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

Author: Teresa Quain

Student ID:  [sba22235](mailto:sba22235@student.cct.ie)

Student email: [sba22235@student.cct.ie](mailto:sba22235@student.cct.ie)

Module: Research & Professional Ethics

Msc Thesis, MSc in Data Analytics

Wordcount:





Research title and Topic Area Introduction

# Introduction

## Problem Statement

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.

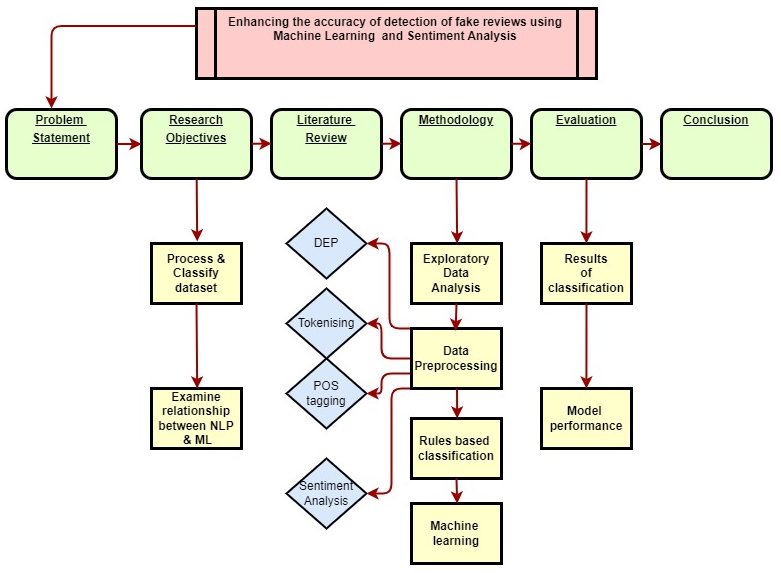
Cite

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

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



The structure of this thesis will follow fig x. 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.
* 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.

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.

# 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 Bayesian 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 x. below.

(insert methodology flowchart)

Algorithm: Data Preprocessing Steps

***Step 1: Import Libraries***

***Import pandas as pd***

***Import outscraper library***

***Step 2: Load Dataset***

***datafile = outscraper.load\_dataset()***

***dataset = pd.DataFrame(datafile)***

***Step 3: Drop Unnecessary Columns***

***Define columns\_to\_drop as a list of unnecessary column names***

***For each column in columns\_to\_drop***

***Drop the column from the dataset in place***

***Step 4: Check for Missing Values***

***Compute the sum of missing values for each column in the dataset***

***Store the missing values count in the missing\_values variable***

***Step 5: Check Dataset Shape***

***Get the number of rows (num\_rows) and columns (num\_columns) in the dataset***

***Step 6: Check Data Types***

***Get the data types of each column in the dataset***

***Store the data types in the data\_types variable***

***Step 7: Display Results***

***Print "Missing Values:"***

***For each column in missing\_values***

***Print column name and its corresponding missing value count***

***Print "Dataset Shape:"***

***Print "Rows:" followed by num\_rows and "Columns:" followed by num\_columns***

***Print "Data Types:"***

***For each column in data\_types***

***Print column name and its corresponding data type***

Libraries

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

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

## Dataset

The chosen dataset was sourced form outscraper.com . It consists of 25 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 variables, while the rating, the owner answer, the number of reviews likes and the actual review text are dependent variables. 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.

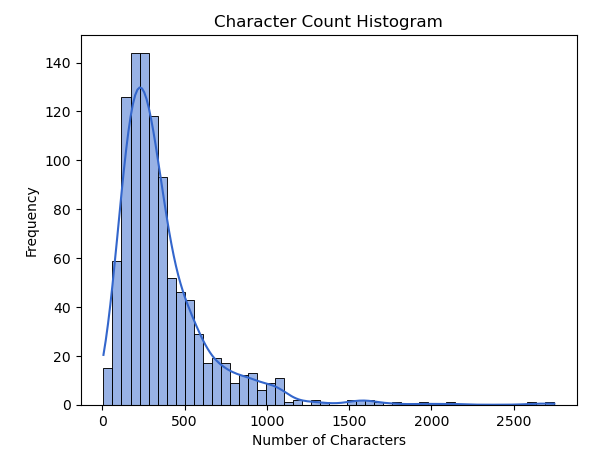
## Exploratory data analysis

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.

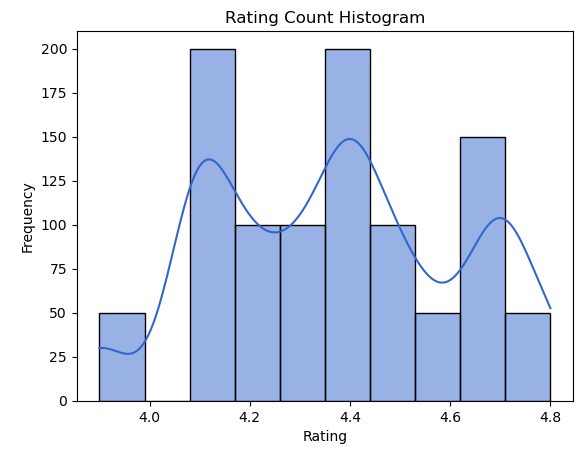
A diagram of data processing process

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.



Character count per review ranged from 0 to 2500 characters with a distribution skewed postivly to the right. 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 postive and to the point with no issues highlighted. Further analysis is required to interpret the meaning of this graph.



The above histogram shows the distribution of ratings in the dataset. These ratings are chosen by each individual reviewer based on their inputing a figure between 0 to 5 in their review. The range in this dataset is very high, with all almost all ratings between 4 and 5. 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.

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.

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.

A function ‘count\_puntuation’ was written to calculate the given number or punctuation marks per review, which can be valuable to interpret the writing style and tone of the review in a quantitative manner as seen in the algorithm below.

* ***Inputs: DataFrame df with 'review\_text' and 'punctuation\_count' columns***
  + ***Step 1: Add New Column 'check 4'***
  + ***For each row in df***
    - ***Set 'check 4' value to 0***
    - ***Step 2: Iterate Through Rows***
    - ***For each index, row in df***
      * ***If row's 'punctuation\_count' > 10***
      * ***Set 'check 4' value to 1 for that row***
      * ***Step 3: End***

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. 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 was then fed into the model as an 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.

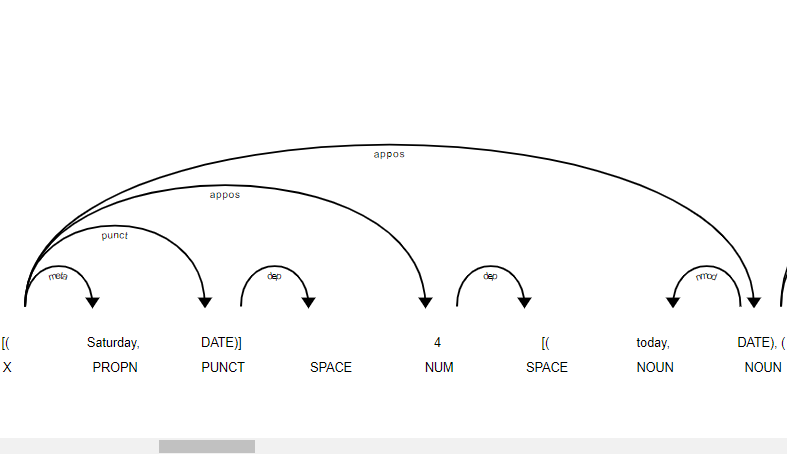
The spaCY package was used to classify the column ‘review\_text’ into named entities and their labels such as locations, cardinal numbers, noun’s and dates. The displacy visualizer allows a sentence to be broken up and it’s dependencies to be examined. This package was useful is analyzing the structure 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’.

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

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

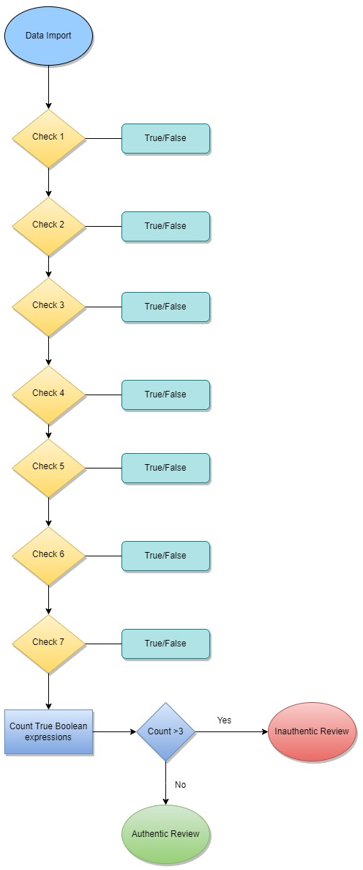
Vadar compound shows a mean sentiment of 0.78, a min of -0.97 and a max of 0.99.

Analyzing the vadar sentiment results showed 930 positive reviews, 62 negative reviews and 8 neutral reviews . Like the other visualisations, 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.

**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. 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.



* ‘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.

The final check takes the sum of the Boolean results of checks 1 to 7 and tests if the integer is greater than 3, that is, that the review has succeeded positively in 3 of the 7 tests at least. If yes, the review is labelled ‘fake’, otherwise it is labelled ‘true’.

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

**check1 = False**

**check2 = False**

**check3 = False**

**check4 = False**

**check5 = False**

**check6 = False**

**check7 = False**

**# Perform checks and store results**

**check1 = perform\_check\_1()**

**check2 = perform\_check\_2()**

**check3 = perform\_check\_3()**

**check4 = perform\_check\_4()**

**check5 = perform\_check\_5()**

**check6 = perform\_check\_6()**

**check7 = perform\_check\_7()**

**# Initialize count**

**count = 0**

**# Count Trues**

**if check1:**

**count += 1**

**if check2:**

**count += 1**

**if check3:**

**count += 1**

**if check4:**

**count += 1**

**if check5:**

**count += 1**

**if check6:**

**count += 1**

**if check7:**

**count += 1**

**# Check count and perform actions**

**if count > 3:**

**data\_import()**

**authentic\_review()**

**else:**

**inauthentic\_review()**