BT5153 Topics in Business Analytics

Project Report on

Text Mining and Reply Generation of Migraine App Reviews

Group 12 - Accelerator

Supervisor: Dr Zhao Rui

Team Members:

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

Table of Figures

Table of Supplementary Figures

List of Tables

List of Supplementary Tables

1. Abstract

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 scraped and analyzed customers' review about the app on Google Play store and tried to find out insights of customer's key concerns and top reasons for giving positive/negative reviews. Multiple text mining techniques, such as sentiment analysis and topic modeling were used. Additionally, in order to enhance the customer experiences and cut-down the manpower deployed in replying customer's feedback, we also explored to automate the generation of relevant replies to users' review using machine learning and deep learning technology.

2. Introduction

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 a significant role in improving quality of service and preventing of customer churn. This simple truth is recognized by most companies, but inspecting and replying tens of thousands feedbacks manually could be very labour-intensive for small and medium enterprises.

In this project we studied customer review data for an Android app named Migraine Buddy which was developed by Healint. Healint is a Singapore-based medical data analytics startup company founded in 2013. Its vision is to build intuitive solutions that empower individuals to manage their chronic pain conditions. Its first product Migraine Buddy was launched in 2014, and immediately overtook the hundreds of apps for migraine and chronic pain to become the world's leading app of its kind. More than one million customers from all over the world have downloaded the app and thirty-one thousand have left their reviews. Currently most reviews were read and replied manually by Healint's customer support staff. By using text mining and natural language processing (NLP) technology, we wish to automate business insight extraction and reply generation from customer's textual review, henceforth relieve the support staff from tedious and repetitive work.

3. Related Work

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).

4. Dataset

4.1 Data Source and Collection

The Migraine Buddy is available on both Android and iOS. Since it has much more downloads on the former platform than the latter, we only focused on customers' reviews and replies on Google Play. Similar to many other modern websites, Google uses Asynchronous Javascript and XML (AJAX) on their webpage to optimize the bandwidth consumption and load of the server. It means the webpage will only display limited information at first loading. It requires users to scroll down to bottom of the webpage, or occasionally click the "Load More" button to retrieve more information. Such design prevents simple web scraping as the webpage only shows about 40 reviews and replies for the initial display.

To address this issue, we used Selenium together with Beautifulsoup to build the web crawler. Selenium is a very popular web automation tool which is normally used to test web application, especially the human interaction part. It can automatically emulate human's web surfing behaviors, such as scrolling and clicking. The crawler with Selenium can auto scroll down to the bottom of webpage every 10 seconds. When the "Load More" button appear, it can also detect the object and perform clicking by itself. It eventually automated more than 280 scrolling and clicking interaction with AJAX, and successfully retrieved 9206 reviews and replies from Google Play. All the texts were read and parsed by Beautifulsoup using html parser.

4.2 Data Description

The parsed information is consolidated into Pandas dataframe in following format:

S/N	Variable	Type	Description
1	Customer name	String	The name of user who provided the review
2	Ratings	String	Customer's rating about the app, from 1 star to 5 stars
3	Helpfuls	Integer	Number of customers who think this review is helpful
4	Reviews	String	Contents of customer's review
5	Replies	String	Replies from Healint Staff

4.3 Data Exploration

Language Filtering

As the Migraine APP is available for users from all over the world, the review messages could also come in different languages. To identify the languages in which the reviews were written in, Spacy's Language Detector was applied to label and classify these directly. This revealed that 96% of the reviews or 8852 reviews out of 9206 were written in English (Supplementary Figure 1). An interesting find was reviews with language label 'UNKNOWN' which contained only emojis (Supplementary Figure 2). For this project, only English reviews would be applied to the models.

Ratings

Most migraine app users found the app to be useful, and 82% gave the app 5 stars rating (Supplementary Figure 3). To get a brief understanding the users' reviews, 3 word clouds were created (unigram, bi-gram, tri-gram) with the tri-gram being the most useful in showing users' positive feedback for the app – "helps track migraines", "app easy use" etc. (Supplementary Figure 4).

Text Processing

We adopted different levels of text cleaning 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, tokenization and stemming.

Sentiment Analysis

The next step was to also understand the sentiment of user's reviews, this was done by applying Textblob package. From the boxplot of sentiment across the ratings, several outliers for 5 stars were identified which could be attributed to the context/ background of the app – many users were describing their symptoms – 'suffer', 'acute', 'pain constantly' which led to negative sentiment (Supplementary Figure 5). Nevertheless, sentiment score is seen as a useful added feature hence was included in the models.

Topic Modeling

Finally, topic modelling was carried out on both positive and negative sentiment reviews in which Latent Dirichlet Allocation from the sklearn package was applied. 5 topics with 10 most popular words were generated, for which the positive sentiment review did not indicate immediately distinct topics, whereas the topics were slightly more distinct for the negative sentiment review – data issues, account problem, login issues (Supplementary Figure 6). Although the groupings make sense, the idea was to automatically categorize the reviews into logical topics in which replies could be applied to each particular group. This would be further explored in detail below.

5. Reply Generation

5.1 Topics Matching Model

Assumptions

With the data collected, the underlining assumption for the topics matching model is that (1) the current replies that are done manually for each review are done so correctly and accurately (2) the types of reviews and the types of replies to these reviews can be generalized into few key areas of interest. Based on these, the model makes use of the idea that certain types of reviews are matched to certain types of replies. Therefore, by identifying the type of reviews of the grouping or reviews, the corresponding replies based on historical data can be automatically matched and generated.

Topics Modeling Optimization

For a practical implementation of the model, the numbers of topics within the reviews and replies needs to be determine automatically and optimized. Here, the coherence score is used to determine the optimal number of topics for each group. Coherence score is chosen here as the better the scoring would show that the topics are better interpreted by humans. Since these matching of topics is mainly for reviews to replies, interpretability seems to make sense as the judging criteria.

The optimization uses the Gensim package as a wrapper to run LDA Mallet. This builds the LDA model with different number of topics and selects number with higher coherence score as the optimal model. Looking the review coherence score graph, it can be seen that 5 topics are selected in this case. By selecting the number of topics at the end of a rapid growth, the topic coherence usually offers meaningful and interpretable topics. As for the optimal number of topics for the replies, 4 is selected even though 7 would have a higher coherence score. This is done as picking a higher number of topics can sometimes provide more granular sub-topics and over specifying the of topics and reply types.

It is worth mentioning that repeated runs of the topics optimization (without setting random states) does show that the optimal numbers rest in the range of 4-5 topics for both reviews and replies.

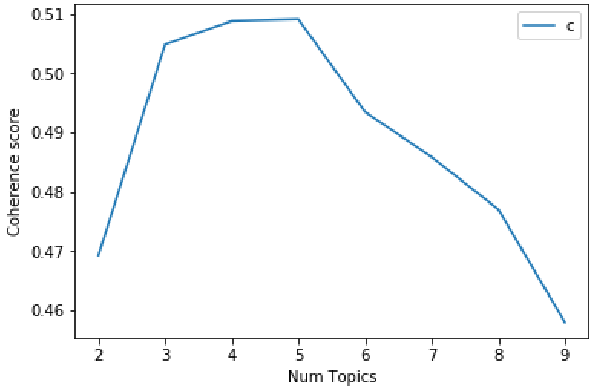


Figure 1: Reviews Coherence Score by Topics

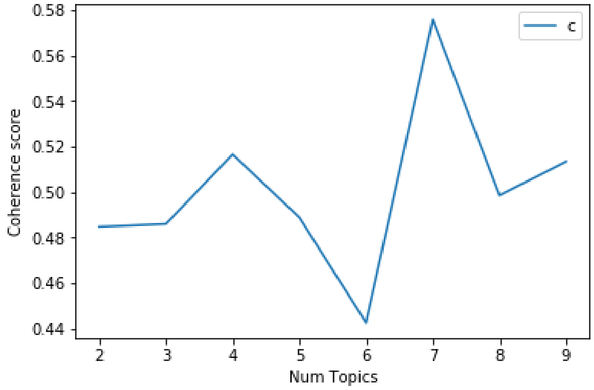


Figure 2: Replies Coherence Score by Topics

Model Generation

After the reviews and replies has been categorized into topics, additional features are added. Firstly, the baseline features from the data source such as "helpful" and "ratings" are added into the models. The sentiment score which as mentioned above was also used as a feature.

Considering the importance of key words that indicates the nature of the review in identifying the correct replies. The review text undergone tokenization to create features for each word. The TF-IDF method was selected to tokenize the text in order to pay more attention to rare key words that are more indicative as compared to comment words such as "great". The data sets are then split using the model selection package in sklearn with model generation based on training data sets and evaluation on test data sets to determine the accuracy.

Model Results

Setting a baseline performance, a dummy classifier was ran based on the most frequent reply topic as the answer. The baseline performance shows that the accuracy is performing at the mid 30% range.

Moving on, Logistic Regression and Random Forest was used for model generation. The model generated based on the additional features was able to achieve a 50% accuracy for both the logistic regression and random forest (Figure 3). More than 10% improvement made on top of the baseline model. It is also worth noting that, if the tokenized text features were removed, the accuracy would only reach 40% which is only slightly better than the dummy classifier. As for model generation, Naïve Bayes model was not used as the sentiment analysis would input negative parameter into the model, which Naïve Bayes cannot accept.

Improvements

With an accuracy of 50%, there still shows some significant gap to operate at an acceptable level of feasibility. Though this method allows for automated classification and replies, the issue with the performance is attributed to the degree of accuracy for topics modeling. In the next model, it explores the manual tagging of topics and with the further generation of additional features.

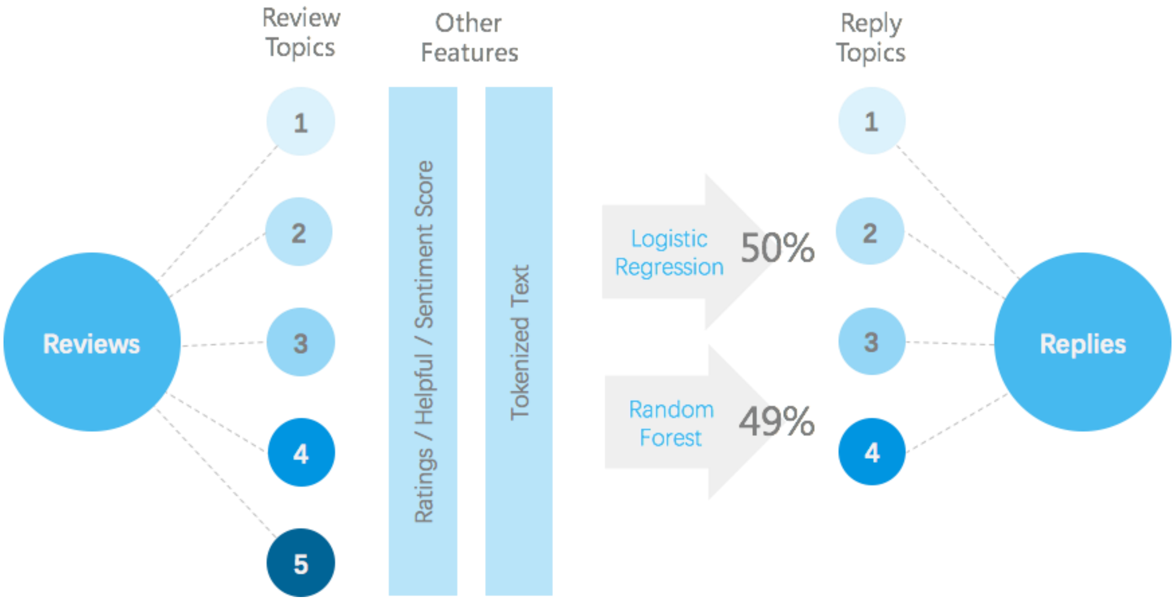


Figure 3: Topics Matching Model

5.2 Assign Pre-defined Replies According to Review Content Classifications

In the dataset, the replies to reviews with 5-star rating are similar in meaning and most of them expressed gratitude to the positive reviews. The replies to reviews with 4-star rating resemble each other by appreciating users' support and demonstrating determination to improve the app. However, reviews with 1-3 star ratings mostly complained about various aspects of the app and replies are quite different in nature. Thus, reviews were firstly separated into 3 broad groups based on ratings (5-star in one group, 4-star in one group and 1-3 stars in one group) and then treated differently to assign pre-defined replies.

5.2.1 Replies to 5-star reviews:

The number of each unique reply is counted for all 5-star reviews. The top 30 most frequent replies were used to construct the reply library for 5-star reviews (Supplementary Table 3). When a new 5-star review is logged, a reply will be randomly selected from the 5-star reply library to form auto-reply.

Auto-reply = Dear first name of reviewer, *selected reply from 5-star reply library*.

For example, a reply generated for a 5-star review "This is an amazing app for anyone who has migraines or chronic headaches" will be "Dear Ruth, *thanks so much for your lovely review! We will keep on improving the apps and addressing any issues surfaced! Jenny*".

5.2.2 Replies to 4-star reviews:

Like 5-star reviews, replies to 4-star reviews were counted accordingly and the top 20 most frequent replies were used as a reply pool for 4-star reviews. When a new 4-star review comes in, a reply will be randomly selected from the 4-star reply libraries to form the auto-reply.

Auto-reply = Dear first name of reviewer, *selected reply from 4-star reply library*.

For example, a reply generated for a 4-star review "Great app. would like there to be a daily diary bit so you can look back and see if there are any triggers you've written about or experiences you had." will be "Dear Alexandra, *thank you! the fifth star would encourage our team of developers who work hard to make this app the perfect tool for migraine sufferers. We will keep on improving the apps and addressing any issues surfaced! Jenny"*.

5.2.3 Replies to 1,2 or 3-star reviews:

Manual Labeling of 1, 2 or 3-star Reviews into Four Categories

Based on the content, reviews with 1 to 3-star ratings were manually labeled as one of the following four categories, namely "Not user friendly", "Technical", "Account set up" and "Ask for better reviews" (Table 1).

Category	**Review Content**
Not user friendly	Complaining about the app is confusing, difficult to use, not enough functions, poor function design, etc.
Technical	Mention about crash, freeze, battery draining, missing logs, update messed up, cannot open, etc.
Account set up	Talk about how to start using the app, account registration, access code, linking email, etc.
Ask for better reviews	No specific issues mentioned or somewhat sound positive.

Table 1: Manual classification of reviews with 1, 2 or 3-star ratings into four categories.

Feature Engineering

To enhance the accuracy of classification models, 7 new features were engineered from the reviews (Supplementary Table 1). Length of reviews might reflect the mood of the reviewers and the topics the reviews are about. Character per word might be associated with the sentiment as well as the topic of the reviews. For instance, an angry reviewer might just say "I hate it" while someone complaining about a specific issue of the app might use longer words. Specific words such as battery, account and crash are directly related to the categories. As these words might be dropped out during n-gram tuning in models, the appearance of these known special words is used as categorical variables. In addition, sentiment score might also be associated with the four categories.

In addition to the manually prepared features, review content is lower-cased, punctuation removed, stemmed and then tokenized. Features from bag of words were combined with manually prepared features as inputs for following model building and tuning.

Model Construction and Parameter Tuning

Reviews with 1-3 star ratings were randomly split into training (80%) and test data (20%). Data were trained on 4 level-one models - Naïve Bayes, Random Forest, Gradient Boosting, Support Vector Machine. Model-specific parameters and parameters involved in tokenization such as n-gram, maximum features and minimum threshold occurrence were tuned via 3-fold cross-validation approach. After tuning, training accuracies were all above 0.9380 for all models (Table 2). The test accuracy was lower, ranging from 0.5932 (support vector machine) to 0.7677 (Naïve Bayes). This is probably due to the small data size and overfitting of training data despite the use of cross-validation.

The predicted probabilities for each category from the 4 level-one models were used to build the stacking model by KNN. Stacking significantly improved the overall training accuracy to 0.9884 and test accuracy to 0.8136 which are better than any of the level-one models.

Accuracy	**Dummy**	**Naïve Bayes**	**Random Forest**	**Gradient Boosting**	**Support Vector Machine**	**Stacking (KNN)**
Training	*0.4380*	0.9380	0.9612	0.9884	0.9845	**0.9884**
Test	*0.4746*	0.7627	0.6441	0.7288	0.5932	**0.8136**

Table 2:Training and testing accuracy of different models.

Insights from Most Predictive Features

In Naïve Bayer model, the most predictive features for each category revealed the general topics associated with the categories (Table 3). The top features for "Technical" are all related to technical issues including crash due to "update", "battery" drain, it "used to" work, etc. For "Account set up", *access code, account* and *email* are all required to *create an* account. For "not user friendly", reviewers usually express that the app *would be* better *if* it has certain function and they *wish* to have certain functions. While for "ask for better reviews", usually the reviews were saying the user *just start* to use the app and the app is so *far* so good. In Random Forest and Gradient Boosting models, *code*, *update*, *email* and *now* are the top hit words which align with Naïve Bayer model outputs.

Category	**Top features**
Not user friendly	would be, if, to be, wish, day
Technical	update, battery, fix, now, last update, used to
Account set up	code, access code, email, account, create an
Ask for better reviews	just start, keep track, me keep, far, that help, effort

Table 3: Most predictive features in Naïve Bayes model for the 4 categories of reviews with ratings below 4-star.

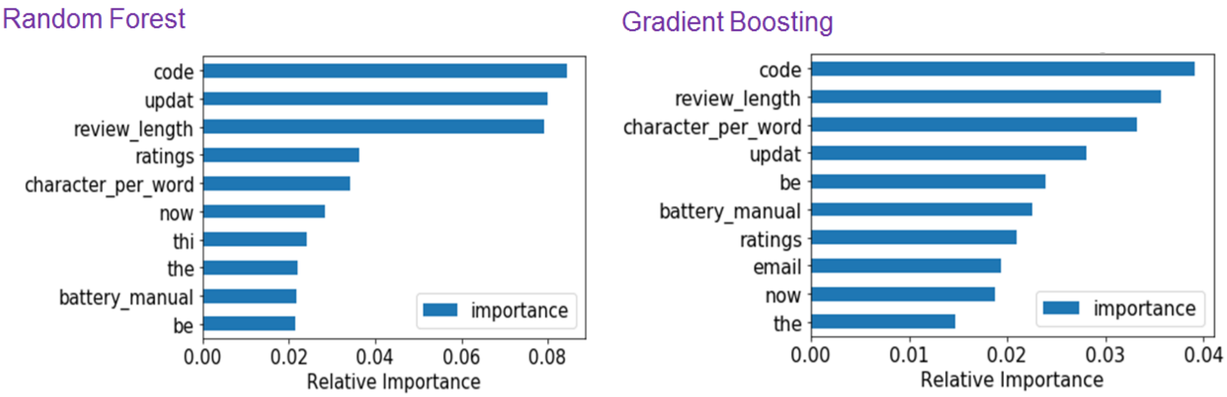


Figure 4: Important features in Random Forest and Gradient Boosting models.

For manually prepared features, both Random Forest and Gradient Boosting indicated that length of review, character per word and word battery are crucial in predicting the category of reviews, demonstrating our engineered features were extremely helpful in classifying the reviews (Figure 4).

Reply Generation

A reply was pre-defined for each category (Supplementary Table 2). The auto-reply will be formed as

Auto-reply = Dear first name of reviewer, *pre-defined reply for the respective category*

For example, an auto-reply for 1-star review “I just downloaded this app and am trying to create an account. I tried signing up with Google, through Facebook, and also with another email, and nothing will work. It swirls and says "Synchronisation" and then I get an error message about connection.” Will be “Dear Shawna, *how is your account working now? Please send an email to contact@healint.com and we will help you out right away. Thanks for your patience! We hope you give us another try as we have released many new improvements to the app since then! Jen*”.

5.3 Seq2seq Model with Attention

A sequence to sequence model, also as known as an encoder-decoder model, was firstly introduced by Google for machine translation tasks. As its name suggests, it takes a sequence as input and a corresponding sequence as output through an encoder and a decoder. Figure 5 shows a general structure of such a model.

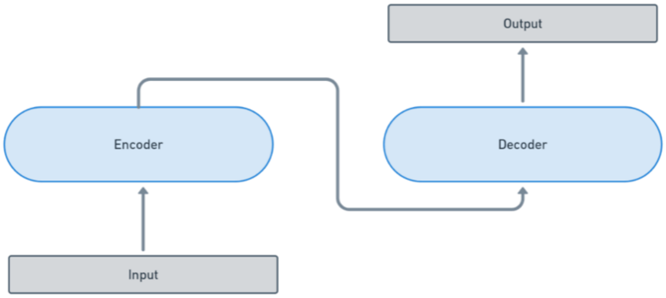


Figure 5: General Structure of a Seq2seq Model

Encoder usually consists of several layers of advanced RNN layers, such as long short-term memory (LSTM) and Gated Recurrent Units (GRU) to transform a sequence of text to corresponding hidden states. Each hidden state represents both the current input text and its context.

Similarly, decoder also consists of multiple layers of LSTM or GRU. It differs from encoder since encoder takes as input a sequence of text only, while decoder takes as input not only the output of the encoder but also its own hidden state and current text from the target sentence.

Google Brain claimed that “Attention Is All You Need” at the year of 2017. Attention mechanism lays between the encoder and decoder. It assigns weights to the output of the encoder and generate weighted hidden context vectors, which allows the decoder to pay its attention to its input strategically.

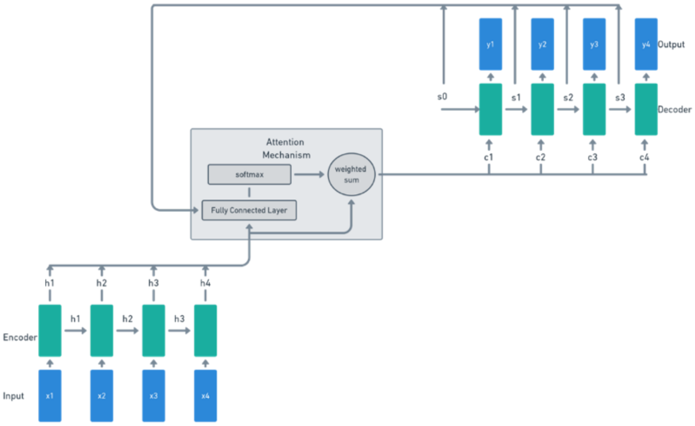


Figure 6: Attention Mechanism

Since the raw data scrapped from Google Play store is unbalanced, e.g. reviews with high ratings are much higher the ones with lower ratings. We decided to train the model on both datasets before and after applying over-balancing strategy. The training loss and test loss is shown in Figure 7.

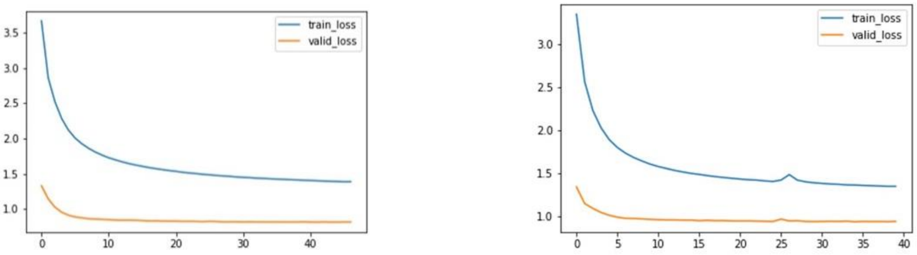


Figure 7: Training and Validation Loss Before (Left) and After over-sampling (Right)

We also generated replies for two test samples which are shown in the appendix (Supplementary Table 5). From the results, we discovered the following issues:

1. With limited input data and consistent replying style, both the models are under-fitted; whilst the generated replies followed grammar rules to some extent.
2. Both the input and output consist various number of sentences. The models are unable to identify the separate sentences and generated corresponding different number of sentences.

6. Future Work

To further extend our work, Emojis could be incorporated into the sentiment analysis to provide enhanced sentimental scores. Emojis is used by users as add-on to express their feeling in the review, it could be helpful to address the emotion further by taking it into consideration for sentiment analysis. Besides that, spelling corrections is another aspect which could be factored in to correct the review text before modellings. Both of those could help to improve the accuracy of the sentiment score and data quality as the input of data modelling. Additionally, the pre-defined replies for 4-star and 5-star reviews could consider the similarity between the review content with the historical reviews to select the best reply from the most frequent reply pool. In another words, the same reply could be used for very similar reviews and the similarity could be measured using metrics such as Cosine similarity, Euclidean distance and Minkowski distance. Lastly, this project could be expanded to generate reply for other languages which have different syntax and semantics from English.

Bibliography

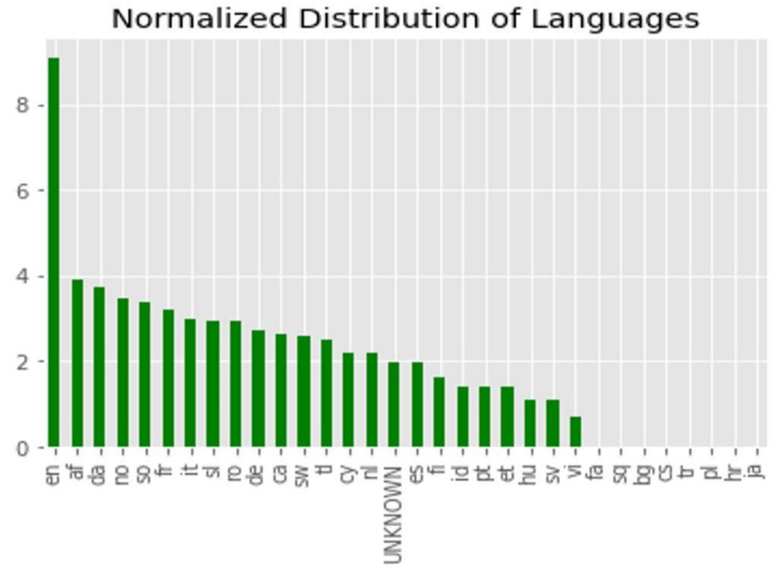
Oracle. (2012). *2011 Customer Experience Impact Report*. Retrieved from Oracle: http://www.oracle.com/us/products/applications/cust-exp-impact-report-epss-1560493.pdf

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

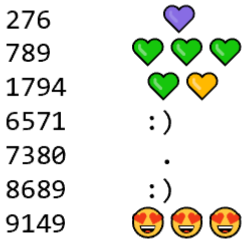
Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. et al. (2019). *Attention Is All You Need.*

Appendix

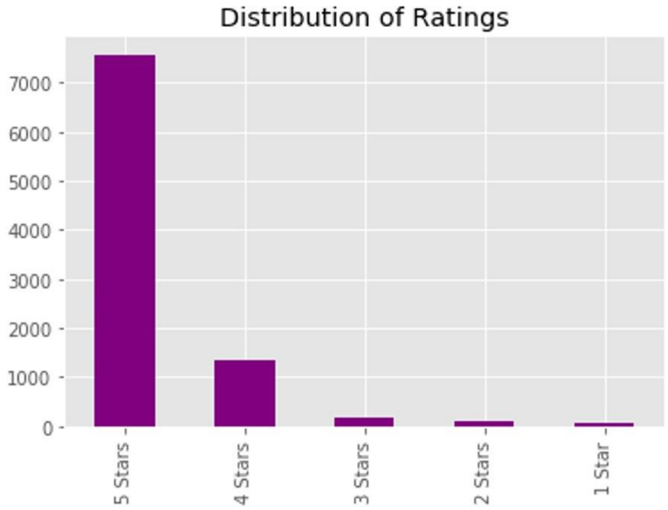


Supplementary Figure 1: Normalized distribution of Review Languages

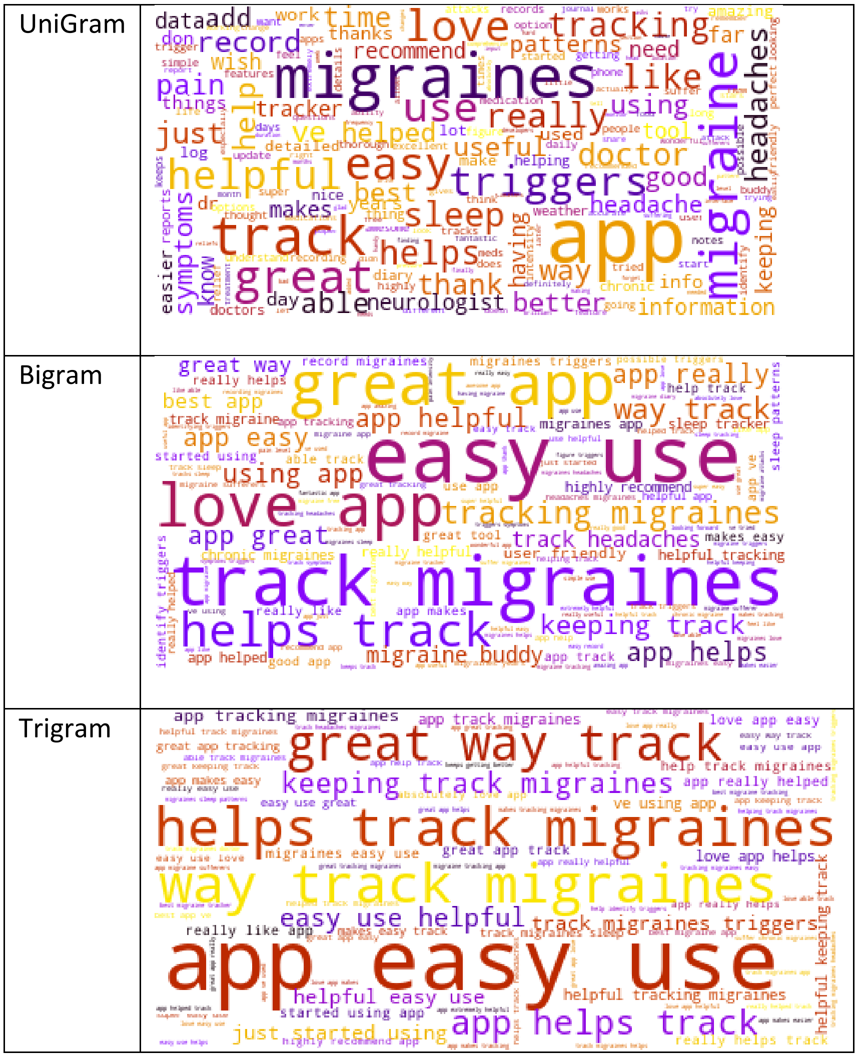


```
276        💜
789       💚💚💚
1794       💚💛
6571       :)
7380        .
8689       :)
9149      😍😍😍
```

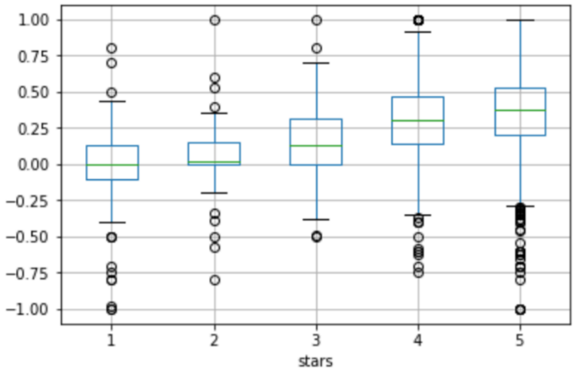
Supplementary Figure 2: Reviews with tagged as ‘UNKNOWN’ language label



Supplementary Figure 3: Distribution of Ratings in dataset



Supplementary Figure 4: Unigram, Bigram, Trigram WordCloud based on English reviews

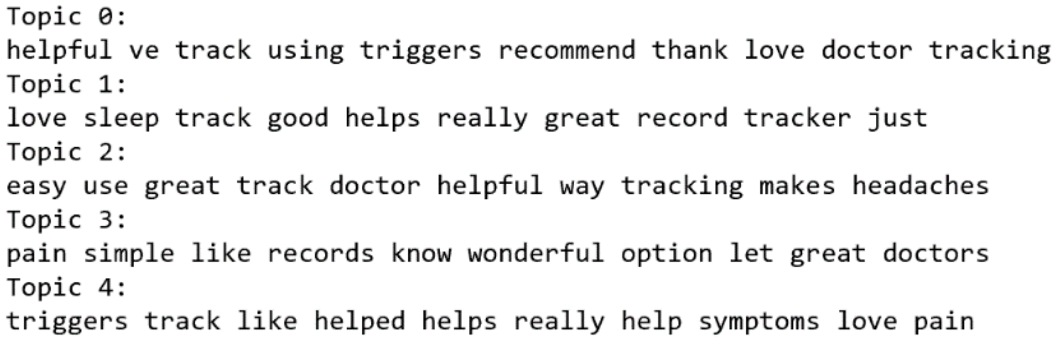


I suffer from the dreaded never-ending migraine. I am always in pain. it wants me to put an end time. I never have an end time. I've been in pain constantly for 4 years now. but the intensity changes every day and sometimes multiple times a day. this makes it very hard for me to use just about any electronic migraine diary, including this one. If there was a way I could just keep a running record of pain intensity changes along with everything else you have listed in this app it would be perfect for me. – *Review 8715*

I have been using this because I suffer frequent acute migraines but no diagnosis. Also i suffer memory loss so this really helped when going back to my doctor. My doctor loved the app and has started recommending it to her patients. – *Review 8621*

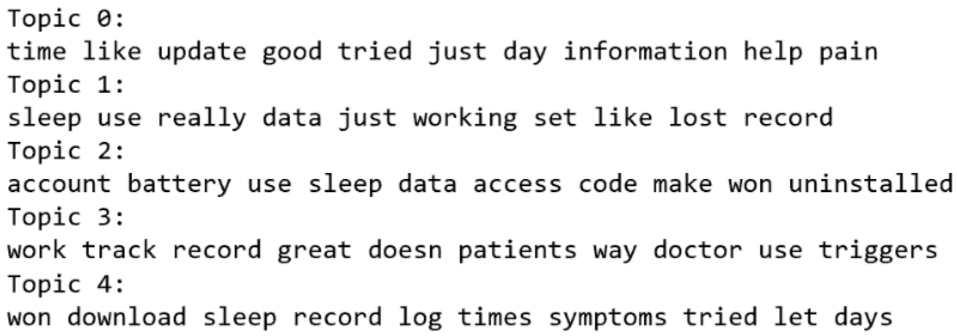
Supplementary Figure 5: Boxplot of sentiment against ratings & sample reviews indicating outliers for 5 star rating

Topics for Positive Sentiment Reviews



```
Topic 0:
helpful ve track using triggers recommend thank love doctor tracking
Topic 1:
love sleep track good helps really great record tracker just
Topic 2:
easy use great track doctor helpful way tracking makes headaches
Topic 3:
pain simple like records know wonderful option let great doctors
Topic 4:
triggers track like helped helps really help symptoms love pain
```

Topics for Negative Sentiment Reviews



```
Topic 0:
time like update good tried just day information help pain
Topic 1:
sleep use really data just working set like lost record
Topic 2:
account battery use sleep data access code make won uninstalled
Topic 3:
work track record great doesn patients way doctor use triggers
Topic 4:
won download sleep record log times symptoms tried let days
```

Supplementary Figure 6: Top 5 topics & top 10 keywords in each topic based on application of Latent Dirichlet Allocation for Topic modelling

Features	**Type**	**Meaning**
ratings	Categorical	The rating number extracted from “Rated x out of 5 stars”
review_length	Numeric	Number of characters in the review
character_per_word	Numeric	Average number of characters per word
battery_manual	Categorical	1 if word “battery” appeared in the review
account_manual	Categorical	1 if word “account” appeared in the review
crash_manual	Categorical	1 if word “crash” appeared in the review
sentiment	Categorical	1 if Afinn sentiment score >0; 0 if Afinn sentiment score <0

Supplementary Table 1: Features engineered from reviews.

Category	Standard Reply
Not user friendly	we are so sorry to hear that you have not found the application to be user friendly! We would love to make your experience better; please let us know how we can earn back those extra stars. We're happy to speak you to and address any specific concerns you have with the app.
Technical	sorry for the inconvenience caused and thank you for informing us! We have looked at the problem and issued update. Do give it a try again and let me know if it's working or not at jenny@healint.com. We will assist you ASAP.
Account set up	how is your account working now? Please send an email to contact@healint.com and we will help you out right away. Thanks for your patience! We hope you give us another try as we have released many new improvements to the app since then!
Ask for better reivews	thank you for using MigraineBuddy. Please contact me at jenny@healint.com and let me know how I can further assist you and earn the missing stars.

Supplementary Table 2: Pre-defined replies for four categories of reviews with 1, 2, or 3-star ratings.

1. thank you for your support!
2. thank you!
3. thank you very much for your support!
4. thanks so much!
5. thank you so much!
6. thank you very much!
7. thank you for your support! glad the app helps you better understand your migraines.
8. thank you so much for your support!
9. thanks very much!
10. thanks!
11. thank you for the perfect score! if you have any suggestions on how we can further improve the app, please feel free to write to me at @healint.com. thanks! =d
12. i'm glad you find migraine buddy helpful! thank you! if you have any suggestions for how we may improve the app, please let me know anytime at @healint.com.
13. thanks so much for your support!
14. thank you for your support! glad the app helps you better communicate with your doctor.
15. thank you very much for your support! glad the app helps you better understand your migraines.
16. thank you for the perfect review. it means a lot to the team and motivates everyone to move faster. please feel free to reach out to me at @healint.com anytime with feedback or suggestions for how we may improve the app! =d
17. thank you! if you have any suggestions for how we may improve the app, please let me know anytime at @healint.com.
18. many thanks :)
19. thank you :)
20. thank you for the perfect score! if you have any suggestions how we can improve the app, please feel free to share with me at @healint.com. thank you! =d
21. thank you for your support! glad the app helps you better manage your migraines.
22. thank you for your support! :)
23. thanks so much for the lovely review!
24. thank you
25. thank you so much for your support! glad the app helps you better understand your migraines.
26. thank you very much for your kind support.
27. thank you for your kind compliments! i'm glad you find the app helpful! if you have any suggestions how we can improve, do share with me at @healint.com! =d
28. thanks so much for your lovely review!
29. thank you for your continuous support!
30. thank you so much for this great review!

Supplementary Table 3: Reply library for 5-star reviews.

1. thank you! the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers. :-)
2. thank you! the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers.
3. thank you for your kind words. the fifth star would encourage our hard-working developers who make this free app the perfect tool for migraine sufferers. :-)
4. i'm glad you find the app useful! ^^ i noticed you gave us 4 stars which usually indicate minor problems with the app. =(did you encounter a problem with the app? please let us know what we can do to earn the missing star! i would love to hear from you at @healint.com. thank you!
5. i'm glad you find the app helpful! ^^ i noticed you gave us 4 stars which usually indicate minor problems with the app. =(did you encounter a problem with the app? please let us know what we can do to earn the missing star! i would love to hear from you at @healint.com. thank you!
6. thank you for your support! our goal is to deliver our users a 5 stars service. feel free to contact me at @healint.com for suggestions that will help us grab this missing star back :-)
7. thank you! the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers! :-)
8. thank you for your kind words. the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers. :-)
9. thank you! the fifth star would encourage our developers who work hard to make this free app the perfect tool for migraine sufferers.
10. thank you, the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers. :-)
11. thank you! the fifth stars would encourage our developers who work hard to make this free app the perfect tool for migraine sufferers.
12. thank you! how can we improve to make it a 5 star app for you?
13. thank you! the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers!
14. thank you! the fifth star would encourage our team of developers who work hard to make this app the perfect tool for migraine sufferers.
15. thanks so much! how can we earn the missing one star back? it would be great to hear your ideas! your support and 5 stars helps us to improve the app much faster! feel free to email me at @healint.com :)
16. thanks! how can we improve to make it a 5 star app for you?
17. thank you! anything i can do to grab the missing star? :-)
18. thank you! anything i can do to grab the missing star ? :-)
19. hello, and thank you for the feedback! please let us know how we can earn that extra star. feel free to email me any time at @healint.com and i would be happy to answer any questions you have.
20. thank you for your support! the fifth star would encourage our team of developers who work hard to make this free app the perfect tool for migraine sufferers. :-)

Supplementary Table 4: Reply library for 4-star reviews.

Reviews	**Replies generated before over-sampling**	**Replies generated after over-sampling**
It keeps crashing.	Thank you so much	thank you for your review we are working on the app of the app and we are working on the pain intensity of the migraine buddy the time is helping you to let me know what you can email me at <email>.
Helpful in tracking of attacks and possibly triggers	Thank you so much	thanks for your review we are working on the sleep diary of the migraine buddy

Supplementary Table 5: Sample reviews & replies for Seq2seq Model