AY 2019/20 Semester 1

Nanyang Technological University

H6751 - Text And Web Mining

Group Assignment Proposal

Team Members

1.	Huang Qifan	MSc. Information Systems, Year 1	qhuang010@e.ntu.edu.sg
2.	Loh Zi Bin Robin	MSc. Information Systems, Year 1	zloh012@e.ntu.edu.sg
3.	Oh Guo Wei	MSc. Information Systems, Year 1	goh096@e.ntu.edu.sg

Project Motivation

Online shopping has become an integral part of our everyday life, largely due to convenience where the product will be delivered to your doorstep with a click of a button. To enhance business experience and to understand consumer needs, most of the online shopping platforms provide consumers an avenue to feedback on the overall shopping process.

Analysing these feedback may yield valuable and interesting insights, which ranges from identifying the fashion trends, to understanding the clothing preferences from the various age groups. These insights are important especially on influencing and predicting customer's future purchasing decisions.

Project Objectives

We aim to build a training model so that from a customer's review, we can predict which customer's age group he/she belongs to. The age groups will be categorised based on a 5 year period (e.g. 15 to 20 years old, 21 to 25 years old).

In this project, we will be performing sentiment analysis on customer reviews (from a women's clothing e-commerce site) to review hidden information from a large set of reviews. During this process, we will identify the top prominent terms mentioned by the various age groups. We foresee that the clothing types might become one of those prominent terms, so we intend to find associations of clothing types with other terms, which might be useful to predict the customer's age group better.

Dataset Source

The dataset source is from Kaggle - Women's E-Commerce Clothing Reviews by nicapotato on 3 February 2018.

Although there are 10 feature labels in the dataset, we will limit the scope of this group assignment to analysing these features so that we can apply various text mining concepts extensively.

- Customer reviews
- Age
- Clothing Type

The APA citation for the dataset source is as follows.

Nicapotato. (2018, February 3). Women's E-Commerce Clothing Reviews. Retrieved from https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/home

Approaches and Methodologies

Text mining approaches will be applied to the customer reviews, such as (but not limited to) text extraction and pre-processing, frequency text analysis and feature extractions. Then, we will apply the classification models such as (but not limited to) logistic regression and random forest to analyse the results. Finally, we will validate the prediction performance of our training model.

The 6 steps below will be applied to achieve our project objectives.

Step 1 - Data Collection and Preparation

Using the dataset from Kaggle, it will be divided into Training Set and Test Set using sampling techniques.

- To ensure that the Test Set is sufficiently large enough to obtain meaningful results, 80% of the original data will be placed in the Training Set, while the remaining 20% will be placed in the Test Set.

- Simple random sampling is suffice as the original data set is sufficiently large and the Test Set will be representative of the original data set.

The Training Set will be used mainly to train and tune the model, while the Test Set will be used to test the trained model.

Step 2 - Text Parsing

Using the Training Set, we will pre-process the text to:

- Tokenize words from sentences into its component parts
- Tokenize sentences using n-gram algorithm and/or binary classifier algorithm

Step 3 - Text Filtering

Using the Training Set, we will filter the text to:

- Filter irregular and irrelevant terms or lingos
- Customise the start and stop lists
- Minimise stop words in the sentences
- Standardise word formats using stemming and lemmatization techniques

Text filtering and parsing will minimise the data dimensions and improve the statistical significance of the data.

Step 4 - Data Transformation

The data attitudes in the Training Set will be inserted into a “Document-Term Matrix” (DTM) or “Term-Document Matrix” to understand the frequency of the terms than appear in each document.

To extract useful features in the Training Set, the following algorithms will be attempted:

- Singular Value Decomposition - To transform the original matrix into special matrices, so that the large matrix can be broken down into different, smaller components.

- Principal Component Analysis - To transform the data in the matrix (in a linear fashion) to new features which are not correlated to each other.

Step 5 - Building of Training Model

The Age column (in the original dataset) will be transformed into the Age Group column which follows a 5 year period. This Age Group column are the “correct outputs” (i.e. Y-values) which will be added into the Training Set.

Various machine learning algorithms will be used using the Training Set. This includes (but not restricted to):

- Classification and Regression Trees (CART)
- Logistic Regression
- Random Forest

Step 6 - Evaluation of Training Model

Using the Test Set, we will test the Training Model. The confidence levels for each algorithm can be compared for analysis, which will be elaborated further in the Machine Learning Results section.

Expected Results

1. Text Mining Results

Most and Least Frequent Terms Analysis

From the document term matrix, we will identify the least 20 and most 20 frequent terms, which measures the degree of importance in the data set. Terms which appear more frequently will generally be of interest for further analysis, while terms which rarely appear will not be of significant value to analyse further.

In order to minimise the sparsity of data, we will set a threshold of factor 0.99 so that we will analyse the terms which appear at least 2% of all the documents.

Word Cloud

To interpret the most prominent terms quickly, we will use the word cloud so that we can find out the terms which are useful for this project.

Correlation of Terms in the Document Term Matrix

Based on the clothing types, we will find out the correlation (between 0 to 1 inclusive) with the other terms in the Document Term Matrix. The correlation values will be represented in a correlation matrix. Higher correlation value implies that the 2 terms are more associated with each other.

2. Machine Learning Results

The performances of all the machine learning algorithms will be evaluated using the measures below:

- Accuracy - Overall correctness of the machine learning algorithm.
- Precision - The correctness of the prediction when the model predicts positive / true.
- Recall - True positive rate.
- F1 Score - A weighted average of the precision and recall.

The formula to calculate each measure is as follows:

$$Accuracy \ = \ \frac{True\ Positives\ +\ True\ Negatives}{Total\ Population}$$

$$Precision \ = \ \frac{True\ Positive}{True\ Positive\ \ +\ \ False\ Positive}$$

$$Recall \ = \ \frac{True\ Positive}{True\ Positive\ \ +\ \ False\ Negative}$$

$$F1\ Score \ = \ \frac{2\ \times\ Recall\ \times\ Precision}{Recall\ +\ Precision}$$

Conclusion

Through this project, we aim to explore the various text mining techniques to interpret hidden results in the customers' reviews. The purpose of this project is not to obtain decent confidence values in our machine learning results, but rather to explore the text relationships between the age groups based on the customer's reviews. We recognise that this project does not cover every aspect of text mining, and we hope to use this experience to get more in depth with our analyses on future projects.

End of Document