Summarization of user reviews on e-commerce websites

- Sai Shibi M R (2012103056)
- Manikandan P (2012103554)
- Pradeep T (2012103570)

Project Guide:

Dr. Arockia Xavier Annie,

Assistant Professor,

DCSE, CEG

Abstract

- E-commerce is a rapidly growing industry at present, with the number of regular users of these sites increasing day by day.
- A recent survey suggests that about 80% of users read the user-review forums of a product before they buy it.
- But the problem with these reviews is that they are not organised. Some users write about the product, while some people write about the service (like delivery, packaging etc.) and for popular products these reviews are big making it hard to read.

Abstract *(contd.)*

- These product reviews can further be categorized into positive, negative or both.
- Our aim is to organize the product reviews from all popular e-commerce websites and summarize those reviews, so that users can get a clear idea about the product.
- Among many e-commerce websites we extract reviews from Flipkart , Amazon and Snapdeal since these websites have many number of users.

Problem Statement

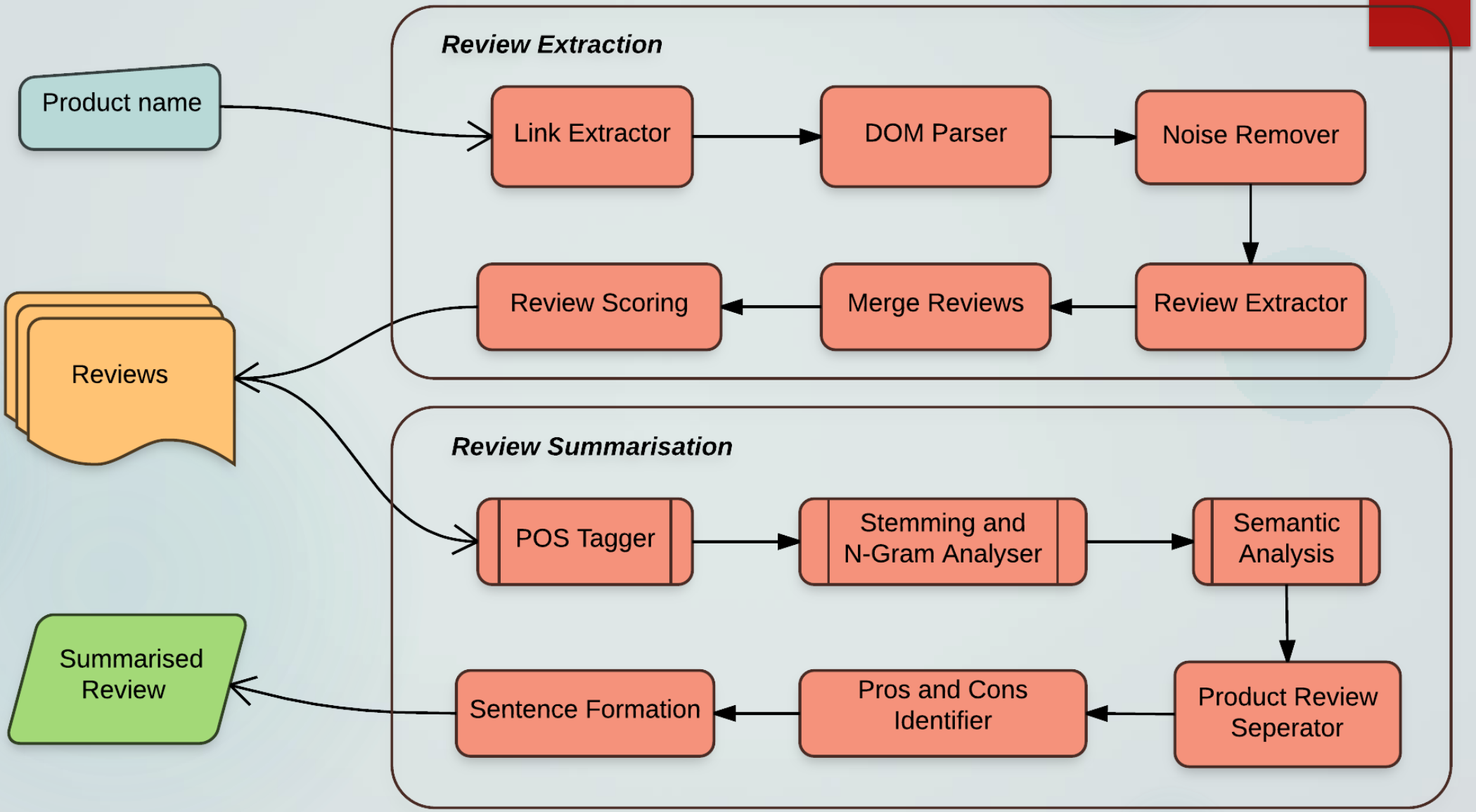
- To build a system to retrieve users' feedback (reviews) from e-commerce websites (like Amazon, Flipkart, Snapdeal etc.) and summarize the extracted reviews into pros and cons.

Problem Domain

- Data Mining
- Natural Language Processing

Module Description

- Extraction of reviews
- Summarization of the reviews


Review Extraction
Product name
Link Extractor
DOM Parser
Noise Remover
Review Extractor
Merge Reviews
Review Scoring
Reviews
Review Summarisation
POS Tagger
Stemming and N-Gram Analyser
Semantic Analysis
Product Review Seperator
Pros and Cons Identifier
Sentence Formation
Summarised Review

Link Extractor

- The input of this module is the product name.
- There are many e-commerce websites of which Amazon , Flipkart and Snapdeal are popular.
- This module aims at extracting the link of the review page of the given product.

Noise Remover

- A tree structured output given by DOM parser is the input of this field
- This removes unnecessary contents like images, advertisements etc..,

Review Extraction

- With some sort of scoring, the reviews are extracted from the web page.

Merging of Reviews

- Last two modules are repeated for all the three websites and the reviews from individual websites are merged.
- One complete sub module(Information Extraction) is done.

Modules Description

Fetching user reviews:

- Every product contains reviews written by the customers. This module brings all those reviews.
- Reviews can be categorized into product reviews which are actually the pros or cons about the product, delivery reviews which tells about the service.
- We do not need to worry about the delivery reviews which are out of our project's aim.
- So the module also involves extraction of the product reviews separately which actually involves Natural Language Processing

Modules Description *(cont.)*

Summarizing reviews:

- This module involves elimination of unnecessary sentences in the extracted reviews. Natural Language is an ocean.
- So rather than eliminating unnecessary sentences, pick(include) the necessary sentences that actually describes the product.
- This is done with “phrase match” and “frequency”. Phrase match actually plays a great role in this module.
- Thus at the end of this module, the summary is available which then pass through another module where the pros and cons are separated.

Modules Description *(cont.)*

Separating Pros and Cons:

- The project basically aims at separating the pros and cons. In this module it is done.
- For this, decide whether a particular review is positive review or negative review.
- The decision is made by matching with set of positive and negative words.

Modules Description *(cont.)*

Sentence Formation:

- With the set of positive and negative words or reviews found so far, sentences are formed and appended to respective sections(Append positive sentences to pros and negative sentences to cons).

Literature Review

1. "Recommender systems based on user reviews: the state of the art"

Li Chen ·Guanliang Chen ·Feng Wang

- Advanced text analysis and opinion mining techniques enable the extraction of various types of review elements, such as the discussed topics, the multi-faceted nature of opinions, contextual information, comparative opinions, and reviewers' emotions.
- In this article, they provide a comprehensive overview of how the review elements have been exploited to improve standard content-based recommending, collaborative filtering, and preference-based product ranking techniques.
- The review-based recommender system's ability to alleviate the well-known rating sparsity and cold-start problems is emphasized. This survey classifies state-of-the-art studies into two principal branches: *review-based user profile building* and *review-based product profile building*.
- In the user profile sub-branch, the reviews are not only used to create term-based profiles, but also to infer or enhance ratings. The product profile can be enriched with feature opinions or comparative opinions to better reflect its assessment quality.

Literature Review *(cont.)*

2. "Information Extraction from Web Pages" Novotny, Róbert; Vojtas, P.; Maruscak, Dušan

- This paper presents a chain of techniques for extraction of object attribute data from web pages which contain either multiple object data or detailed data about a single object.
- We discover data regions containing multiple data records, which will be extracted with help of extraction ontology.
- Furthermore, we present an additional algorithm for detail-page extraction based on the comparison of two HTML subtrees.

Literature Review *(cont.)*

3. "Thai herb information extraction from multiple websites" Chainapaporn, P.; Netisopakul, P.

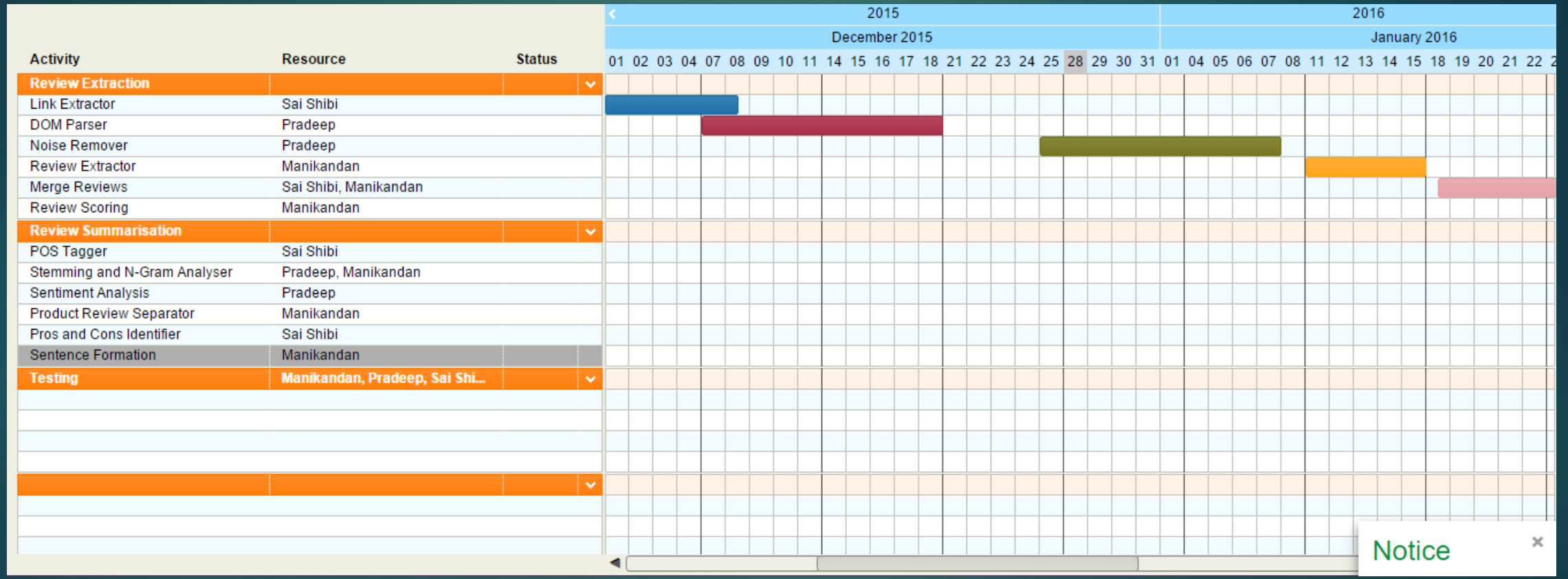
- Thai herbs have increasingly gained public attention. Recently, there are a number of Thai herb websites.
- Each website has similar information but quite different details. For example, some webpages do not provide information indicating which part of Thai herb can treat the specified symptom.
- In order to collect more complete Thai herb information, we have developed information extraction process to extract Thai herb information from multiple websites.
- The process employed a HTML parser and file templates to recognize useful information in various webpage formats.
- Preliminary experiments gave satisfactory precision and recall over 85 percent

Literature Review *(cont.)*

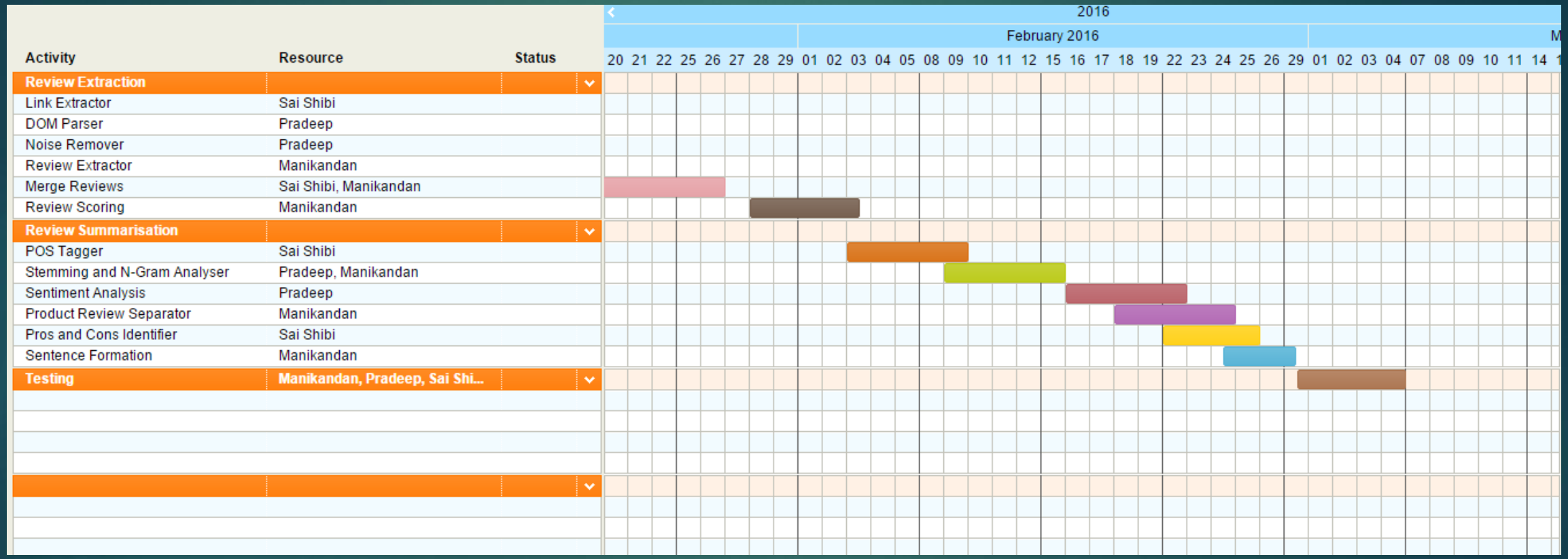
4. "Classification and Summarization of Pros and Cons for Customer Reviews" Hu, Xinghua; Wu, Bin

- As e-commerce is becoming more and more popular, the number of customer reviews for online products grows rapidly. For a popular product, there can be hundreds of reviews. This makes it difficult for a potential customer to read all of them in order to get as much information as possible and to make a decision on purchasing.
- Therefore, a summarization of product reviews would make purchase more convenient and reliable. The conventional way of summarizing a review is to select or rewrite a subset of the original sentences from the review, which is inefficient.
- In this paper, we propose to summarize all customer's reviews of a product as a list of phrases named pros and cons list, which can be perceived and understood at a glance.
- We employ a score algorithm which considers the strength of a word towards positive or negative orientation to calculate and weigh the sentiment of a sentence. To assess our algorithm, a number of existing classifiers are also presented. Our experimental results show that our Sentence Weight classifier is more accurate and effective than those compared.

Gantt Chart


2015
2016
December 2015
January 2016
01 02 03 04 07 08 09 10 11 14 15 16 17 18 21 22 23 24 25 28 29 30 31 01 04 05 06 07 08 11 12 13 14 15 18 19 20 21 22
Activity
Resource
Status
Review Extraction
Link Extractor
Sai Shibi
DOM Parser
Pradeep
Noise Remover
Pradeep
Review Extractor
Manikandan
Merge Reviews
Sai Shibi, Manikandan
Review Scoring
Manikandan
Review Summarisation
POS Tagger
Sai Shibi
Stemming and N-Gram Analyser
Pradeep, Manikandan
Sentiment Analysis
Pradeep
Product Review Separator
Manikandan
Pros and Cons Identifier
Sai Shibi
Sentence Formation
Manikandan
Testing
Manikandan, Pradeep, Sai Shi...
Notice

Gantt Chart *(contd.)*



References

- [1] Raut V.B and Londhe D.D :“Opinion Mining and Summarization of Hotel Reviews” Computational Intelligence and Communication Networks (CICN), 2014 International Conference on Year: 2014
- [2] Amiya Kumar Tripathy, Revathy Sundararajan, Chinmay Deshpande, Pankaj Mishra, Neha Natarajan “Opinion Mining from User Reviews” International Conference on Technologies for Sustainable Development (ICTSD-2015)
- [3] Novotny, Róbert, Vojtas, Maruscak and Dušan : “Information Extraction from Web Pages” Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences.
- [4] Chainapaporn P and Netisopakul P : “Thai herb information extraction from multiple websites” Knowledge and Smart Technology (KST), 2012 4th International Conference.