natural language processing techniques and or machine learning. These falsified reviews are intended to generate additional business for a product or service. The enhanced and additional information, which this model would provide, would allow for more accurate financial projections of a business and promote a stronger customer base. Moreover, it would also enhance a company's reputation online and create more opportunities to conduct targeted marketing campaigns and increase sales.

Through this analysis, the following research objectives will be pursued:

1. To process and classify a text dataset in depth of online reviews from restaurants in Ireland using NLP techniques such as POS tagging, entity identification and semantic analysis
2. To implement a rule-based classification system to detect if online reviews left are completed by authentic patrons.
3. To generate several machine learning models on the trained dataset and compare their performance results and overall effectiveness with the rule-based classification system

# Literature Review

## Literature Summary Table

# Methodology

The analysis will be set out in the following format as seen in fig 2. below. This methodology will describe the initial dataset, the EDA, exploratory data analysis that was performed, the preprocessing necessary for this dataset, the rules-based classification system based on predefined NLP methods and a comparison to alternative machine learning methods.

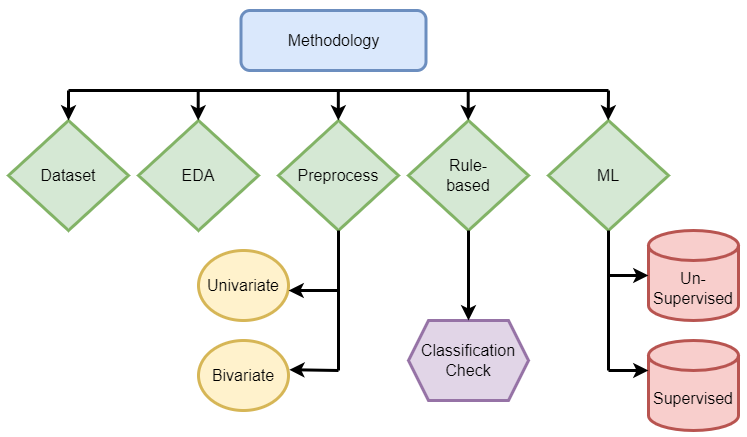


Figure Methodology structure

## Dataset description

The chosen dataset was sourced form outscraper.com . It consists of 26 data columns and 1328 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 contains 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 variables, while the rating, the owner answer, the number of reviews likes and the actual review text are dependent variables..

The following libraries were imported for various data analysis, preprocessing, and machine learning tasks: NumPy (numeric computations), pandas (data manipulation), seaborn (data visualization), Matplotlib (plotting), NLTK (Natural Language Toolkit for text processing), scikit-learn's TfidfTransformer and CountVectorizer (feature extraction for text data), and train\_test\_split (data splitting for machine learning). Additionally, warnings have been filtered out, and the inline display of Matplotlib plots is enabled. Libraries related to classification reporting and confusion matrix are also included. String processing tools, such as tokenization, stemming, and lemmatization, are available through the NLTK library.

Top of Form

## Exploratory data analysis

Exploratory data analysis was conducted to perform the initial investiagtions on the dataset to discover pattern, trends and correlation between variables. It was set out in the format of fig. 3 below. The information gained allowed summary statistics and graphical representations of the data to be 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 and allow an overview of the data. Both univariate and bivariate analysis were undertaken, to understand each variable individually and how they are related to another.

A diagram of a diagram of a diagram

Description automatically generated

Figure Exploratory data analysis structure

Missing data can distort results and reduce statistical power. Reviews are the main information in this dataset. If any row is missing the review, it was deleted. Estimating or imputation of the missing data would not make sense in this instance. The code *df.isnull().sum()* was used to calculate and display the count of missing values in each column the DataFrame. Other columns with missing data included: ‘review\_img\_url’ and owner\_answer\_timestamp. These were also not required for the dataset analysis and were deleted as part of the data reduction.

**Algorithm 1:** Exploratory Data Analysis, EDA

[Google Colors - Hex, RGB, CMYK, Pantone | Color Codes - U.S. Brand Colors (usbrandcolors.com)](https://usbrandcolors.com/google-colors/)

#Univariate Analysis

1. Describe the dataset , data types, summary statistics, data frame rows and columns

***df.info()***

***df.describe()***

***df.shape()***

1. Create histograms with seaborn of single variables, google hex colour scheme

**sns.distplot(df["rating"], color = '#0F9D58')**

**sns.distplot(df["review\_likes"], color = '#0F4B400)**

**sns.distplot(df["char\_count"], color = '#04285F4)**

1. Create word clouds for reviews based on start rating

**consolidated=' '.join(word for word in df['review\_text'][df['rating']<4].astype(str))**

**wordCloud=WordCloud(width=1600,height=800,random\_state=21,max\_font\_size=110)**

**plt.figure(figsize=(15,10))**

**plt.imshow(wordCloud.generate(consolidated),interpolation='bilinear')**

1. Tabulate tables insert code, most common words

#Bivariate Analysis

1. Create correlation heatmaps

**heatmap = sns.heatmap(df.corr()[['vadar compound']].sort\_values(by='vadar compound', ascending=False), vmin=-1, vmax=1, annot=True, cmap='Blues')**

1. Insert pair plot code

[Step-by-Step Exploratory Data Analysis (EDA) using Python - (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/)

Summary statistics in table 1 below using the .descibe() function provided a high level overview of certain information such as such as count, mean, standard deviation, minimum and maximum values and the quantiles of the reviews themselves, the rating and the review likes.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **reviews** | **rating** | **author\_id** | **owner\_answer \_timestamp** | **review\_rating** | **review \_timestamp** | **review\_likes** | **reviews\_id** |
| **count** | 1.0E+03 | 1.0E+03 | 1.0E+03 | 1.7E+02 | 1.0E+03 | 1.0E+03 | 1.0E+03 | 1.0E+03 |
| **mean** | 2.7E+03 | 4.4E+00 | 1.1E+20 | 1.7E+09 | 4.4E+00 | 1.7E+09 | 7.7E-01 | 2.9E+17 |
| **std** | 2.5E+03 | 2.4E-01 | 5.4E+18 | 1.4E+07 | 1.1E+00 | 3.2E+07 | 1.5E+00 | 5.1E+18 |
| **min** | 8.5E+02 | 3.9E+00 | 1.0E+20 | 1.6E+09 | 1.0E+00 | 1.5E+09 | 0.0E+00 | -9.1E+18 |
| **25%** | 1.6E+03 | 4.2E+00 | 1.0E+20 | 1.7E+09 | 4.0E+00 | 1.7E+09 | 0.0E+00 | -4.3E+18 |
| **50%** | 2.1E+03 | 4.4E+00 | 1.1E+20 | 1.7E+09 | 5.0E+00 | 1.7E+09 | 0.0E+00 | 7.1E+17 |
| **75%** | 3.0E+03 | 4.5E+00 | 1.1E+20 | 1.7E+09 | 5.0E+00 | 1.7E+09 | 1.0E+00 | 4.0E+18 |
| **max** | 1.3E+04 | 4.8E+00 | 1.2E+20 | 1.7E+09 | 5.0E+00 | 1.7E+09 | 1.7E+01 | 9.0E+18 |

Table Summary statistics of dataset

After removing the missing data, the dataset contained 1328 reviews.

* The minimum rating is 2.8 stars, and the maximum featured rating is 4.9 stars.
* The mean number of review likes is 0.66, (less than 1 per review)

#### Univariate Analysis

For the univariate analysis, charts such as histograms, wordclouds and density plots were used to visualize the data using Matplotlib and Seaborn libraries.

A graph with green lines

Description automatically generated

Figure Histogram of Ratings distribution

Review ratings showed a range of between 3.8 and 5 stars, with a mean value of 4.3 in fig 4.

* The ratings are normally distributed with most of the reviews having a rating of between 4.2. and 4.6.
* This may indicate a high number of falsly positive reviews if the reviews are found to be inauthentic
* It may also indicate that the standard of restaurants in the datast is high and consumers are satisfied with the experience.

The range of ‘review likes’ is spread from 0 to 17.5 in table 2 below, with a skew to the right.

* This shows a low level of engagement among reviewers. Using the .value\_counts() function shows 608, 46% reviews of the total 1328, have 0 likes. 233, 18% have 1 like and 75, 0.06% have 2 likes.

|  |  |
| --- | --- |
| review\_likes | count |
| 0 | 860 |
| 1 | 281 |
| 2 | 95 |
| 3 | 38 |
| 4 | 24 |
| 5 | 12 |
| 6 | 8 |
| 7 | 5 |
| 17 | 2 |
| 13 | 1 |
| 14 | 1 |

Table Count of review likes

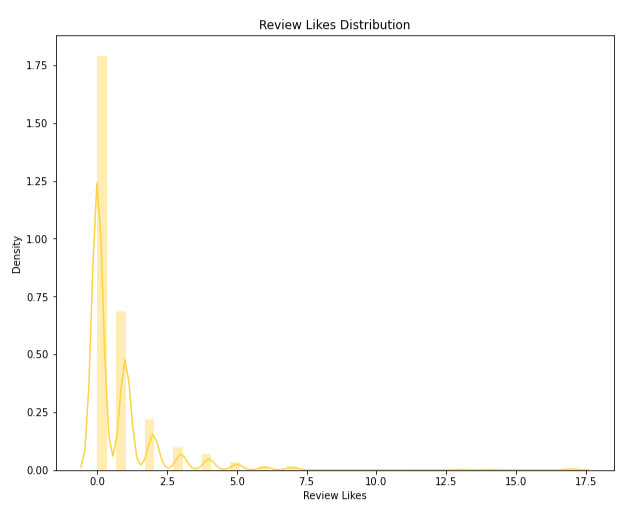


Figure Histogram of review likes distribution

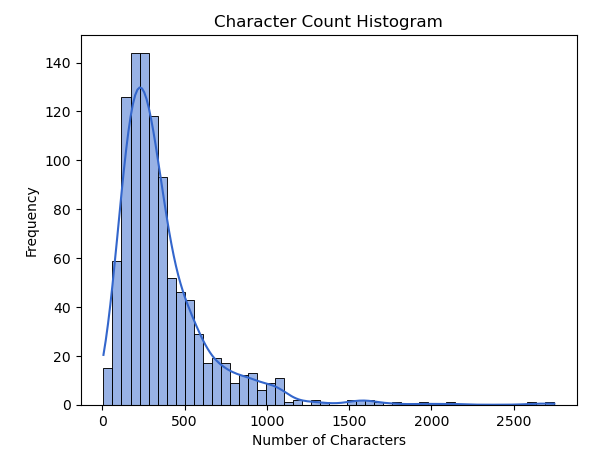


Figure Histogram of character count distribution

Character count per review ranged from 0 to 2500 characters with a distribution skewed postivly to the right in fig. 6. The distribution's tail is skewed to the right, indicating that there are relatively fewer extreme values on the right side and a concentration of lower values on the left side. Over 140 reviews have a character count between 0 and 500. This can indicate either the reviewers are leaving less detail per review, or that the reviews are positive and to the point with no issues highlighted.



Figure Word count of frequent words

From the above word cloud in fig. 7, the following observations were made

* The most frequented words are food, service, staff and Dublin. This suggest that the quality of the food, the service in the restaurant from staff and the location are important to customers.
* A lot of the words in the word cloud have positive connotations such as tasty, friendly, good, quality, excellent, incredible and highly recommend . This suggest most of the reviews are highly rate and have a positive sentiment rating

#### Bivariate Analysis

For the bivariate analysis a combination of correlation heatmaps, pairplots and countplots were used to understand how the variables are related to one another and asses the dependency relationships.

A blue and white graph

Description automatically generated

Figure Heatmap of variable correlation with Sentiment

The above correlation heatmap in cmap colour ‘Blues’, fig 8 shows each variable and it’s correlation with the ‘vadar compound’, (sentiment) variable

* Review rating is the darkest at 0.66 correlation, followed by review likes. This indicates that a positive review tends to get more ratings and likes, than a negative one.

A green bar chart with white text

Description automatically generated

Figure Heatmap of variable correlation with Review likes

Fig. 9 in cmap colour ‘Greens’ theme shows each variable’s level of correlation with ‘Review likes’.

* ‘Check 5‘ (checks if the owner has replied to the review) and ‘check 7‘ (checking number of details) are negativly correlated at -0.0064 and -0.19. This indicates that the polarity of a review has no effect on whether or not the owner of the buisness engages with the review or is related to the number of details they have included in their review.
* Review length and punctuation count are similarly correlated at 0.24 and 0.22 respectivly. This indicates that other reviewers appreciate a longer review with more detail.

Fig 10. below shows a seaborn countplot of names of restaurants and the number of stars per restaurant.

* Purple is the most predominant colour showing restaurants in the dataset are highly rated with 5 stars. The Chaper One restaurant has over 40, 5 star reviews
* Sophies restaurant is rated 15 times with 4 star reviews

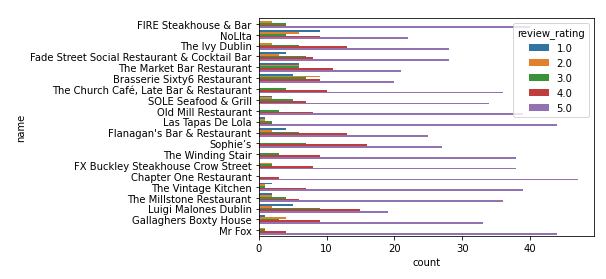


Figure Count plot of restaurants vs. review rating

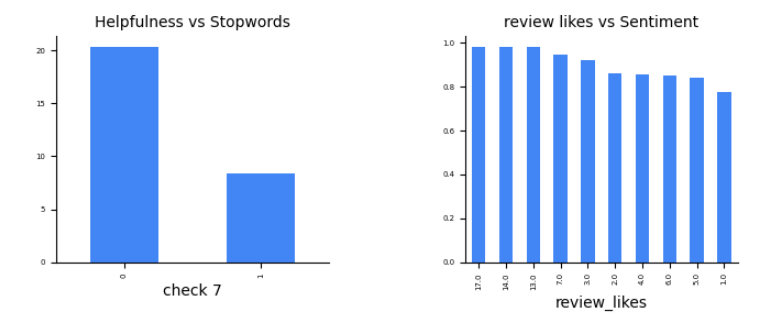


Figure Subplots of helpfulness, stop words, review likes & sentiment

Figure 11 above shows 2 subplots generated with the matplotlib library. The first shows check 7 vs. the number of stop words. Check 7 (review contains at least 10 details, that would be helpful to other customers reading the review). Reviews which meet check 7 (1) have a lower number of stop words, than those that don’t (0). The second subplot shows the number of review likes vs the sentiment value. As expected, reviews with a higher number of likes (14 and 17) have a higher positive sentiment rating 0.9 to 1. Reviews with less likes have a slightly lower rating (0.8) which is still very positive.

## Data preprocessing

The text pre-processing was completed as part of the preparation for the application of Natural Language processing, classification, and machine learning. The frequency of punctuation symbols as well as the syntactic and lexical category quantities in each review adds valuable information and contributes to its correct classification. The below flowchart in fig. 12 demonstrates the libraries and methods involved in the preprocessing of the dataset.

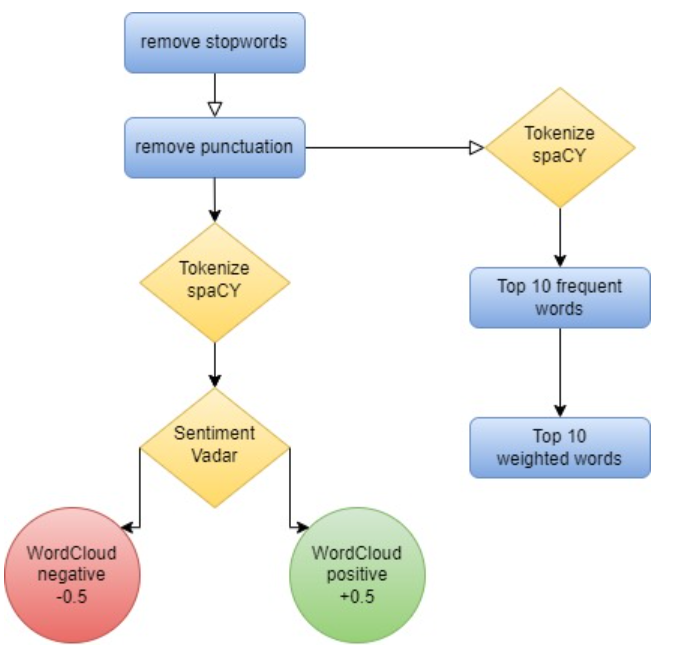


Figure Data Preprocessing structure

The [‘review\_text’] column was also modified to lower case and stop words were removed to allow for more accurate analysis.

Since stop words do not contribute any meaningful information to this text classification model, they were removed from the corpus. This allows for a smaller, more focused dataset. The ‘describe’ function counted 1328 stop words in total in the dataset. This lower-level information also prevents significant words that may contribute to the sentiment analysis later in the analysis from being treated.

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. This included exclamation marks, comma’s, dollar signs, hashtags, underscores, emoticons, non-word characters, at symbols, full stops, brackets, and colons. Inauthentic reviews typically have typos, either an excessive amount or complete lack of punctuation in relation to their word count and poor grammar. The punctuation marks of each line were also accumulated in a dictionary and each element was transposed into data frame columns for a greater overview. Three additional columns were added: punctuation count, punctuation list and accumulated punctuation dictionary. This information will be reused later for the rule-based classification.

The Tf-idf library was employed to provide more information on the content of each review. The proportion of occurrences of a certain term to the total number of that term in the dataset provides an insight into emphasis and importance a reviewer has attributed to term. This preprocessing permits a model to learn relationships between words as it is now represented as a vector. Count vectorizer from the sklearn library was used to convert the collection of text into a matrix of tokens. N-gram count of the number of unique words is 976. The count vectorizer was also used to create a bag-of-words representation of the text data which results in a matrix of the reviews. 1328 rows representing the number of reviews and 976 columns representing the unique words. The matrix has non-zero count of 30047 and a relatively low sparsity of 2.32%.

* *sparse matrix shape: (1328, 982)*
* *nonzero count: 30319*
* *sparsity: 2.32%*

The TF-IDF transformed weights were converted to a NumPy array type dataset to present the keywords, weights and sums for each review and demonstrate their context in the complete corpus. The top 10 most frequent words are shown below in table 3. ’Food’ is by far the most common and is featured 1062 times, and ‘good’ is the second with 563 times. Other frequent words featured in the top 10 list have generally positive connotation such as ‘great’, ‘nice’ and ‘delicious’.

|  |  |
| --- | --- |
| **Term** | **Occurrences** |
| food | 1062 |
| good | 563 |
| service | 555 |
| great | 529 |
| staff | 439 |
| place | 417 |
| restaurant | 347 |
| nice | 331 |
| would | 284 |
| delicious | 278 |

Table Top 10 frequent words

The top 10 most heavily weighted words are shown below in table 4. ‘Food’ is naturally the heaviest weighted word at 0.0625 and ‘good’ being second at 0.0485. Other heavily weighted words include staff, place and more positive adjectives which provides evidence that the reviews in the dataset are relevant to the subject and overall positive in nature.

|  |  |
| --- | --- |
| **Term** | **Weight** |
| food | 0.062 |
| good | 0.048 |
| great | 0.046 |
| service | 0.042 |
| staff | 0.036 |
| place | 0.035 |
| nice | 0.034 |
| restaurant | 0.029 |
| delicious | 0.029 |
| amazing | 0.027 |

Table Top 10 weighted words

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 as seen in fig. 13 below. 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 specific 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’.

The code **‘displacy.render(nlp(str(df['named\_entities'])), jupyter=True, style='ent')’** allows the named entities to be annotated based on their entity type and displayed using and ‘ent’ stype visualisation

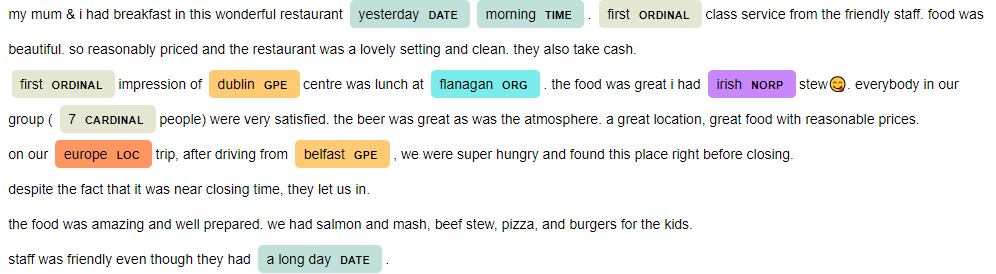


Figure Review with displacy visualizer

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 in fig. 14 below 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.

A diagram of a diagram

Description automatically generated

Figure Named entity structure visualization

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

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. Through consumer sentiment analysis, companies can detect eh polarity of the review, gage the reaction of competitors and gain insight from their consumers.

A graph with blue squares

Description automatically generated

Figure Sentiment score vs. Rating

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

|  |  |
| --- | --- |
|  | vadar compound |
| count | 1328 |
| mean | 0.783034 |
| std | 0.360879 |
| min | -0.9493 |
| 25% | 0.7964 |
| 50% | 0.92 |
| 75% | 0.9638 |
| max | 0.9991 |

Table Sentiment summary statistics

Analyzing the vadar sentiment results in table 6, there is 1242 positive reviews, 66 negative reviews and 20 neutral reviews.

|  |  |
| --- | --- |
| vadar sentiment | count |
| positive | 1242 |
| negative | 66 |
| neutral | 20 |

Table Sentiment count

Like the other visualization’s, this graph shows a single side to the data. The high count of positive reviews with a vadar sentiment score of between 4 and 5 indicate that the reviewers are generally happy with the experience or they maybe be gushing or overexaggerating their feelings to present a different picture of the restaurant.



Figure Word cloud of positive reviews

According to the word clouds for positive (above) in fig.16 and negative words(below) in fig. 17 reviews, the words service, friendly, delicious, recommend, atmosphere and staff appeared most frequently among the positive reviews, while ‘staff, rude, poor, disappointed, expensive and unwelcoming’ appeared most frequently in the negative reviews. These indicate that the staff are performing well, are welcoming and the food is appreciated in the restaurants which have highly rated reviews, however there is controversy over the staff manner and the price of the meals.



Figure Word cloud of negative reviews

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. They are all equally weighted. 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.

A diagram of a check process

Description automatically generated

Figure Rule based classification structure

* ‘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. Highly polarized reviews are another red flag, that may indicate the reviewer is biased in their opinion and is not basing the the review on a general experience.
* ‘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.

The final check takes the sum of the Boolean results of checks 1 to 7 and tests if the integer is greater than 4, that is, that the review has succeeded positively in 4 of the 7 tests at least (more than half). If yes, the review is labelled ‘inauthentic’, otherwise it is labelled ‘authentic’ in a new column ‘label’.

|  |  |
| --- | --- |
| state | count |
| authentic | 1067 |
| inauthentic | 261 |

Table Results rule based classification

The pseudocode for the rule-based classification is shown below.

**Algorithm 2** Rule Based Classification Model

Initialize the check values

**check1 = 0**

**check2 = 0**

**check3 = 0**

**check4 = 0**

**check5 = 0**

**check6 = 0**

**check7 = 0**

Perform checks and store results

**1. If row[‘count\_of\_places\_reviewed’] < [‘review\_count’]:**

**check 1= 1**

**else**

**check 1 = 0**

**2. if row[‘review\_count’] >1**

**check 2 =1**

**else**

**check 2 =0**

**3. if row[‘avg\_vadar\_compound’] <= -0.6 or row[‘avg\_vadar\_compound’] >=0.95**

**check 3 =1**

**else**

**check 3=0**

**4. if row[‘punctuation\_count’] >10**

**check 4=1**

**else**

**check 4=0**

**5. df[‘check 5’] =np.where(pd.isna(df[‘owner\_answer’]) | (df[‘ownder\_answer’] == ‘’), 1,0)**

**6. if row[‘char\_count’] <150**

**check 6=1**

**else**

**check 6=0**

**7. if row[‘noun\_count’] +row[‘date\_count’] +row[‘ordinal\_count’]+row[‘location\_count’]**

**+row[percent\_count] < 5**

**check 7=1**

**else**

**check 7=0**

Initialize count

count = 0

**8. def sum\_check(row)**

**return if row[‘check 1’]+row[‘check 2’] +row[‘check 3’] +row [‘check 4’] +row[‘check 5’] +row[‘check 6’] +row[‘check 7’] >3**

**9. df[‘label’]=df.apply(sum\_check, axis+1)**

# Check sum count and count number of each label, true/ fake

**10. df[‘label’]. value\_counts()**

## Machine Learning

The dataset was further subset for applying the machine learning to ensure only the relevant columns were included. Columns: ‘Name’, ‘rating’, ‘author\_id’, ‘label’, and ‘review\_text’ were saved as ‘dataset 3’. An additional column called target, which will hold our target variable was created. Inauthentic reviews will be assigned a 1, and authentic reviews will be assigned a 0. The ‘review\_text’ column was split into the training and test datasets with 30% being assigned to the test datast. The model will be trained to predict the target value based on this column alone, identical to the rules-based classification method.

Both supervised and unsupervised machine learning approaches were undertaken.

### Supervised Learning Method

A dictionary of classification models was created, which included XGBClassifer, CatboostClassifier, LinearSVC, MultinomialNB, LGBMClassifier, RandomForestClassifier, DecisionTreeClassifier, ExtraTreeClassifier, AdaBoostClassifier, KNeighborsClassifier, RidgeClassifier, SGDClassifier, BaggingClassifier, BernoulliNB. Cross validation was performed using the different classifiers and their performance was evaluated in terms off the ROC AUC (Receiver Operating Characteristic Area Under the Curve). Since the AUC is a widely used measure of the accuracy of the diagnostic test, it is suitable for this application The high AUC value indicates the binary classifiers are capable of distinguishing between the different classes by measuring the separability. A loop records the ROC AUC score of each classifier including their run time.

A set of hyperparameters were defined as ‘param\_grid’ and ‘GridsearchCV’ from the sklearn package was used to methodically search through the best possible combination in order to improve the ROC AUC score.

The performance of the classifiers was evaluated with metrics such as accuracy, the precision, the recall and the ROC/AUC score.

### Unsupervised Learning Method

An unsupervised learning algorithm, Kmeans was employed to divide the review data into clusters and perform detailed analysis of the clusters to categorize reviews as fake or real. This method of vector quantization aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean.

Dataset, ‘df1’ was created with the columns: ‘punctuation\_count’, ‘review\_likes’, ‘rating’, ‘avg\_vadar\_compound’, ‘char\_count' and ‘check 7’ (number of details/review) as a subset of the original dataframe, ‘df’. Max clusters limit was set to 7 and the elbow technique was carried out to select the optimal number of clusters. In fig. 19 below, the within-cluster-sum-of-squares (WCSS) values are shown on the y-axis and the number of clusters (k) is shown on the x-axis. The elbow point and optimal value of K in this case was selected at value 4, after which the value of WCSS remains constant to the x-axis.

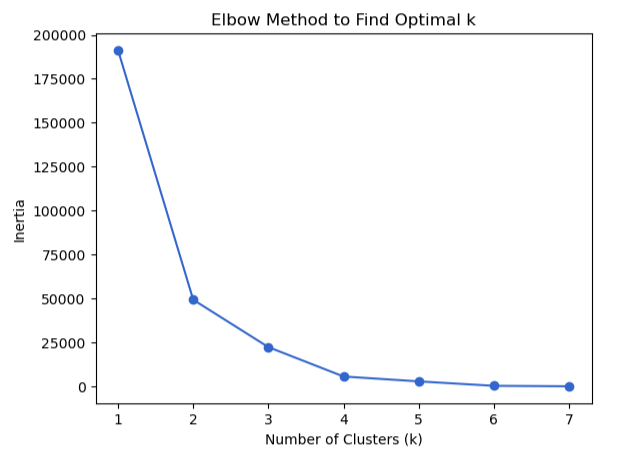


Figure Elbow method for optimal k

The KMeans object was initialized to 4, and the random seed for reproducibility was set to 42. The cluster assignments were added as a new column ‘cluster’ to the dataframe ‘df1’. The cluster centers of each cluster were obtained using the ‘kmeans.cluster\_centers\_’ command, which provides further insight into the characteristics of each cluster. As seen in table 8 below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | punctuation\_ count | review\_likes | rating | avg\_vadar\_ compound | char\_count | check 7 |
| 0 | 5.666667 | 0.666667 | 4.233333 | 0.764273 | 219 | 0.666667 |
| 1 | 13 | 0 | 4.4 | 0.764273 | 584 | 0 |
| 2 | 3 | 0 | 4.1 | 0.764273 | 83.5 | 1 |
| 3 | 11 | 0 | 4.4 | 0.764273 | 376 | 0 |

Table Cluster characteristics

Seaborn was utilized to provide some subplot visualisations of each cluster, to understand their characteristics. Below in fig. 20 is a histogram with KDE showing the distribution of likes across the dataset. The boxplots display the distribution of likes across each cluster.

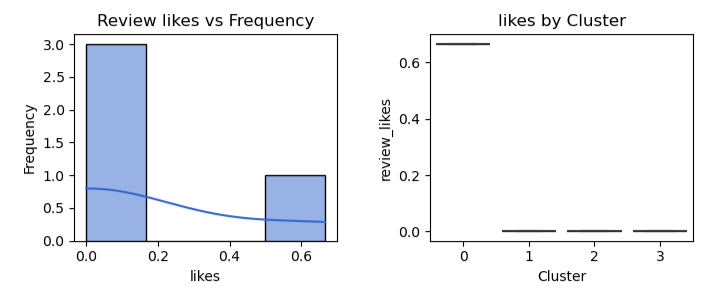


Figure Review likes vs Frequency per cluster

In this case, approximately 70% of the reviews have less than 0.2 likes on average, which 30% have between 0.5 and 0.65 likes. Cluster 0 has more likes compared to other clusters which are quite extreme values. Cluster 0 is largely different from other clusters. A high quantity of likes is another flag for inauthenticity.

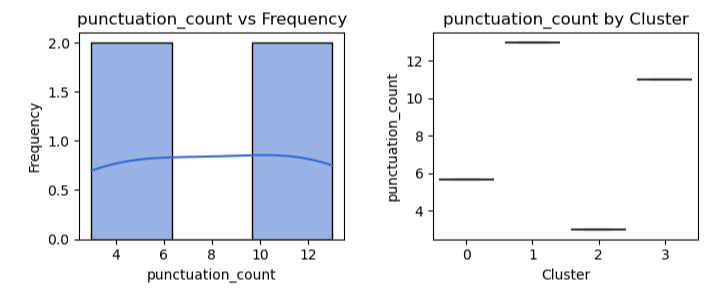


Figure Punctuation count vs Frequency per cluster

The punctuation count above in fig 21. above is equally split, 50% of the reviews have between 0-6 punctuation counts per review. The other 50% have between 10 and 12 per review. Cluster 0 and cluster 2 have less punctuation compared to 1 and 3. A high punctuation count is another flag for inauthenticity.

|  |  |
| --- | --- |
| cluster | punctuation count |
| 0 | 5.7 |
| 1 | 13 |
| 2 | 3 |
| 3 | 11 |

Table Punctuation count per cluster

Table 9 above show cluster 1 and 3 have punctuation counts of 13 and 11 respectively, while cluster 0 and cluster 2 have counts of 5.7 and 3 respectively.

A comparison of a blue square with a blue square with black text

Description automatically generated with medium confidence

Figure Sentiment vs Frequency per cluster

Fig. 22 above shows no change in sentiment data across the clusters. This indicates that each cluster contains an equal spread of both positive and negative reviews.

A comparison of a graph

Description automatically generated

Figure Level of detail vs Frequency per cluster

Fig. 23 above shows the distribution of column ‘check 7’ across the dataset and per cluster. Check 7 counts the number of named entities such as nouns, dates and place names per review, in other words the number of details. Cluster 0 and 2 contain the highest number of details per review, while cluster 1 and 3 contain lower number of details per review. A lack of details indicates generic comments and is a sign that the reviewer hasn’t physically visited the restaurant.

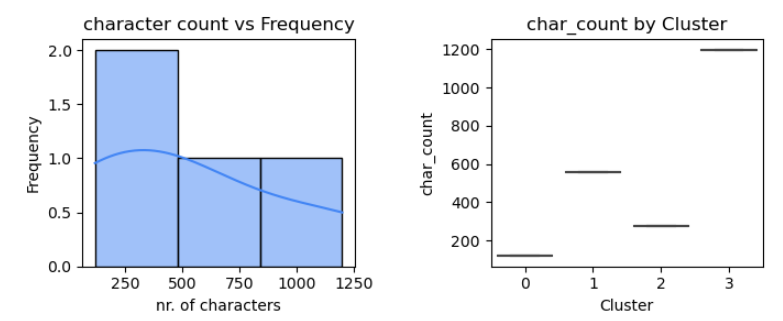


Figure Character count vs Frequency per cluster

The character count above in fig. 24 shows a higher character count in clusters 1 and 3. A long winded, irrelevant review is another flag for inauthenticity.

# Results and Evaluation

## Data pre-processing

Both univariate and bivariate analysis were carried out through the exploratory data analysis. From the univariate analysis, the following can be observed:

* The dataset contains 1328 rows and 26 columns after missing data removed
* Histograms generated show a distribution of ratings from 2.8 to 4.9 stars
* The number of likes per review in the dataset show 46% reviews of the total 1328, have 0 likes. 233, 18% have 1 like and 75, 0.06% have 2 likes
* Character count per review ranged from 0 to 2500 characters. Over 140 reviews have a character count between 0 and 500
* Word clouds from the column ‘review\_text’ showed the most frequented words are food, service, staff and Dublin. This suggest that the quality of the food, the service in the restaurant from staff and the location are important to customers

From the bivariate analysis:

* The variables correlated to a high sentiment review were the star rating (0.66) and the review likes (0.078)
* The level of review likes was positively correlated to the character count (0.24) whether the owner had responded to the review and the level of details (check 7) in the review (0.22). This indicates that other reviewers appreciate a longer review with more detail.
* A countplot of the restaurants vrs. review rating showed most restaurants have very highly rated reviews. . The Chaper One is by far the highest with over 45 5 star reviews. La Tapas and Mr. Fox is the next highest with over 40 5 star reviews

## Rule-based classification

* Check 1, Has the reviewer submitted more than 1 review in dataset?

75 reviewers were flagged for leaving multiple reviews in the dataset. 842 reviewers submitted just 1 review

* Check 2, Has the reviewer submitted multiple reviews for the same place?

3 reviewers were flagged for leaving multiple reviews for the same place, the majority (914) were not flagged.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **author\_id** | **review\_count** | **count\_of\_places\_reviewed** | **check 1** | **check 2** |
| 1.05589E+20 | 3 | 3 | 1 | 0 |
| 1.02056E+20 | 2 | 2 | 1 | 0 |
| 1.15269E+20 | 2 | 2 | 1 | 0 |
| 1.01958E+20 | 2 | 1 | 1 | 1 |
| 1.06966E+20 | 2 | 2 | 1 | 0 |
| 1.15179E+20 | 2 | 2 | 1 | 0 |
| 1.01845E+20 | 2 | 2 | 1 | 0 |
| 1.04228E+20 | 2 | 2 | 1 | 0 |
| 1.04233E+20 | 2 | 2 | 1 | 0 |
| 1.06839E+20 | 2 | 2 | 1 | 0 |

Table Top 10 reviewers ordered by review count

Table 10 shows the top 10 reviewers, ordered by reviewed the review\_count. From table 10, we can see author\_id 1.05589E+20 has submitted the highest number of reviews in the dataset (3), for 3 different places. This means check 1 is flagged but not check 2. Author 1.01958E+20 has submitted 2 reviews for the same place and is flagged in check 1 and 2.

* Check 3, Does the review have an extreme sentiment polarity value?

Of the 1328 reviews, 25 were deemed to have an extreme sentiment value. That is an average vadar compound value of greater than 0.95 or less than -0.06.

* Check 4, Does the review have a high punctuation count ( =greater than 10)?

719 reviews have punctuation counts less than 10. 281 reviews have punctuation counts greater than 10 and were flagged for check 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **name** | **author\_id** | **punctuation\_count** | **check 4** |
| Mr Fox | 1.0728E+20 | 79 | 1 |
| NoLIta | 1.0320E+20 | 51 | 1 |
| The Winding Stair | 1.1640E+20 | 51 | 1 |
| Luigi Malones Dublin | 1.1336E+20 | 45 | 1 |
| Old Mill Restaurant | 1.0787E+20 | 43 | 1 |
|  |  |  |  |

Table Top 5 restaurants order by punctuation count

Table 11, is ordered by punctuation count. Author id 1.728E+20 left a review for the Mr. Fox restaurant with the highest punctuation count (79).

* Check 5, Has the business owner verified the review by responding to it?

In 172 reviews in the dataset, the owner has responded to the review and engaged with the reviewer. In the majority of other reviews, (828), the owner has not responded to the review, and the reviews were flagged for check 5.

* Check 6, Is the review less than 150 characters? If the feedback is excessively short, it was taken into consideration that the reviewer did not genuinely experience the restaurant

854 reviews had character counts were greater than 150 characters. 146 reviews were less and were flagged for check 6.

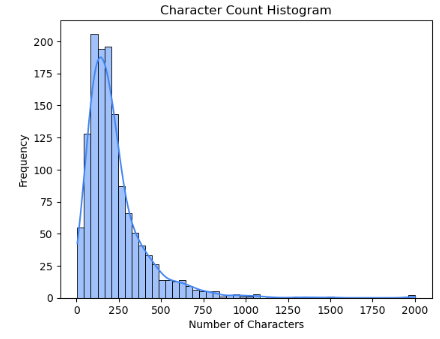


Figure Character count distribution

* Check 7, Is the review helpful to other potential customers by containing detailed information?

Check 7 takes a count of the entities of the review, gained from the spaCY package in the pre-processing section, that is nouns, numbers, dates, locations. If the sum is less than 10 ie. If there were less than 10 details in the review, the review is flagged for check 7. Using the value\_counts function in python showed 738 reviews have more than 10 details, 590 reviews have less than 140 characters.

* Overall 1067 reviews passed the authenticity check, that it, the review met less than half of the criteria for inauthenticity. 261 reviews met at least 4 of the checks for inauthenticity.

## Machine learning

## Supervised

Due to the size of the dataset and a limited hardware, the dataset was reduced to a smaller sample size to execute the machine learning. 200 random rows were selected to be used for the machine learning.

A machine learning pipeline was used, consisting of two main components: TF-IDF vectorizer for text feature extraction on the ‘review\_text’ column and SGD classifier for classification on the authenticity of the review. The pipeline was fitted to the training data (X\_train, y\_train) to make predictions on the test data (X\_test).

The set of hyperparameters ‘param\_grid’ were defined for the grid search CV. Learning rate range was specified between 0.1 and 0.01 based on the weight at the end of each batch. The number of estimators was set between 50 and 300 decision trees. Max depth for the gradient boosting was evaluated at odd values between 1 and 7 (1,3,5,7), and the subsample range was chosen between 0.6 and 1.2. The hyperparameter were chosen based on an optimum ROC output value.

The grid search outputted the best parameters as below in Table 12. The best ROC score was 0.6333, which is a poor result. A value of 0.63 means it performs better than a random example only 63% of the time. The value is relatively close to the upper left corner of the graph where specificity and sensitivity are close to 1.

|  |  |
| --- | --- |
| **Parameters** | **Value** |
| learning rate | 0.1 |
| n estimators | 50 |
| max depth | 5 |
| subsample | 0.6 |

Table Supervised ML parameters

|  |  |  |  |
| --- | --- | --- | --- |
| **model** | **run\_time** | **roc\_auc** | **roc\_auc\_std** |
| MultinomialNB | 0 | 0.706845 | 0.028274 |
| SGDClassifier | 0 | 0.696429 | 0.14881 |
| LinearSVC | 0 | 0.65625 | 0.049107 |
| AdaBoostClassifier | 0.01 | 0.643601 | 0.008185 |
| RidgeClassifier | 0 | 0.642857 | 0.047619 |
| RandomForestClassifier | 0.01 | 0.62872 | 0.042411 |
| DecisionTreeClassifier | 0 | 0.616071 | 0.03869 |
| CatBoostClassifier | 0.29 | 0.592262 | 0.056548 |
| BaggingClassifier | 0 | 0.590774 | 0.004464 |
| XGBClassifier | 0 | 0.555804 | 0.073661 |
| ExtraTreeClassifier | 0 | 0.528274 | 0.052083 |
| LGBMClassifier | 0.01 | 0.523065 | 0.02753 |
| KNeighborsClassifier | 0.01 | 0.504464 | 0.081845 |
| BernoulliNB | 0 | 0.342262 | 0.104167 |

Table Model results, Supervised ML

Table 13 shows a list of the 14 models ranked by ROC AUC score. MultinomialNB was ranked highest with a score of 0.71 followed closely by SGDClassifier with a score of 0.696. All of the results are generally quite poor to fair in the range of 0 to 0.7. Further hyperparameter tuning might improve a model result above 0.8. The final model accuracy was estimated at 0.87, with a precision of value 0, a recall of value 0.0 and an ROC AUC value of 0.5.

Accuracy: 0.8666666666666667

Precision: 0.0

Recall: 0.0

ROC/AUC: 0.5

[How to Tune the Number and Size of Decision Trees with XGBoost in Python - MachineLearningMastery.com](https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/)

## Unsupervised

On the basis of analysis from pipeline of k means clustering,

* Clusters 1 and 3 have high punctuation counts
* Cluster 0 has a higher number of likes compare to cluster 1,2 and 3
* Clusters 1 and 3 have low level of detail
* Cluster 0 and 2 have a low character count
* Cluster 3 has a high character count but low level of detail
* Sentiment and star rating is equal across the board

Clusters 1 have 3 have more flags for inauthenticity than clusters 0 and 4 overall.

# Conclusion

The paper discusses how sentiment analysis facilitates online buying power, aided by machine learning techniques. Sentiment analysis enables businesses to understand customers’ aggregate opinions and attitudes towards certain products by distinguishing the polarity of the reviews, and helps customers make the correct and informed decision by supporting their research. according to the word cloud for positive and negative words in reviews. The paper gives some clear suggestions for tackling with the most frequently appeared complaints - the staff manner and the price of the meals based on data submitted by real customers.

The data was sourced from a combination of google reviews for restaurants in Galway and Dublin Ireland between February 2016 and April 2023. The exploratory data analysis shows a large and varied dataset with ratings between 2.8 to 4.0 stars and and sentiment polarity scores of -0.9 and +0.99. This indicates a range of satisfied and unsatisfied customers. Different types of restaurants are included ranging from upmarket steakhouses such as Fire Steakhouse & Bar to pub gastronomy such as Flanagan’s Restaurant & Bar. The Ivy in Dublin is a stylish brasserie while the Fox is a contemporary irish restaurant with a French influence offering an attractive vegetarian menu. The Quay Street kitchen is a small bustling restaurant serving up a range of cuisines in the heart of Galway city.

An in-depth literature review identified potential features which can help tell an authentic review from an inauthentic one such as the review length, the sentiment score, the helpfulness of the review, verification, coherence and readability. These features were taken into account for the rule-based classification design.

There are many types of inauthentic reviews, they range from computer generated to human generated and can be either positive or negative depending on the creator’s intention. The intention of this report is solely to identify methods to improve the accuracy of categorizing these reviews into 1 of 2 boxes; authentic or inauthentic. Several different analytical methods have been employed to accomplish this aim such as sentiment analysis, rule-based classification and machine learning.

This thesis is primarily focused on the actual review text column of the dataset. Further analysis extracted more quantitative information from this column such as the character and punctuation count, the sentiment polarity, the helpfulness score of the review. Other factors such as the number of likes a review had and its star rating were documented and visualized but not taken further.

The rule-based classification was judged to be the more successful approach however since this was completed with unlabelled data (regarding whether or not the review is authentic or not), there is no performance statistic to compare the accuracy. Overall, 1067 reviews passed the authenticity check, that it, the review met less than half of the criteria for inauthenticity. 261 reviews met at least 4 of the checks for inauthenticity. These rules can easily be adjusted to be made stricter, such as tightening the acceptable sentiment limits or reducing the limits for character count and punctuation. Additional rules can also be added easily to test for other characteristics of inauthentic reviews.

The supervised machine learning method returned a final accuracy of 0.87 which is a good result for predicting the authenticity of a review against the rule-based classification method. The unsupervised method pointed towards clusters which could then be broken down further by identify author id’s and used to identify problem areas.

Future improvements could involve expanding the rule-based classification such as testing for sarcasm, or going into more detail on the level of helpfulness in a review. The algorithms could also be tested on unknown datasets or compare the performance against a labelled dataset. More machine learning could also be undertaken with different models and further hyperparameter tuning.

# References

1. CNET. (n.d.). *How to spot fake Amazon reviews*. [online] Available at: <https://www.cnet.com/deals/spot-fake-reviews-amazon-best-buy-walmart/>.
2. Khalifah (n.d.). *Fakespot | Analyze and identify fake reviews and counterfeits*. [online] www.fakespot.com. Available at: <https://www.fakespot.com/>.
3. Federal Trade Commission (2021). *FTC Puts Hundreds of Businesses on Notice about Fake Reviews and Other Misleading Endorsements*. [online] Federal Trade Commission. Available at: <https://www.ftc.gov/news-events/news/press-releases/2021/10/ftc-puts-hundreds-businesses-notice-about-fake-reviews-other-misleading-endorsements>.
4. Gain, V. (2022). *Amazon sues 10,000 Facebook groups over fake reviews*. [online] Silicon Republic. Available at: <https://www.siliconrepublic.com/business/amazon-fake-reviews-facebook-groups>.