**Project Brief: Amazon Fine food reviews Analysis**

**Team 2**

**Abhishek Shivaprakash**

**Shengjie Shi**

**Xinlei Ma**

**Yue Shangguan**

**Viraj Karnik**

**Vivek Jayaraman**

**Project Summary**

Amazon is the fastest growing companies in the USA with focus of their online marketplace always being on capturing market share and getting the best output from each of its users. Fifty-five percent of shoppers start their buying research on Amazon, [a survey by marketing firm BloomReach found](http://bloomreach.com/amazon/), and half of all shoppers say they rely primarily on Amazon for reviews, according to Market Track, an e-commerce analysis firm. It is estimated that around 20% of the sales are driven by reviews and one-third of the buyers don’t buy a product that has not received a positive review. The customer reviews are also a great tool to find out the customer perspective so that better recommendations can be provided to them in the future. They also provide a means to judge how influential their reviews are on other customers and how their reviews can be used to increase the output from other customers.

The dataset under consideration contains reviews of Amazon Fine foods from Oct 1999 to Oct 2012. There are around 568,454 reviews given by 256,059 users. There are 74,258 products that are reviewed in this dataset.

The project first and foremost focuses on finding out whether the review is a positive or a negative one. The dataset is also used to find out how well the other users agree on these reviews. Later on, specific users are focussed on. The past reviews are used to generate patterns of likes and dislikes of the customers in fine foods so that they can be recommended better products in the future.

**Overview of Needs**

The project sponsor hopes to find the nature of reviews and how many people are finding these helpful. They also hope to get a better understanding of the requirements or preferences of the people reviewing the products.

**Project Deliverables**

The following are the proposed deliverables for this project.

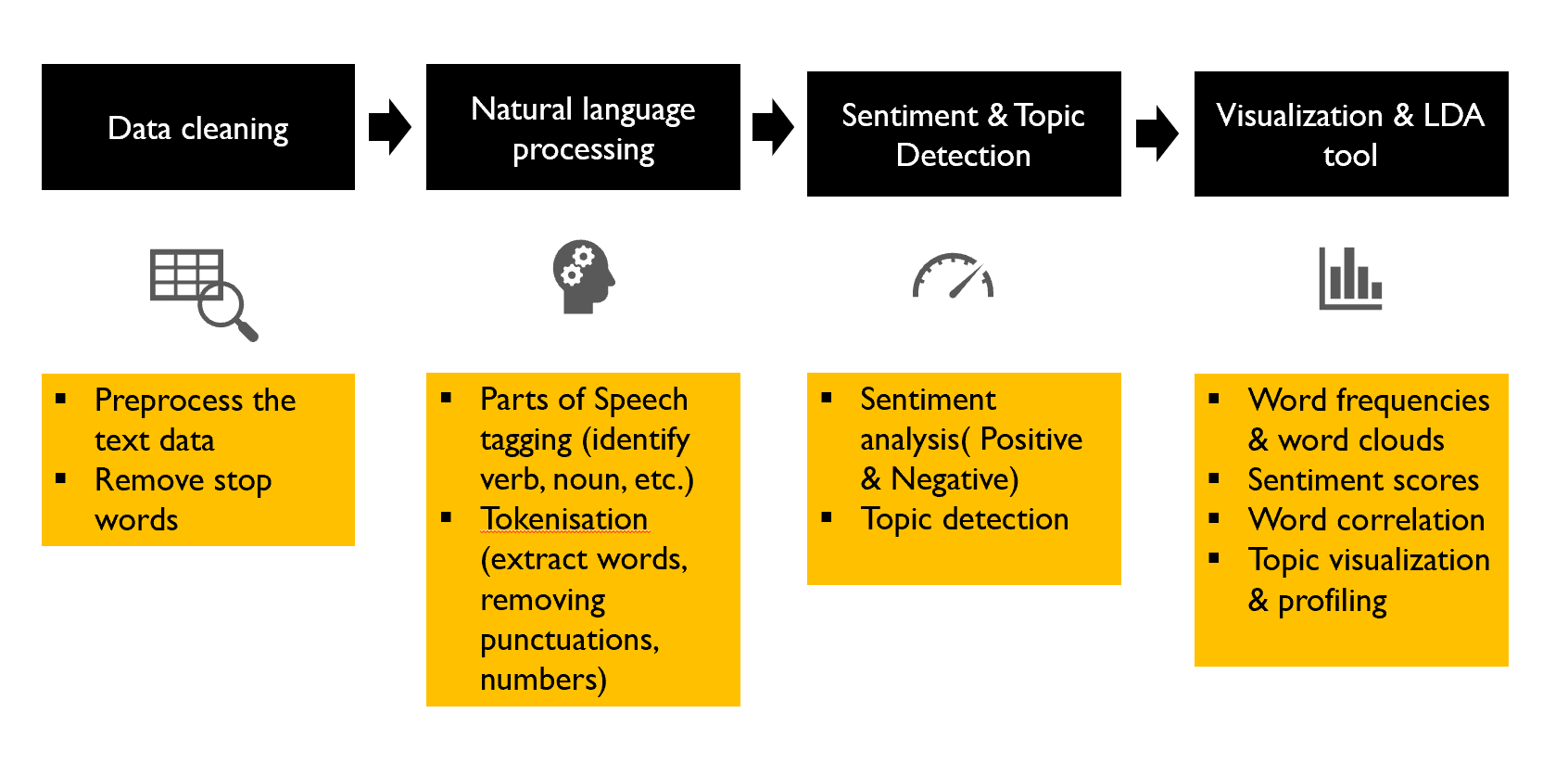
1. Project brief
2. Project proposal presentation
3. R Script
4. Final presentation

**Data Summary**

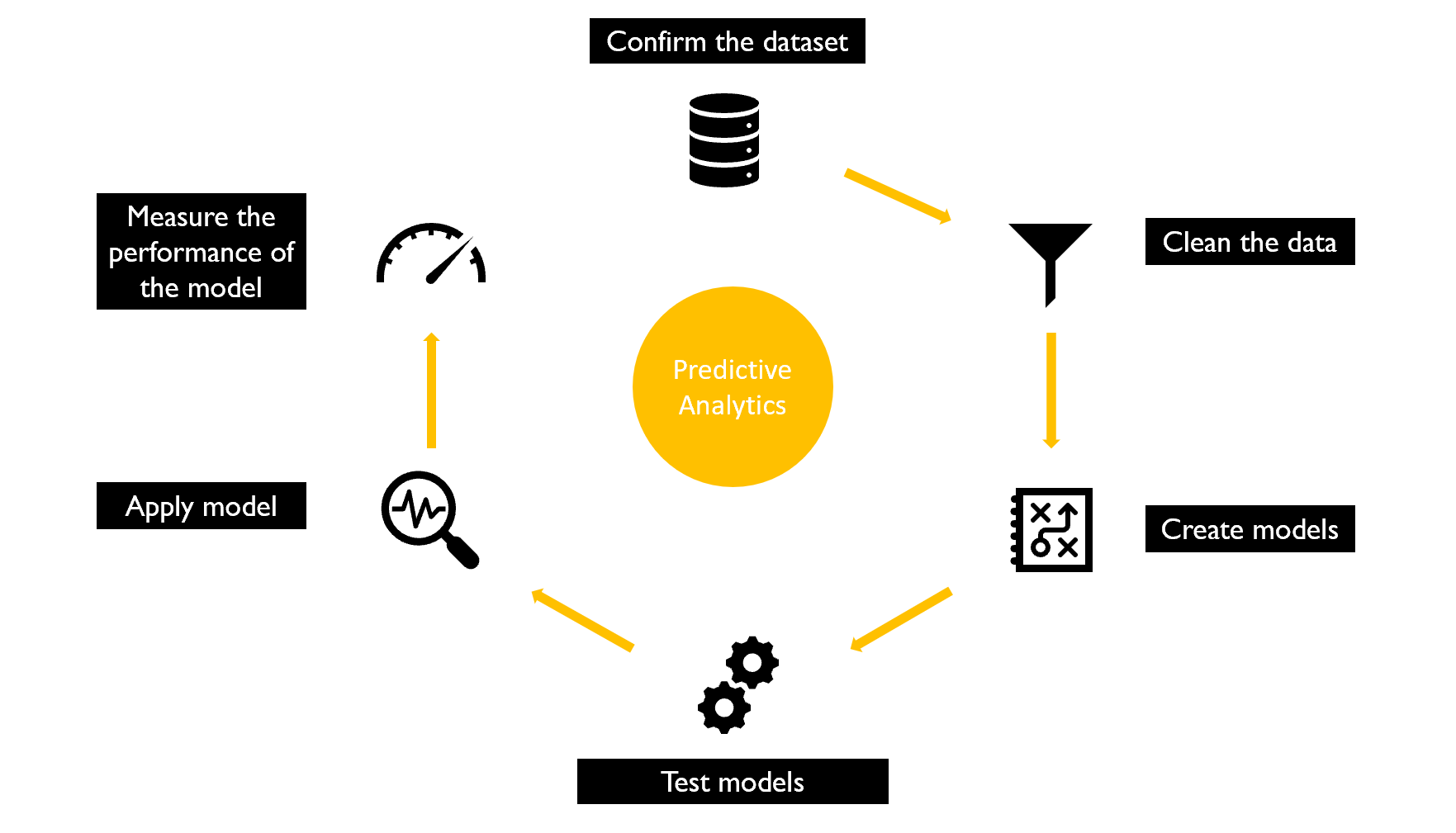
This dataset consists of reviews of foods in amazon from Kaggle, which includes product id, user id, rating scores, time, summaries and 568455 text reviews. That’s enough for us to do the sentiment analysis and user behavior research. However, we do not have the exact product information. The dataset just have brief summary of project and we do not have more information like brand, size, color, etc., which may create some bias for the result. If we have more auxiliary data, we can get more analysis from that. For out-of-score, we cannot use the data to give a specific recommendation for user.

**Proposed Methodology**

Sentiment Analysis Methodology

****

Predictive Analytics Methodology

****

**Project Success**

* Sentiment Analysis Metrics and Evaluation

We will use the most frequently way to understand the performance of our sentiment analysis: cross-validation.

We first split the training data into a certain number of training folds and the same number of testing folds, use the training folds to train the classifier, and test it against the testing folds to obtain performance metrics. Cross-validation helps prevent the overfitting of testing dataset.

Next, we use precision, recall, and accuracy as our metrics. They are standard metrics used to evaluate the performance of a classifier. Most frequently, precision and recall are used to measure performance since accuracy alone does not say much about how good or bad a classifier is.

Furthermore, in order to compare our learning algorithms, we will build the ROC curve. The curve with the highest AUC value will show our "best" algorithm. After plotting the ROC curve, it would appear that which specific method provides us with the best results. Maybe there exists a conflict between ROC curve performance and AUC value, so we will use ROC curve performance as the final evaluate matric.

* In-depth study evaluation on user behavior

Since there are no numeric tools to score the performance of user behavior analysis, we will build top 10 users’ food preference profile to generate business recommendations and implications to Amazon. We hope that our user behavior analysis can make an exploratory contribution to Amazon's user group analysis database in the future

* Schedule

We will do our best effort to keep the project on the original timeline. Past project experiences told us that we know how hard that is, but it’s a little bit easier if we continually evaluate our progress as we go.

We will update our project schedule regularly, at least once every two days. The schedule evaluation is something we can do more formally at the end of the stage or phase. We will also look at our major milestones and check if they still fall on the same dates as we originally agreed.

**User-Stories**

Sentiment analysis is useful to product analytics because it helps Amazon of the following:

Through our analysis, we can target users as frequent and infrequent reviews and explore their likes through text analysis. In the depth analysis, we can even know the reason about users’ preference. Sentiment analysis provides us a real time track tools to monitor users’ reviews and help Amazon act immediately based on reviews.

Our user behaviors analysis aims to build a user preference profile to better help Amazon segment the target user market. Since the profile library established, Amazon can target users to achieve accurate advertising goal.

Sentiment analysis offer an interacted way for customers and product designs. Amazon had real feedback from real customers, directly reaching the ears of the people to whom it mattered most. As any great product team does, Amazon listens to the customers and meet their needs.

However, the dataset just have brief summary of project and we do not have more information like brand, size, color, etc., which may create some bias for the result. If we have more auxiliary data, we can get more analysis from that. For out-of-score, we cannot use the data to give a specific recommendation for user. Besides, the data collected in the dataset is too old and we do not use the crawl technology to crawl the real time reviews. So the contribution to the future analysis might be decreased.

**Project Plan and Milestones**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Deliverables/Milestone | Start Date | Plan Duration | Completion Date |
| 1 | Topic Selection | 04/25 | 1 | 04/25 |
| 2 | Dataset Identification | 04/26 | 1 | 04/26 |
| 3 | Project Brief | 04/27 | 3 | 04/30 |
| 4 | Dataset Cleaning | 04/28 | 4 | 05/02 |
| 5 | Descriptive Analysis | 04/28 | 4 | 05/02 |
| 6 | Project Pitch Slides | 04/29 | 2 | 05/01 |
| 7 | Text Mining | 04/30 | 2 | 05/03 |
| 8 | Sentiment Analysis | 05/01 | 4 | 05/05 |
| 9 | Users Behaviors Analysis | 05/04 | 2 | 05/06 |
| 10 | Business Recommendations | 05/06 | 1 | 05/06 |
| 11 | Final Presentation Slides | 05/06 | 1 | 05/07 |

**Project Roles and Responsibilities**

|  |  |  |
| --- | --- | --- |
| Name | Competency | Responsibilities |
| Abhishek | Problem solving, influencing | User behavior analysis, modeling, sentiment analysis |
| Shengjie | Leadership, decision making, problem solving | Algorithm exploration, modeling, sentiment analysis, presentation |
| Viraj | Critical thinking, communication | Proposal, presentation, dataset cleaning |
| Vivek | Problem solving, influencing | Descriptive analysis, sentiment analysis |
| Xinlei | Problem solving, critical thinking time management | Slides, proposal, description analysis |
| Yue | Critical thinking, perspectives, communication, time management | Slides, proposal, dataset cleaning, business recommendations, presentation |