BT5153 Topics in Business Analytics

Project Proposal on

Text Mining and Reply Generation of Migraine App Reviews

Group 12

Supervisor: Dr Zhao Rui

Team Member:

Tang Han (A0176586L)

Zheng Yiran (A0176608W)

Veronica Hu He (A0057037H)

Derek Li Lingling (A0176652X)

Sophia Yue Qi Hui (A0176615Y)

Jason Chew Guo Jie (A0176614B)

Contents

1. Objective

Migraine Buddy is an advanced migraine headache diary and tracking application (APP) designed with neurologists and data scientists to assist patients to track and better understand their migraine-related patterns. In this project, we will mainly collect data through scraping and analyze the user reviews with the respective responses from Migraine APP in Google Play to gain insights on customer's key concerns and top reasons on giving positive/negative reviews through text mining techniques, such as sentiment analysis and topic modeling. Additionally, we also aim to automate the generating of relevant replies to users' review pertaining to the APP using machine learning to enhance the customer experiences and cut-down the manpower deployed in replying customer's feedback.

2. Background

Excellent customer service has been a pillar of business success for years. 89% of people surveyed (Oracle, 2012) indicated that they switched to use a competitor's services after a poor customer experience was encountered. It is essential to business to obtain honest customer feedback when looking in improving customer's experience. However, it is not easy to capture customers' feelings as 91% of customers don't provide feedback or complain when they are unsatisfied with the product or service, according to the study conducted by Oracle. Customers think that it is not worth the time to provide feedback because the business simply doesn't care. Nevertheless, 81% of customers claimed the willingness to provide the feedback when they knew that there would be an immediate response from the companies. In summary, responding to customer's feedback or review promptly plays an important role in improving customer service and preventing of customer churn.

As most of the customer feedbacks are captured in textual form, text mining and natural language processing (NLP) techniques are highly relevant. Currently, there is a spectrum of methods and techniques for feature extraction, ranging from the simple statistical methods, such as bag-of-words model like the term frequency - inverse document frequency (TF-IDF), to the more complex models that apply neural networks (John, 2017).Increasingly, neural networks models are being developed and applied in natural language processing because of their ability to better identify obscure patterns that are inherent in language, whereas recurrent neural network (RNN) models are said to be more superior at language modeling tasks such as response generation (Goldberg, 2016).

3. Dataset

3.1 Data Source

Google Play is a digital platform which provides digital distribution service for Android operating system users. APP users are eligible to write their reviews and communicate with APP owners through Google Play Store. Reviews from APP users and replies from APP owners can be scraped from Google Play website.

3.2 Data Collection

The whole webpage is written in html. By exploring the website using inspect tool, we discovered that the information belongs to the same category has a similar nested structure and a same class name.

We plan to apply Beautiful Soup to scrape from the website directly, and according to the tags to retrieve the needed information. Meanwhile, we will still face some challenges, such as wrapped reviews, emojis, auto-extend scroll webpage and multiple language leading to decoding error.

3.3 Data Description

From the website, the main content we plan to scrape including the following 6 variables:

S/N	Variable	Type	Descriptions
1	User name	String	The name of user who provided the review
2	Review date	Date	The date when the review was provided
3	Review content	String	The content of the review
4	Number of stars	Numeric	The ranking of the APP which user indicated in measuring the quality of services provided by the APP, from 1 star (lowest) to 5 stars (highest).
5	Reply date	Date	The date when the company replied to the particular review
6	Reply content	String	The content of the response from the company

In total, there are 30,277 reviews and 97% of them have the replies from the company.

4. Project Methodology

4.1 Text Preprocessing

4.1.1 Language Filtering

As the Migraine APP is available for users from all over the world, the review messages could also come in different languages. **Langdetect**, a robust library ported from Google's language-detection will be used to filter all comments in English. The rest non-English reviews will be excluded from our study, which are estimated to be around 20% of the total data.

4.1.2 Text Cleaning

We will adopt different levels of text cleaning process to address the different goals of our project: insight information extraction and reply generation.

For insight information extraction, a deeper level text cleaning will be performed, in order to normalize the text as much as possible and eliminate the interference from non-standard words/symbols and derivatives. The cleaning techniques include:

- Lower case transformation
- Punctuation removal
- Stop words or frequent words removal

- Rare words removal
- Spelling correction
- Tokenization
- Stemming or lemmatization

For reply generation, a much shallower text cleaning will be applied, in order to generate a more natural sentence in the reply. Most components from the gathered text, including stopwords and punctuations will be kept as input to the NLP model. The cleaning process will just do spelling correction and removal of some irregular symbols (such as extra spaces), which is mainly focusing on the error amendment.

4.2 Review Insights

Reviews will provide valuable feedback to different aspects of the APP and give insights for future improvement. To gain insights from the reviews, we aim to conduct sentiment analysis of the reviews and identify the topics attracted most attention in positive reviews and negative reviews accordingly. With the help of our study, the APP developer can strengthen the most mentioned topics in positive reviews and improve the areas picked out in negative reviews.

To enable detailed analysis, we will firstly generate a library of the words appeared in all reviews (Bag of Words) by N-grams approach where N = 1 and 2 respectively. The number of times of a word (or two consecutive words, 2-gram) appearing in a review will be computed (Term Frequency, TF) and normalized according to its overall frequency in all reviews (TF- Inverse Document Frequency, TF-IDF).

4.2.1 Sentiment Analysis

Although lower review score (1 star or 2 stars) will probably indicate a strong negative sentiment, the vast majority of 5 stars cannot help us to identify the reviews with most positive sentiment. Thus, we will use existing sentiment dictionaries to assist in this task.

Two dictionaries, namely SenticNet-4 and Lexicoder Sentiment Dictionary (2015) will be used to assess the positive/negative level of the reviews.

SenticNet is a well-known organization specialized in concept-level sentiment analysis. SenticNet-4 is one of the latest sentiment dictionaries published by them. It introduces the concept of semantic primitives and further extends the knowledge base to 50,000 entries. Each entry has a decimal number between -1 to 1 to indicate the strength of positivity/negativity of the word. The reviews will be broken up into individual word (or 2-gram) and searched within dictionary for matches. The ratings of matched entries will be aggregated to derive the sentiment score of each review.

Lexicoder Sentiment Dictionary (LSD2015) consists of 2858 negative sentiment words and 1709 positive sentiment words. Compared to traditional sentiment dictionary, LSD2015 also contains 1,721 word-patterns indicating a positive word preceded by a negation (used to convey negative

sentiment) and 2,860 word-patterns indicating a negative word preceded by a negation (used to convey positive sentiment). To use this dictionary, the number of four types of words in each review will be counted. The sentiment of review will be computed by the following formula. If overall count is positive, the review is positive, and the review is negative if overall count is negative.

Count = Positive + Negative Negative – Negative – Negative Positive.

4.2.2 Topic Modeling

We aim to identify 5 or 10 most popular topics among the positive and negative reviews respectively. Top 10 words in each topic will also be listed. The choice of 5 or 10 topics will depend on the perplexity score obtained from free package for Latent Dirichlet Allocation (LDA) model.

We will also try to assign weightage according to words in each review according to the review sentiment score so that the topics in more positive/negative reviews will have a higher weightage in the topic selection.

Number of reviews related to each topic and overall sentiment score of related reviews for each topic will also be computed and visualized in dashboard to make our report easier to understand to readers.

4.3 Reply Generation

Recurrent Neural Network (RNN) is a very robust tool in handling the sequential input and output due to its intrinsic structure. It is widely used in Natural Language Processing (NLP) application including text generation based on given input. Hence there is no doubt it is our best choice to develop reply generation function for Migraine APP.

The naïve RNN is poor in handling long sequence due to the gradient vanishing problem incurred from the long chain of input. In practical, most RNN used in NLP is bundled together with Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) to learn how to remember or forget the information carried from early beginning. The GRU unit controls the flow of information like LSTM unit, but without having to use a memory unit. In our project, we will try GRU first as in many scenarios it has on-par performance as LSTM, but more efficient in computational cost. If the outcome is not good enough then we would switch to LSTM or even deep bidirectional LSTM, which could understand the context better at the cost of more computational resource.

Besides the mathematical model, there's also an option of word-level RNN (learn and generate text word by word) against character-level RNN (learn and generate text character by character). We will use word-level RNN because in most cases it displays higher accuracy with lower computational cost than its counterpart, especially when techniques like LSTM are used. The character-level RNN is more popular in languages with a rich morphology such as Finish, Turkish and Russian. If the output from our model does not meet expectation due to insufficient training data, we could attempt other machine learning

technics (e.g. decision tree) to improve the performance. One possible solution is to create a library of 50 standard replies to different topics and let RNN assign replies to reviews.

For the implementation, we will build our test model using Tensorflow with Keras wrapper on Google Colaboratory platform, as it provides free GPU for 12 hours per session. The extensive model training will be conducted on National Supercomputing Centre Singapore (NSCC) platform or one of the commercial clouds (AWS or GCP).

5. Evaluation and Other Work

To determine the performance accuracy, we will use two methods of evaluation. Firstly, we will use bilingual evaluation understudy (BLEU), an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Here, BLEU is used to compare replies generated by our RNN model to the original replies in the validation/test dataset. BLEU score will help us to determine if the customer's comments have been properly responded.

Secondly, for each model, we will randomly pick 60 pairs of review and reply. Six group members will evaluate the appropriateness of replies respectively and an average score will be used as the final result. This human evaluation method will complement our BLEU score as an additional gauging measure.

Together, both measurements would work together to evaluate the quality of the customer comments interpretation accuracy and the quality of the response made to better meet the customer's needs

Bibliography

Oracle. (2012). *2011 Customer Experience Impact Report*. Retrieved from Oracle: http://www.oracle.com/us/products/applications/cust-exp-impact-report-epss-1560493.pdf

John, V. (2017). A Survey of Neural Network Techniques for Feature Extraction from Text. *arXiv preprint arXiv:1704.08531*.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.