Product Defect Mining

Grant Golden, Jack Hall, Thomas Nguyen, Tianchen Peng, Elizabeth Villaflor, Shauicheng Zhang
CS 4624: Multimedia, Hypertext, and Information Access
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Instructor: Dr. Edward A. Fox
Client: Xuan Zhang
Spring 2017

Overview

Project Goal

Problems Faced

Summary of Deliverables

Lessons Learned

Acknowledgements

Project Goal

Use a machine learning algorithm to recognize and identify the different entities that make up a product defect review.

Model	EXPLORER 2002
Description	I have a 2002 Ford Explorer. **The o/d light was flashing and the transmission was slipping.** *We took it to our mechanic and it was diagnosed that the transmission needed replaced..*

Table 1: **Complaint Example** (**bold**: symptom, *italic*: resolution)

Problems Faced

Agreement on label definitions

- What counts as a “resolution”? What should be ignored?

Finding forums with a sufficient number of unique posts (at least 1000 desired)

Finding forums that included solutions for most of the posts

Lack of previous experience with machine learning/classifier training

Summary of Deliverables

Collected Datasets:

- Apple Community: 10,942 records
  - MacBook Air, MacBook Pro, iTunes
- Dell Alienware: 22,726 records
- Samsung: 12,000+ records
- Asus
  - Still being crawled, number of collected records unknown
  - Around 1 million available

Summary of Deliverables (cont.)

Labeled Datasets and Scripts

- Scripts: split records into single sentences, count number of labeled records, calculate percentage of matching labels between two datasets
- Dataset labels are within 90% matching between two people

Classifier Training Script

- Linear SVC Algorithm

Classifier Training Result Summary

Precision: 75-80%

Recall: 75-80%

F1-Score: 70-75%

Classes with fewer sentences performed worse

- Statistics were generally lower when predicting “resolution” sentences, since there were fewer present in the labeled dataset

Lessons Learned

Quantity over quality

- Resolutions don’t necessarily need to be correct
- The more data available, the better the classifier results

Previous experience helpful

- A lot of time dedicated to research, took away from actual classifier training

Acknowledgements

Thanks to Xuan Zhang for all of his patience, support, and willingness to teach throughout this semester.