**D214 Task 2**

Jon Isak

D214: Capstone

Contents

[Data Collection 4](#_in3nqrrrgrzu)

[Data Extraction and Preparation 4](#_82xiarrk83ec)

[Analysis 6](#_v6451p7jl5ji)

[Data Summary and Implications 7](#_48i9vrkg5tja)

**D214 Task 2**

**Research Question**

A. The escalation of fake reviews online has become a pressing issue, impacting both consumers and businesses. This study's focal point is to address this concern by constructing a logistic regression model capable of accurately distinguishing between fake and real reviews within a dataset consisting of 20,000 entries for each category. The significance of this research question lies in the adverse effects of deceptive reviews, misleading consumers, and tarnishing businesses' reputations, while also disrupting fair competition. Detecting fake reviews manually poses significant challenges, necessitating the development of an automated identification system due to the potential harm they cause [(BrewTech Marketing, 2023)](https://www.zotero.org/google-docs/?e5rNLb). Leveraging prior research on fake news detection at the University of Michigan provides a foundational basis for this study, offering an opportunity to build upon existing methodologies [(Linkedin, n.d.)](https://www.zotero.org/google-docs/?J37ame). In today's digital landscape, online reviews significantly influence consumer decisions regarding products, services, and experiences. However, the prevalence of fake reviews has led to an atmosphere of distrust among consumers, exacerbated by competitors or malicious entities seeking to exploit these reviews for their own benefit [(Pike & Lustig, n.d.)](https://www.zotero.org/google-docs/?63kTXJ). Thus, the creation of a robust automated system to distinguish between authentic and fake reviews is vital for maintaining transparency and trust in online platforms. The research hypothesis (H1) asserts that a logistic regression model can achieve an accuracy rate exceeding 70% in identifying fake reviews. This hypothesis is built upon logistic regression's potential capabilities, supported by the prior study. The primary objective is to utilize logistic regression to develop a model that effectively identifies distinctive patterns or features separating genuine reviews from deceptive reviews.

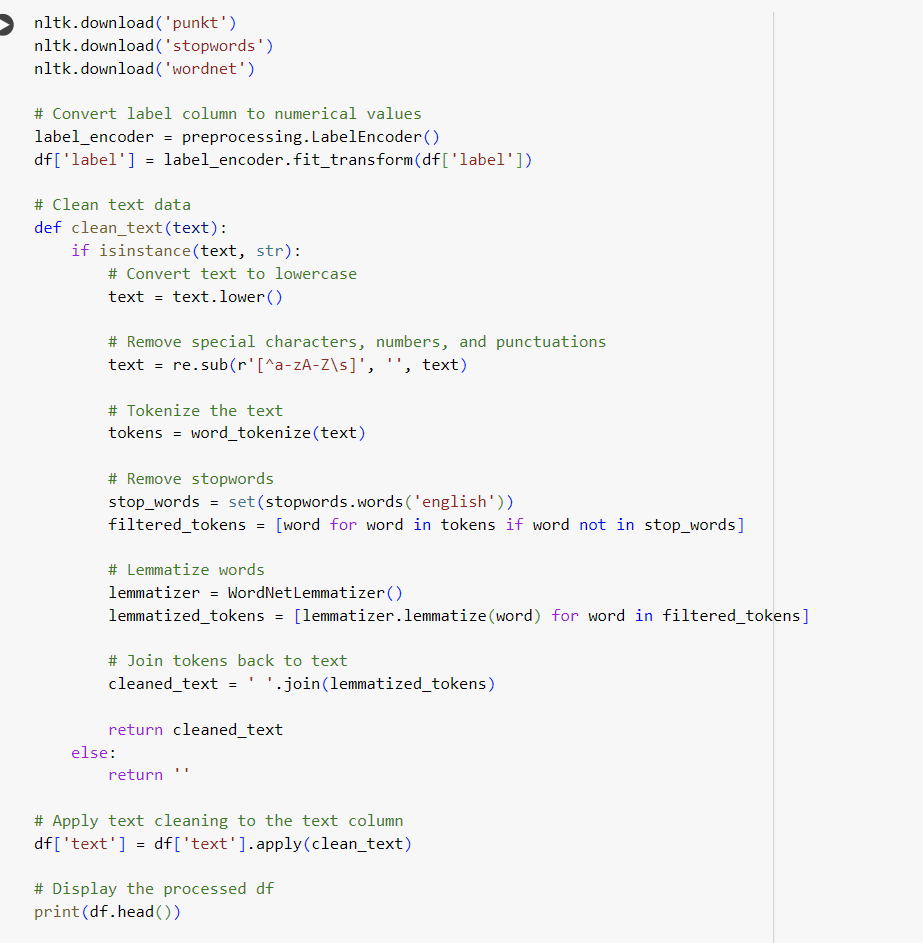
## Data Collection

B. The data collection process involved gathering relevant datasets necessary for addressing the research question of identifying fake reviews using logistic regression. The primary dataset sourced from Kaggle contained 40,433 rows comprising 20,000 real and 20,000 fake reviews, providing a substantial and balanced collection for analysis. One significant advantage of sourcing the dataset from Kaggle is the accessibility and diversity of data. Kaggle hosts a wide array of datasets contributed by the community, offering datasets across various domains and topics. While Kaggle provides a platform for various datasets, ensuring the quality and reliability of the data might be challenging. Datasets on Kaggle come from diverse sources, and there might be inconsistencies or errors within the data. In the case of fake review detection, inaccuracies in labeling reviews as fake or real could impact the model's performance. One challenge was that the text data was not clean and ready for modelling. To overcome this challenge, NLP libraries such as NLTK, were employed to efficiently clean and preprocess the text data.

## Data Extraction and Preparation

C. Pandas was used to load the CSV files into a Jupyter notebook. Pandas was used because it can read a variety of file formats into data frames and allows for data manipulation. For data preparation, the natural language toolkit libraries were used for text preprocessing, including removing special characters, punctuation, stop words, lemmatization, and tokenization. NLTK offers comprehensive text preprocessing capabilities, enabling efficient cleaning and transformation of text data. NLTK's performance might degrade when dealing with large volumes of text data, potentially leading to slower processing times for large datasets [(Pressman, 2017)](https://www.zotero.org/google-docs/?6n0i1C).





## Analysis

D. In the exploratory phase, the fake reviews dataset was reviewed to understand its structure, distributions, and generate statistical summaries, aiding in comprehending its characteristics. Text preprocessing using NLTK was subsequently employed to ready the data for analysis and modeling. The text review column underwent conversion to lowercase, removal of special characters, punctuation, stop words, and lemmatization of words. This technique possesses the advantage of cleaning the text data, making it suitable for modeling. However, excessive cleaning might result in information loss and the elimination of relevant features. Feature extraction was performed using TfidfVectorizer from scikit-learn. This process converted the preprocessed text data into numerical features, signifying the significance of words in distinguishing between fake and real reviews. Nonetheless, a drawback of this technique is its potential failure to capture relationships between words. Subsequently, the data was divided into training and testing sets, and logistic regression was employed as the classification algorithm due to its effectiveness in binary classification tasks. However, one of its disadvantages is its limitation in capturing non-linear relationships within the data. The metrics were then analyzed. On the original dataset, the logistic regression model achieved a high accuracy of 87% in discerning between fake and real reviews. However, for validation, when the model was applied to an external dataset, its accuracy dropped to 50%, indicating a substantial decrease in performance when generalized to a different dataset.

## Data Summary and Implications

E. The data analysis conducted aimed to address the research question of identifying fake reviews using logistic regression. This analysis utilized a dataset comprising both real and fake reviews from diverse product categories. While the logistic regression model achieved a high accuracy of 86% on the original dataset, its accuracy dropped to 50% when applied to the external dataset. Therefore, based on the statistical evaluation, I fail to reject the null hypothesis. This decline in accuracy when evaluating the model on an external dataset highlights a limitation in its ability to generalize across different review domains or contexts. The model may struggle to adapt well to diverse types of reviews, indicating potential issues with dataset variations. Despite attaining high accuracy within the original dataset, the model's diminished performance on an external dataset suggests a need for further investigation and refinement. Prioritizing additional analysis and model refinement to enhance the model's ability to generalize across diverse review datasets is important. This could involve retraining the model with more diverse data or employing domain adaptation techniques. Furthermore, developing domain-specific models trained on particular categories of reviews may enhance performance within specific domains by capturing domain-specific patterns and nuances in fake reviews. Exploring ensemble learning methods or transfer learning techniques could further enhance the dataset analysis, leveraging knowledge from the original dataset to improve the model's performance on new datasets. Employing ensemble models or transfer learning approaches could aid in utilizing information learned from the original dataset to adapt and perform better on diverse review datasets.

F.

[BrewTech Marketing. (2023, September 27). *Unmasking the Dark Side: Why Are Fake Reviews Thriving, and What Is Their Impact?* https://www.linkedin.com/pulse/unmasking-dark-side-why-fake-reviews-thriving-what/](https://www.zotero.org/google-docs/?ys6xDv)

[Linkedin. (n.d.). *What machine learning techniques can you use to detect fake reviews?* Retrieved January 2, 2024, from https://www.linkedin.com/advice/1/what-machine-learning-techniques-can-you](https://www.zotero.org/google-docs/?ys6xDv)

[Pike, M., & Lustig, D. (n.d.). *Your Competitors Can Sue You For Fake and False Product Reviews*. Pike & Lustig, LLP. Retrieved January 2, 2024, from https://www.turnpikelaw.com/your-competitors-can-sue-you-for-fake-and-false-product-reviews/](https://www.zotero.org/google-docs/?ys6xDv)

[Pressman, L. (2017, March 15). *The Python NLP Ecosystem: A Short and Very Opinionated Guide |*. Automated Insights. https://automatedinsights.com/blog/the-python-nlp-ccosystem-a-short-and-very-opinionated-guide/](https://www.zotero.org/google-docs/?ys6xDv)