**Project 5 – Advanced Data Mining Applications**

**CS548 Knowledge Discovery and Data Mining - Fall 2016**

**Prof. Carolina Ruiz**

**Student: Deepan Sanghavi**

|  |  |
| --- | --- |
| Description of the particular problem within the selected data mining topic to be addressed in this project | /15 |
| Description of the approach used in this project to tackle the above problem | /25 |
| Description of the dataset selected | /15 |
| Appropriateness of the dataset selected with respect to this topic/problem | /10 |
| Guiding questions | /10 |
| Preprocessing | /10 |
| **Experiments:**   * Sufficient & coherent | /25 |
| * Objectives, Data, Additional Pre/Post-processing | /20 |
| * Presentation of results | /20 |
| * Analysis of results | /30 |
| Overall discussion, comparisons, and conclusions | /20 |
| TOTAL | /200 |

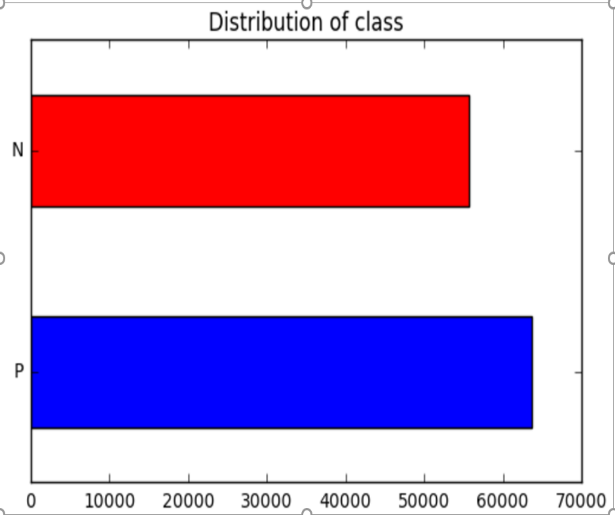
Total Written Report: \_\_\_\_\_\_\_\_\_\_\_\_\_\_/200 = \_\_\_\_\_\_\_\_\_\_\_/100

Class Presentation: \_\_\_\_\_\_\_\_\_\_\_/100

Class participation during project presentation: \_\_\_\_\_\_\_\_\_\_\_/100

*Do not exceed the given page limits for this written report*

**Topic: Text Mining – Amazon user questions and reviews about baby products**

1. **Description of the problem within the selected data mining topic to be addressed in this project:** Develop insight about goodness of baby products by analyzing reviews with high and low ratings. Moreover, summary of these reviews is analyzed to get an overview about user likings and disliking. Furthermore, analyzing user questions which can assist in reducing the queries by highlighting common questions. Also, recommending the baby shower gift that most people like it.
2. **Description of the approach used in this project to tackle the above problem:** Initially, removed the stop words and performed stemming operations on text data. Then converted the unstructured data into structure data using term-document IDF frequency weight and used python scikit-learn toArray package to convert into table format. To load all the text files (each text file represents a review) in Weka, used Simple CLI to load the data. Further, used Weka’s string to word convertor to convert text data into table format. After that, by considering highly correlated attributes to the target (CFS), applied decision tree and association rules to find insight about the data.
3. **Dataset Name**: Amazon baby products Reviews and Queries
4. **Where found:** http://jmcauley.ucsd.edu/data/amazon/ (Baby products reviews)
5. **Dataset Description:** Dataset contains baby product reviews and metadata from Amazon. This was a demo version of the original data set. It had columns like product reviews, reviewer information, price, category (baby), ratings and product ID. But, I got an access to full version of data after requesting the owner (Julian McAuley). This version included additional product meta data information which included product Id, price, name, items frequently bought and description of the product. Ratings were given to each review from range 0 to 5. Moreover, it had additional dataset which consist of user queries, answer to those queries, time, user ID and product ID. Each query was distinguished as Yes/No query or Open Ended Queries

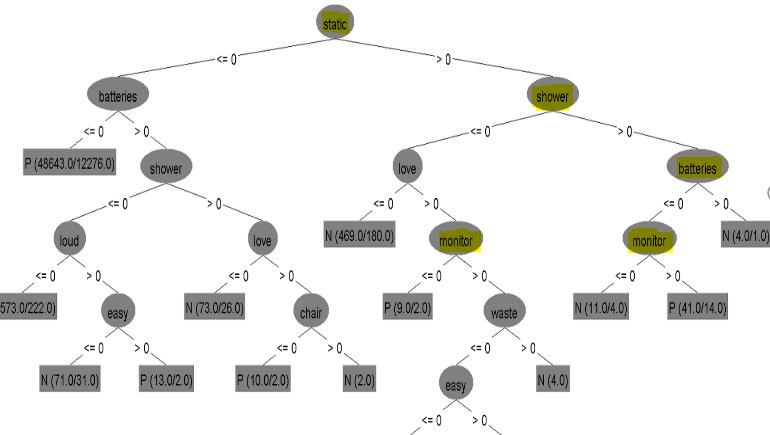
**Initial data preprocessing, if any:** Created a class attribute by giving Positive value to review which are rated above 2 and Negative to others. Merged Price information of each product using product metadata with the original review data set. Also, created one more class attribute by giving expensive tag to product whose price is greater than $45 and others as Cheap. Removed redundant attributes (user information and time date information) which aren’t part of my analysis. Performed Stemming and filtering of stop words using NLTK library in Python. Used toArray, TFIDVectorizer scikit functionality and String to word vector in WEKA to get structured data. Converted each review as text file to process in Python and combined all the files using commands in Simple CLI.

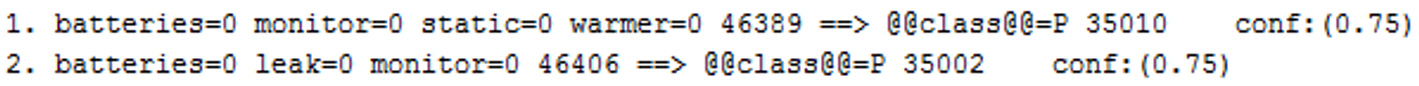
1. **Three Guiding Questions about the dataset domain:**
2. What are the qualities that makes user dissatisfied with the baby product?
3. Is there any specific category of products about which user usually appreciate or complains?
4. Which are the product that people prefer to gift for baby shower?
5. What should be included in the product description that reduces user queries?

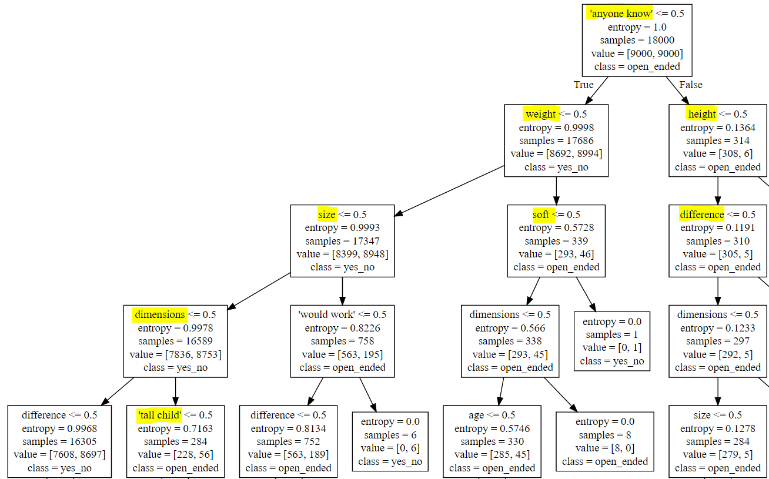
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Tool** | **Pre-process** | **Mining**  **Technique** | **Results** | **Time**  **taken** | **Evaluation** | **Observations about experiment**  **Observations about visualization**  **Interpretation results** | **Resolution or Improvement** |
| 1 | Weka | Considered Summary of Reviews columns and removed user review column.  Ngram= (1).  K=10 folds | J48 Decision tree | Correctly classified instances = 72.33%. | 33.61s | F-score: 64.4%  ROC Area:0.509 | Stinks, static, smelly, breaks are few words which user use often in negative reviews. Acceptable F-score (combination of Precision and Recall) and ROC value. | Company should look avoiding these traits in their product. E.g. Powder are expected to have good smell. Devices should concentrate on static sound issue. Male less fragile baby products. |
| 2 | Weka | Removed columns: Summary of review.  Ngram:1  Filtered stop words based on the file which contained list of stop words.  K=10 Folds | J48 Decision tree (Used K Folds =10) | Correctly classified instances = 74.92%.  Visualization shown below | 72.08s | F-Score: 64.7%  ROC Area:0.512 | Static noises made by sound devices and finding batteries for these devices are the major user issues. User also complains about static noises even after the battery change. But large amount of user love monitor devices as their baby loves it.  Good F-score and ROC Area | Along with these devices, extra batteries can be provided for long run use. Also, good detail about batteries should be provided. Concentrate on sound system while developing a product. People love monitor devices produced by crib CD Player. |
| 2 | Weka | Same as above | Association Rules | Absence of Batteries, monitor, static, warmer, leak, milk = P | 15s | Confidence: 0.75 | Negative user reviews stating the issue in bottle warmer system of making water or milk warmer than expected and leak problems when not kept upright. High Confidence value | Many users do not read direction of use properly and because of that this issue occurs. Can update the product to avoid leak issues and provide simple direction of use. |
| 3 | Python | Filtered reviews which had ‘shower’ word in it.  Class – Cheap (Below 45$) or expensive  Ngram= (1,2)  K=10 Folds | Decision tree Classifier | Cheap: duck, spray, bather, batteries  Expensive: Playtex, monitor, sling carrier and champ  are high entropy words | 3.45s | Precision=72.43%  Recall=60.76%  Score=67.95% | Champ is nappies brand and Playtex is drop ins BPA starter set (Product metadata). Moreover, user prefer to give bather, duck, sprayer, sling carrier and monitor as a shower gift. | Company can promote its product by suggesting these products to users to give as a shower gift. include a section of shower gift and put these products on top |
| 3 | Weka | Same as above | Association Rules | Absence of carrier, champ, genie, monitor, pail, sprayer, static, sterilizer ==> Cheap | 12s | Confidence:0.57 | Apart from above, genie is also a diaper brand which people prefer to give as a shower gift. | Diaper can be in the strong suggestion list of shower gift |
| 4 | Python | Max depth=4  K= 10 splits  Ngram (min=1, max=2) | Decision tree Classifier | Anyone knows, size, soft dimension, tall child, height, age, weight are common high entropy words | 56.23s | Precision=80.6%  Recall=60.34%  Score=70.34% | There are certain Yes/No queries which most of the user have. E.g. Most user wants to know for which age is X product suitable or what are the dimensions of product | Information on all Yes/No queries can be answered by giving more information in the description of product. This will reduce the number of user queries |
| 4 | Weka | NGram= (1,3)  Split=66% | Decision tree Classifier | Apart from above, machine washable, made of | 6.72s | ROC=0.61  F-score=50.4% | Most user want to knows whether the product is washable or its important to know in baby products that what it is made up of | Include this information in description of product though they are part of open ended questions to reduce queries |

**Analysis of Results: (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.

Removing stop words and stemming using scikit and Weka did not provide acceptable results. Hence, I used NLTK library in Python and a file which contains list of stop words in WEKA to remove stop words and performing stemming. Moreover, I varied Ngram parameter to get better insight about results.

Guiding Question 2: After analyzing specific reviews, I got to know that most user complains about static noises (The First Years - Crib CD Player) that few sound devices made even after changing the battery. Also, for devices that they receive as a shower gift, finding batteries for them is a difficult task. But most of the user love monitor (Graco ultra-baby monitor) devices as visual images make their babies happy and they love it.



Guiding Question 4: Data had two categories of user queries, Yes/No and Open ended. Yes/No queries were easy to answer and for a company, it is beneficial if the user has least number of queries. So, my analysis, found few keywords that are found in Yes/No and open ended queries. This analysis can be used to update the description of product. For example, user ask a lot about dimension and so it’s better to add such information in description itself.

**Summary of what you learned in this project:**

* Importance of stop words filtering and stemming in text mining
* Ngram functionality and how it can help in giving better insight
* String to Words Vector and doing Text Mining in Weka. TFID Vectorizer in python, though I did not use it for my results, it helped in finding out important keywords