Exploratory analysis of the forum data provided by Mumsnet

A non-technical description of the problem and the approaches chosen June 2021

General direction

The dataset provided by Mumsnet contains text data collected from various forums where individuals exchange information on a number of topics that are likely to be relevant to parents. In addition, in all of these threads, the participants ask for and receive advice on various topics, which results in the recommendation of specific amazon products.

The data comes from 85 different forums where different topics are discussed, including parenting, as well as style and beauty, and book recommendations. The dataset contains (i) the text of the message, (ii) the title (i.e. the thread), (iii) which forum the message was posted to, (iv) the corresponding timestamp of the message, as well as (v) a unique ID number.

Instead of selecting a specific, a priori defined problem for thorough analysis, I decided to explore the data from multiple angles to get a global understanding of the dataset. I started by analysing user behaviour on the different forums, followed by the exploration of text data using visualisation techniques and a supervised machine learning model.

Part 1. Understanding user behaviour on the different forums

First, I plotted the number of messages posted to each forum (Data_visualisation/Number of messages across forums.pdf). I found that certain forums were far more popular than others; for instance the chat forum, the forum entitled ‘Am I being unreasonable?’ and the theme of Christmas were amongst the most popular ones.

Importantly, however, these messages were unique in the sense that they all contained amazon recommendations, therefore it is possible that certain forums are more likely to be associated with these types of recommendations than others. This question could be further investigated by looking at the popularity of the same topics overall, i.e. in the dataset where no recommendations were made.

Second, I converted the time stamps into hours between 0-24 to identify peak periods of messaging on the different forums (Data_visualisation/Time of recommendations.pdf). This plot might provide an insight into whether certain topics are associated with traffic at certain intervals during the day, and whether some topics elicit a continuous interest throughout the day and well into the night. For instance, the parenting and the chat forums were associated with more night time traffic than the housekeeping or the home decoration forums. In other words, this type of analysis reveals which are the topics that are ‘keeping people up at night’.

Part 2. Exploring the *content* of the threads

In order to get a quick insight into the broad topic of the threads, I first plotted the words used in the titles in the format of a word cloud, which reflects the frequency of the words used (Data_visualisation/Titles_WordCloud.png’). I reasoned that for a quick initial understanding, the words in the title might be a better indication of the overarching topic, rather than the lengthier and more noisy message data.

Indeed, the word cloud reveals that the conversations are centred around giving and receiving advice. The high frequency of the words 'help', 'please', 'anyone', 'last minute panic' is likely to reflect the act of *asking* for advice, while the words 'make life easier', 'think', 'brought joy' might reflect the act of *giving* advice.

Overall, the word cloud reflects the common theme across a large number of diverse topics, namely the presence of product recommendations.

Second, I wanted to explore what the participants might be recommending to each other, and whether the recommended products will differ across the different forums. To that end, I plotted the most frequent collocations (2- and 3-word combinations) in each of the different forums (The folder Data_visdualisation/Frequent_collocations contains all the plots from the 85 forums). These plots appear to be highly informative in relation to the likely candidates of a recommendation: for instance, the word combinations of 'nursing bra' and 'bedside crib' were amongst the 50 most common word combinations on the pregnancy forum, while the 'try get' and 'sorry going' were amongst the most frequent collocations on the teenagers forum (which might have appeared as 'try to get…' and 'I'm sorry you are going through this' in the original texts).

Finally, in order to get a better understanding about the content of these recommendations, I intended to extract information on the recommended products. Since the recommended items usually appeared as common nouns (i.e. "shoe" instead of a specific brand of shoe), looking for proper names ("named entity recognition") could not be used to extract information on the items.

Therefore, I scraped the information contained by these links to find information on what type of item was recommended in each post. I noticed that the information on the *category* of the product that the specific item was listed under (i.e. 'Baby', 'Books', 'DIY') appeared in a predictable pattern. Therefore, once the URLs were accessed, I used these patterns to find the category of the product that was recommended in each post.

This analysis revealed that some of the most common categories that the recommended items were listed under on amazon were 'Kitchen & Home', 'Books', and 'Health & Personal Care' (Data_visualisation/Recommended_products.png').

Part 3. Do the words in a message predict the type of product that is recommended?

Finally, I investigated whether the words used in the messages would predict the category of the item that is recommended in the post. Namely, if the words 'bottle', 'night time', 'feeding' appear, the post is likely to recommend an item that Amazon lists under the category "Baby", rather than "Kitchen & Home" or "DIY", for instance.

I considered this question a multi-class classification problem, and I used a supervised machine learning model (Random Forest Classifier) that used the words in the message, as well as the category of the recommended product to establish meaningful associations between these two types of information. In this analysis, I only included the 7 most frequent product categories, where the number of posts in a category exceeded 20.

In brief, the model performed poorly, at an overall accuracy of 43%, and none of the performance metrics related to the individual topics exceeded 66%.

Although this type of machine learning might provide better results in the future, there may not have been enough data for the model to establish meaningful associations between the words and the product categories. Since many of the messages were relatively brief, there was only a limited amount of linguistic information to map onto the categories. Furthermore, there were 7 different categories used in this analysis, which ultimately resulted in a small amount of information being mapped onto a large number of different topics. However, future models with more data and/or a smaller number of categories might reveal important information on the link between the text and the recommended products.

Limitations and future directions

As mentioned above, this dataset was unique in the sense that all the included posts contained an amazon recommendation – which is likely to represent a minority of the forum data on Mumsnet's website. Therefore, the information on the popularity of a forum (measured via the number of posts overall) might simply reflect the imbalance in the likelihood of making a recommendation on the different forums. In other words, the chat forum may not be the most popular forum overall, but rather, it might just be more likely to contain amazon links. Therefore, only a dataset where no recommendations appear could give a correct estimate of the popularity of each forum.

On a related note, I noticed that a considerable proportion of the messages contained several recommendations involving the same product type. This phenomenon might be investigated to understand the difference between the posts with a single recommendation, as opposed to multiple links to different products from the same category.

With regards to the distribution of the messages over the 24 hours in a day, an important limitation is that these metrics may differ on weekdays as opposed to weekends. In addition, the popularity of the forum on Christmas seems to suggest that the messages come from a time period before or during Christmas, which might not be the most representative time of the year to measure internet traffic.

Regarding the analysis of the text data and the machine learning model, although this appears to be a good starting point, there are numerous limitations to the analysis.

Firstly, when looking at the most frequent word combinations in the different forums, I noticed certain artefacts, such as 'tag', 'ref', and 'ref tag', which come from the links themselves. These uninformative words that are due to the specific context of the data ("corpus-specific stop words") should be removed from the data before plotting or modelling. This work, however, would require some domain-specific knowledge and more time to carry out accurately.

Secondly, when using the text data to predict the type of the product that was recommended, the number of words that are represented in the model should first be restricted. Ideally, a relatively small set of informative words should be used that are most diagnostic of the categories under consideration, while the words that are unimportant to this classification problem should be discarded ("feature engineering"). However, similarly to the previous problem, this could only be carried out with a more thorough analysis and more time exploring the unique properties of text data coming from chat forums.