Text & Web Analytics: Project Proposal

Syndicate Group 2

Introduction

To provide valuable products and services, it is critical for companies to understand the needs of their customer base. With the advent of e-commerce stores, online reviews have become an important indicator of product quality. According to the Australian Competition & Consumer Commission, many consumers rely on online reviews to inform their judgement about whether or not to make a purchase.[1] As such, reviews can provide valuable signals about the strengths and weaknesses of specific products.

Extracting information from customer reviews can be a major challenge for online retailers. Since reviews are typically formatted as free-form text, it can be difficult to efficiently identify common issues across product lines. This is especially problematic for large e-commerce stores such as Amazon, where individual products may receive many thousands of reviews. While star ratings can provide a rough measure of consumer attitudes, more advanced techniques are required to derive useful data from reviews.

For our syndicate project, we propose to apply sentiment analysis to a set of customer reviews from an online clothing store. Sentiment analysis is a type of natural language processing (NLP), where text is analysed to classify the opinion or attitude of the author. With respect to online reviews, sentiment analysis can be used to measure customer satisfaction with a given product. As such, NLP techniques can provide valuable inputs to both product design and market research.

Task Specification

Our core dataset is a collection of 23,486 customer reviews that were scraped from an online women's clothing store in 2016.[2] The dataset contains 10 variable fields, which are summarised in table 1. Amongst these features, there are three dependent variables (DVs) of interest:

1. The **rating score**, which is a measure of the customer's attitude towards the product.
2. The **recommendation binary variable**, which indicates whether or not a customer would recommend the product to others.
3. The **positive feedback count**, which measures the degree of positive feedback or 'upvotes' that a review receives from other consumers.

---

[1] ACCC, 2013. *What You Need to Know About: Online Reviews – A Guide for Business and Review Platforms*.

[2] The dataset is publicly available at https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Table 1: Variables from Kaggle dataset

Variable	Type	Description
Clothing ID	Integer	Unique ID of the product
Age	Integer	Age of the reviewer
Title	String	Title of the review
Review Text	String	Body text of the review
Rating	Integer	Star rating awarded to the product by the customer, ranging from 1 (worst) to 5 (best)
Recommendation IND	Binary	Dummy variable indicating whether the customer will recommend the product, where 1 = recommended and 0 = not recommended
Positive Feedback Count	Integer	Units of positive feedback or 'upvotes' that the review has received by other shoppers
Division Name	String	Name of the division that the product is in (e.g. *General, Intimates*)
Department Name	String	Name of the department that the product is in (e.g. *Tops, Dresses*)
Class Name	String	Type of product (e.g. *Skirts, Sweaters, Pants*)

The rating and recommendation variables are both measures of review **sentiment**. As a result, they are likely to be highly correlated. If a reviewer awards a high rating to a product, it is likely that they will also recommend the product to others. In contrast, the positive feedback count is a measure of review **utility** — that is, it can be viewed as a measure of whether or not other customers find the review helpful for their own purchase decision. Accordingly, our task will have two main components:

1. **Sentiment:** explore the features of the review corpus to identify textual elements that are associated with positive and negative reviews. Use this data to build a classification model that can categorise reviews according to both rating and recommendation.
2. **Utility:** explore how word choice, semantics and syntax affect whether the review is trusted by other customers, as measured by the positive feedback count.

Methods

In order to predict customer satisfaction and utility from the customer reviews, we will adopt a similar approach to past literature in this field.[3] This approach involves three main steps, which are detailed below.

Step 1: Pre-Processing

Firstly, the review data needs to be pre-processed into a suitable format for sentiment analysis. This pre-processing may include the following steps:

- Removing non-alpha characters
- Case folding
- Removing punctuation & special symbols
- Tokenization
- Removing stop words

Step 2: Model Development

After pre-processing the review data, we will use NLP techniques to extract lexical and syntactic features from the review text. This may include:

- **Token frequency:** generate a 'bag of words' for each review and calculate the TF*IDF frequency of each token.
- **Parts of speech:** extract and classify the parts of speech (such as verbs and adjectives) from the review text.

Using these textual features, there are at least two possible methods of predicting customer satisfaction and utility:

1. **Regression:** use the textual features as predictors in a regression model.
    - To predict rating (on a scale of 1 to 5), we could use an ordinal logistic regression. This type of model would produce a latent score as a function of the textual features in each review text. This score could then be converted to a star rating based on threshold values.
    - To predict positive feedback count, we could use a Poisson regression or a Tobit regression (where the DV is bounded below at zero).
2. **Classification:** alternatively, we can use an ensemble of different classification methods such as Naïve Bayes, Support Vector Machines or Random Forests to classify each review according to rating level or positive feedback count.

---

[3] Liu, D., Chai, Y., Zheng, C. and Zhang, Y., 2017. *Rating Prediction Based on TripAdvisor Reviews*. URL: https://pdfs.semanticscholar.org/51b7/a2cdfd350b0d5d8e3fd71e2941d495fda2b8.pdf

Step 3: Model Selection & Evaluation

Sample data will be split into *training*, *validation* and *test* sets. The test data will comprise approximately 20% of the whole dataset. Our models will be constructed using the training data and the test data will be used to evaluate the predictive performance of each model.

For example, suppose that our DVs are coded as binary variables (e.g. *positive* or *negative* sentiment). In this case, we can evaluate model predictions using the following steps:

1. **Prediction:** Predict the value of the DV for each test observation, using threshold values to categorise classes where required.
2. **Evaluation:** Construct a confusion matrix to compare the predicted classes to the true classes in the test data. Evaluate the model performance using an appropriate measure, such as:
    a. **Specificity (True Negative Rate):** the proportion of actual negatives that the model correctly identifies as negative. For example, suppose that we are interested in identifying products that receive poor reviews or fail to be recommended. In this case, we would prioritise models with a high specificity rate so that we can identify as many negative cases as possible.
    b. **Recall (True Positive Rate):** the proportion of actual positives that the model correctly identifies as positive. For example, suppose that we are interested in identifying the textual characteristics of useful reviews. In this case, we would prioritise models with a high recall rate so that we can identify as many useful reviews as possible.
3. **ROC curve:** Generate a ROC curve by plotting true positive rate against false positive rate as a function of different cutoff thresholds. Calculate the area under the curve (AUC) for each model. The AUC is a measure of the model's capacity to distinguish positive and negative cases, where better models have an AUC score that is closer to 1.