**Feature Based Classification of Product Reviews**  
Shruti Vangari, Kashyap Uppuluri, Mukesh Kumar Chippa  
Department of Computer Science  
The University of Akron  
[sv47@zips.uakron.edu](mailto:sv47@zips.uakron.edu), [ku10@zips.uakron.edu](mailto:ku10@zips.uakron.edu), [mc80@zips.uakron.edu](mailto:mc80@zips.uakron.edu)

**Abstract**

*With the advent of technology in this era, manufacturing companies have been changing the product features depending on the customer requirements from time to time. The customer reviews of these products also changed with the features proportionately. Hence product reviews based on the increasing features became an integral part of a product’s success. In this paper, we have considered the customer reviews from amazon for two products Nokia 6610 (released in 2004) and Nokia 6600 (released in 2009). We have applied techniques from data mining and conclusions are drawn after comparing the results from 2 tools – ROSE2 & WEKA.*

1. **Introduction**

With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are buying products on the Web. In order to enhance customer satisfaction and their shopping experiences, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that they buy. With more and more common users becoming comfortable with the Internet, an increasing number of people are writing reviews. As a consequence, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites. This makes it very hard for a potential customer to read them to help him or her to make a decision on whether to buy the product.

Based on the customer reviews collected from Amazon for the 2 Nokia phones, this project aims at the following:

* To find out important features of interest for a particular product and year
* To compare the results from WEKA (using J48) and ROSE2

1. **KDD Process**

Knowledge Discovery in Databases (KDD) Process refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database. These customer reviews served as our input to the knowledge discovery (KDD) process. The KDD process for our project is defined as shown below:

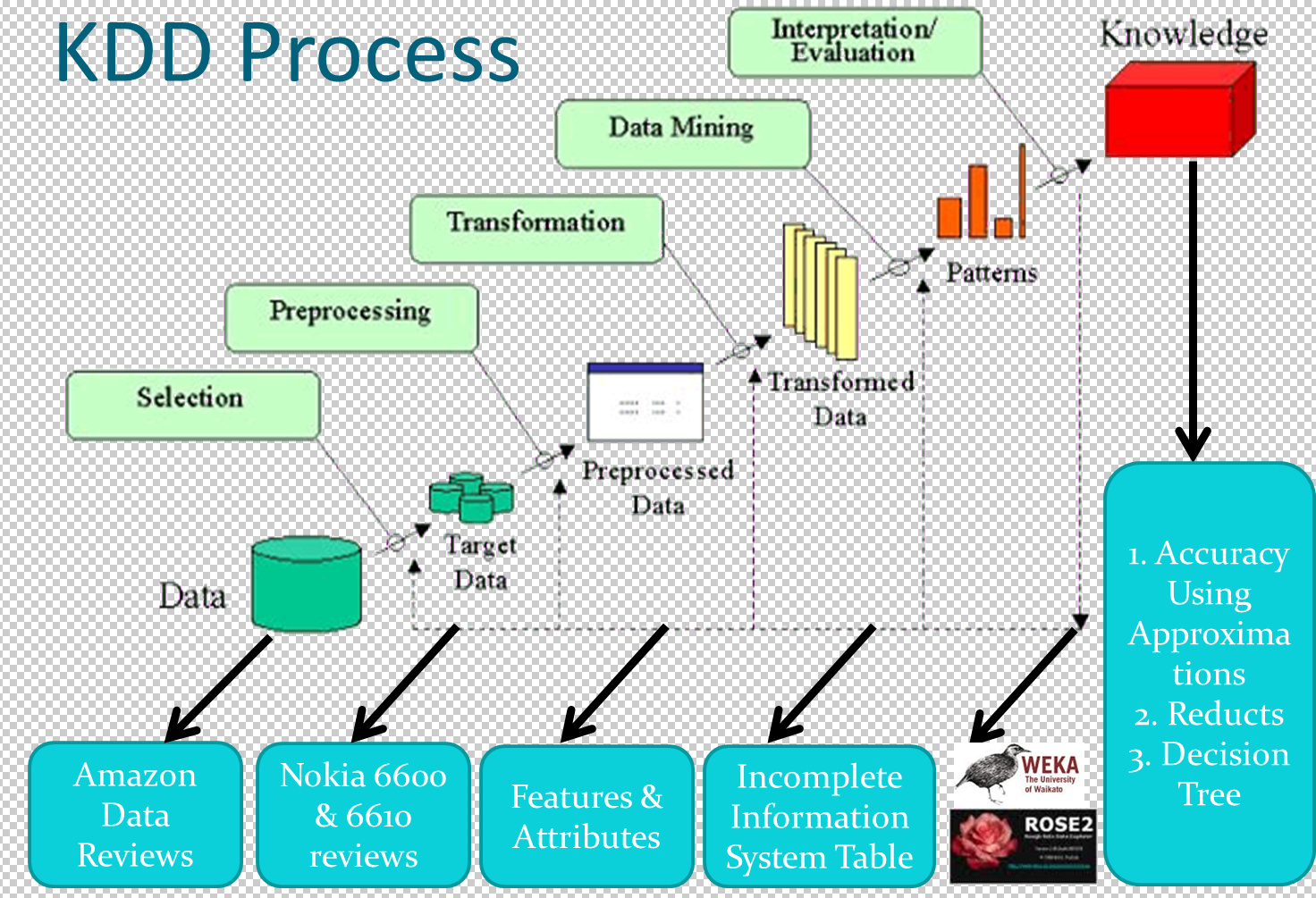
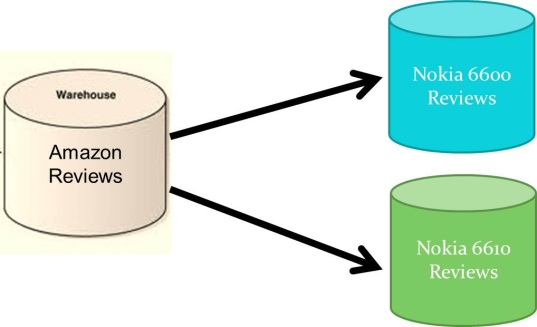
****

Fig 1: KDD process for this project

* 1. **Data Selection**



Number of reviews collected for Nokia 6610= 95

Number of reviews collected for Nokia 6600= 100

Fig 2: Data Selection

* 1. **Preprocessing**

The reviews collected have been cleaned by removing the following manually:

* Stop words
* Emoticons
* Unwanted words
* Unrelated reviews

List of stop words and emoticons:

* Prepositions – in, at, as, besides, before, between…….
* Symbols - ( ) , : # ‘ - . $ + [ ] { } “ ? % @ / \_
* Articles – a, an, the
* Other words – about, all, almost, none, now, rather……

Reviews those were not relevant for e.g.

*“[t]sound good but complicated . ##do too many stuffs to get to a free phone after rebate . ##also , if i want a family plan , which mean i need tw phones on a plan .. .*

*##so how can i do that ... are n't that free for two after the rebates . ##and how can i do that ... . ##print those coupons twice for two phones ?”*

These reviews contained unwanted data and are not related to this project and were hence not considered.

Number of reviews after removing stop words is

For 6610 = 65 For 6600 = 60 reviews

* 1. **Transformation**

We have also collected the feature list for each product from the manufacturer’s specifications list. The attributes for each feature have been defined. Depending on the attribute value for each feature, the entire review has been classified as either positive or negative or neutral.

* Collect the list of features
* Define the domain of the attributes

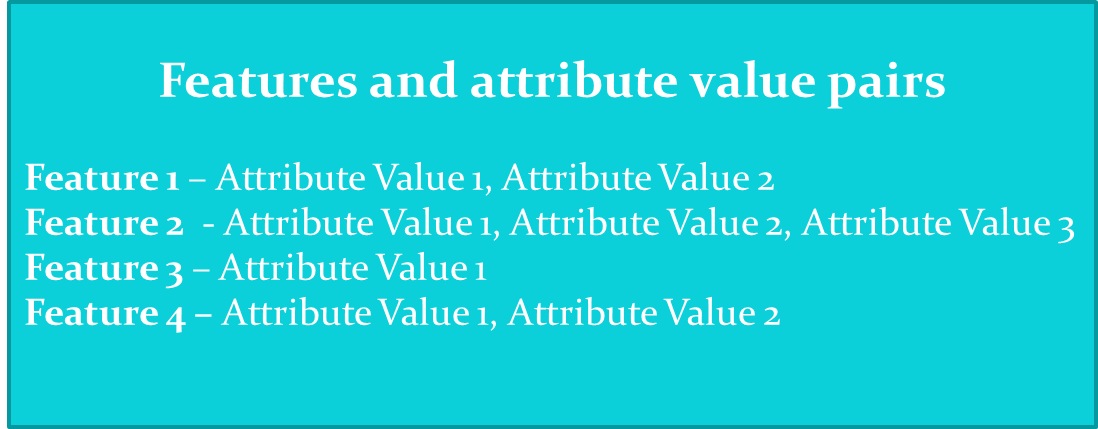


Fig 3: Features and Attribute value pairs

Number of features collected are

For Nokia 6610 – 47

For Nokia 6600 – 52

**List of features and their domain values for Nokia 6610**

**Feature Name Domain of Attributes Feature Name Domain of Attributes**

Phone {Excellent, Good, Bad}

Radio {Excellent, Good, Bad}

Sound\_Quality {Excellent, Good, Bad}

Speaker\_Phone {Excellent, Good, Bad}

Width {Small, Medium,Large}

Depth {Small, Medium,Large}

Height {Small, Medium,Large}

Weight {Light,Medium,Heavy}

Installed\_games {Excellent, Good, Bad}

Platform\_supported {Excellent, Good, Bad}

Screen {Excellent, Good, Bad}

GPRS\_Indicator {Good, Bad}

Connection\_type {GSM, CDMA}

Technology {Excellent, Good, Bad}  
Antenna {Excellent, Good, Bad}  
Signal\_strength {Excellent, Good, Bad}

GPRS\_Indicator {Good, Bad}

Infrared {Excellent, Good, Bad}

Bluetooth {Excellent, Good, Bad}  
Memory {Excellent, Good, Bad}

SMS {Excellent, Good, Bad}

Keypad {Excellent, Good, Bad}

MMS {Excellent, Good, Bad}

GPRS {Excellent, Good, Bad}

Volume {Good, Bad, Average}

Battery\_life {Excellent, Good, Bad**}**

Sim\_card\_lock {Yes, No}

PC\_sync {Yes, No}

Alarm\_clock {Yes, No}

Calendar {Yes, No}

Reminder {Yes, No}

Calculator {Yes, No}

Converter {Yes, No}

Caller\_ID {Yes, No}

Voicemail\_capability {Yes, No}

Voice\_recognition {Yes, No}

Phone\_lock {Yes, No}

Vibrating\_alert {Yes, No}

Call\_timer {Yes, No}

Conference\_call\_capability {Yes, No}

Count\_down\_timer {Yes, No}

Stopwatch {Yes, No}

Multimedia\_features {Yes, No}

Divert\_Indicator {Yes, No}

Silent\_ring\_indicator {Yes, No}

Voice\_message\_waiting {Yes, No}

**Mapping from the features**

The following list gives the mapping definitions that were used to map the words from reviews to the domain attribute values definitions described above

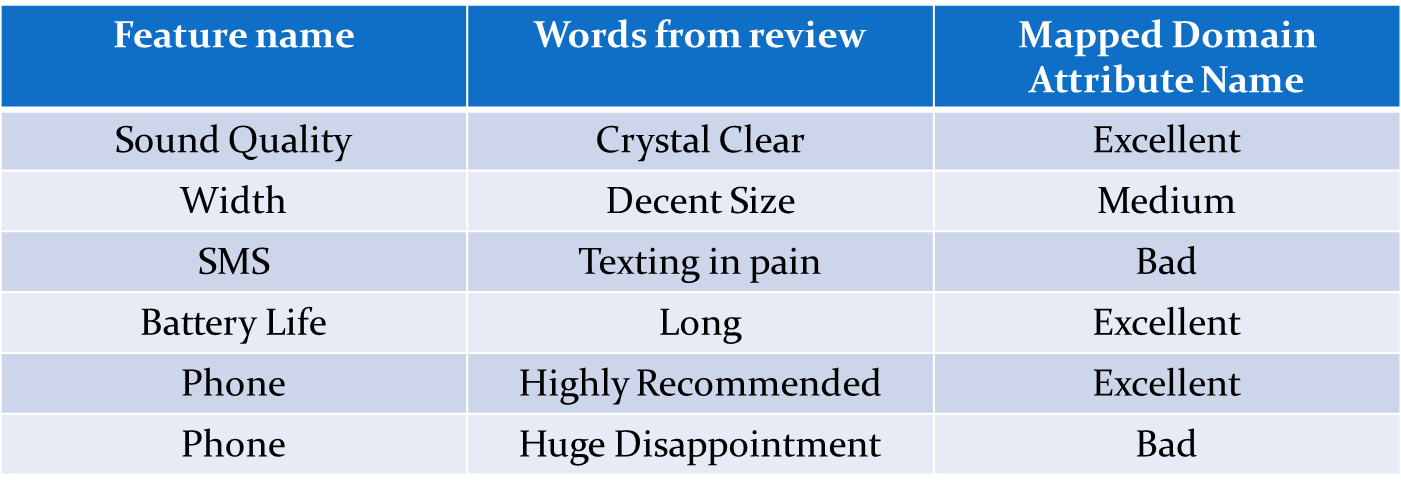


Fig 4: Mapping of the data

Summarizing the mapping list from all features we get:

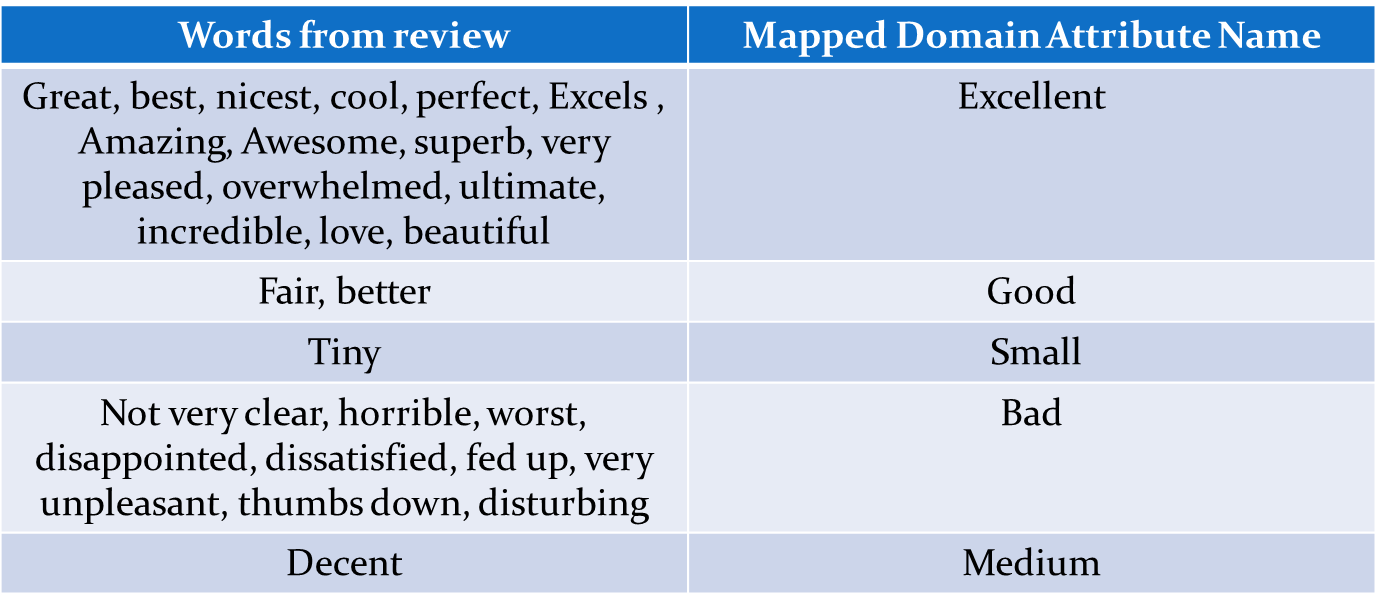


Fig 5: Summary of the mapping of the data

From the reviews, we have extracted review number, the features and their attribute values. This data has been written into an incomplete information system format as shown below where \* denotes a missing value.

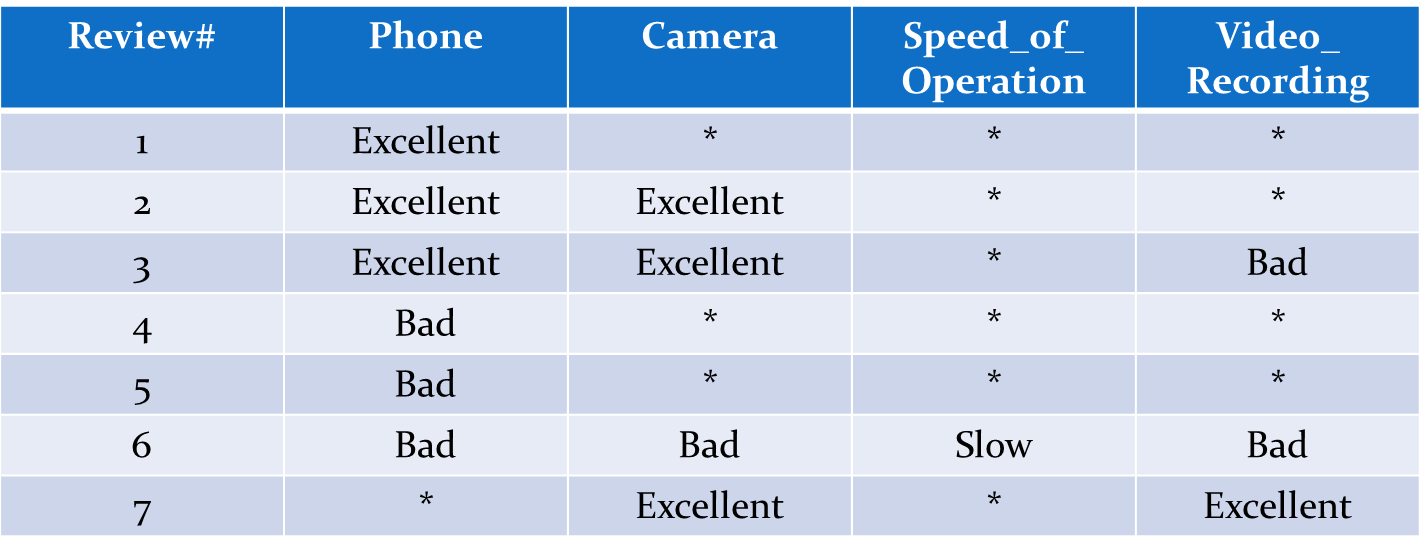


Fig 6: Sample table representing preprocessed data

This data has been transformed into a format so that it can be passed to the data mining tools.

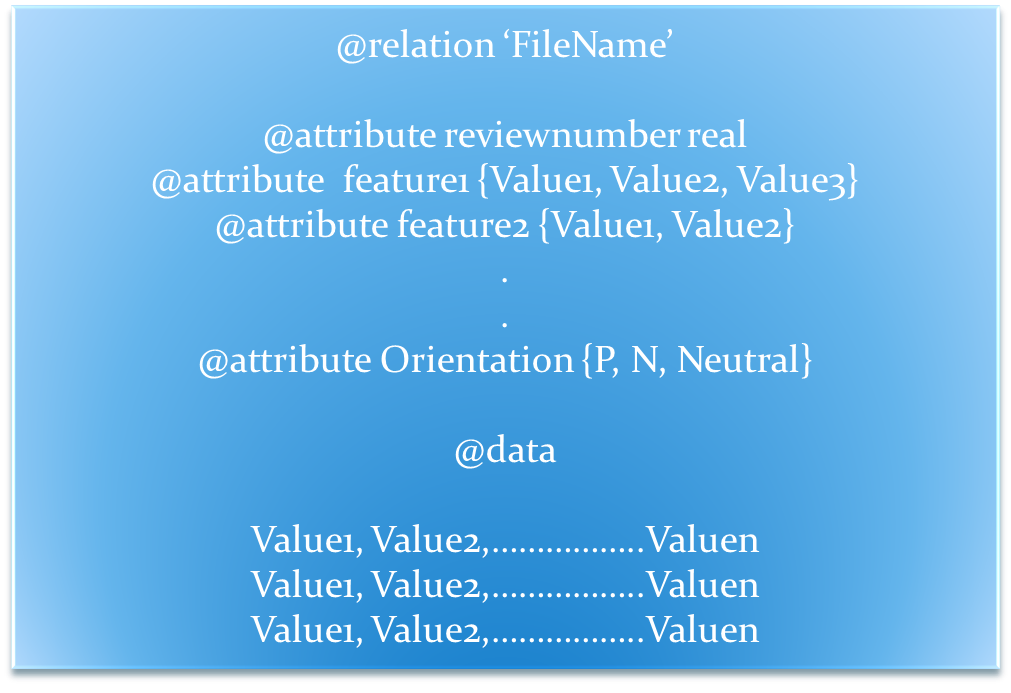
* 1. **Data Mining – Experimental Setup**

We have compared the results based on the performance of the 2 tools – WEKA and ROSE2

* + 1. **Using WEKA**

WEKA stands for Waikato Environment for Knowledge Analysis. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The input file for WEKA should be in a .arff format as shown below. Hence the data collected has been written in format shown below:

  
  
Fig 7: Representation of the input file in .arff format

The following 4 steps need to be followed while processing the data using WEKA:

Step 1: Open WEKA and click on Explorer  
Step 2: Click on Preprocess and Open the file  
Step 3: Click on Classify and select the filter used to classify (J48 in this case). Also, select the cross-validation to 10 folds  
Step 4: Ensure that the decision attribute is selected and click on start to generate the output.

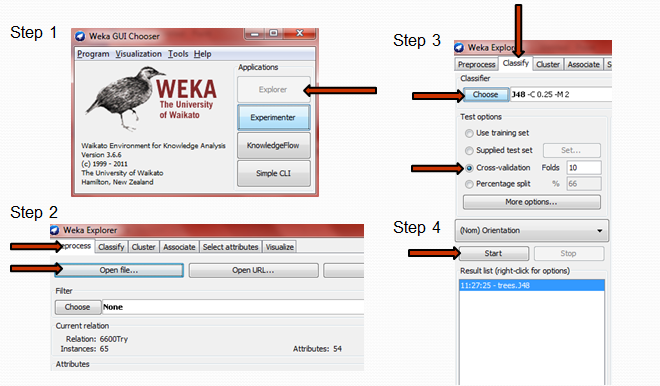


Fig 8: Steps to follow in WEKA

* + 1. **Using ROSE2**

ROSE2 (Rough Sets Data Explorer) is a software implementing basic elements of the rough set theory and rule discovery techniques. The system contains several tools for rough set based knowledge discovery, e.g.:

* Data preprocessing, including discretization of numerical attributes
* Performing a standard and an extended rough set based analysis of data
* Search of a core and reducts of attributes permitting data reduction
* Inducing sets of decision rules from rough approximations of decision classes
* Evaluating sets of rules in classification experiments
* Using sets of decision rules as classifiers.

The input file for ROSE2 should be in a .isf format as shown below. Hence the data collected has been written in the below format. This data has been passed through ROSE2 and we have found the lower approximations, higher approximations and the accuracy along with the reducts. These results have been again presented in the results section.

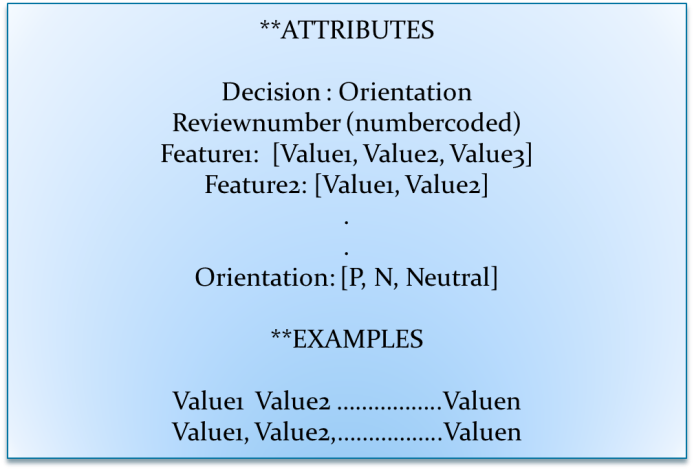
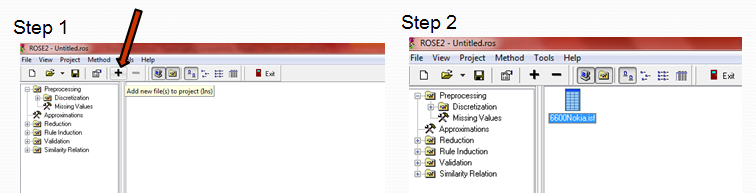


Fig 9: Representation of the input file in .isf format

The steps to be followed using the ROSE2 tool is as follows:

  
Fig 7: Opening the file in the project in ROSE2

To generate the approximations, one should follow the following steps:

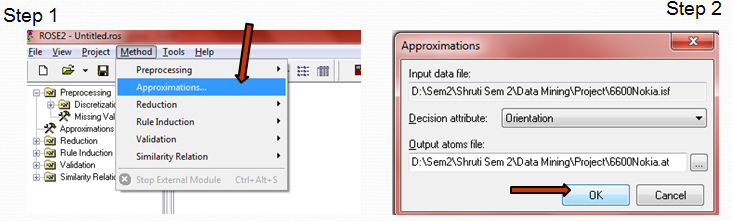


Fig 10: Generating Approximations

To generate the reducts, follow these steps:

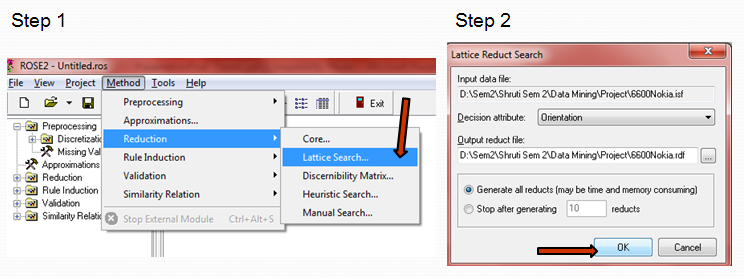


Fig 11: Generating Reducts

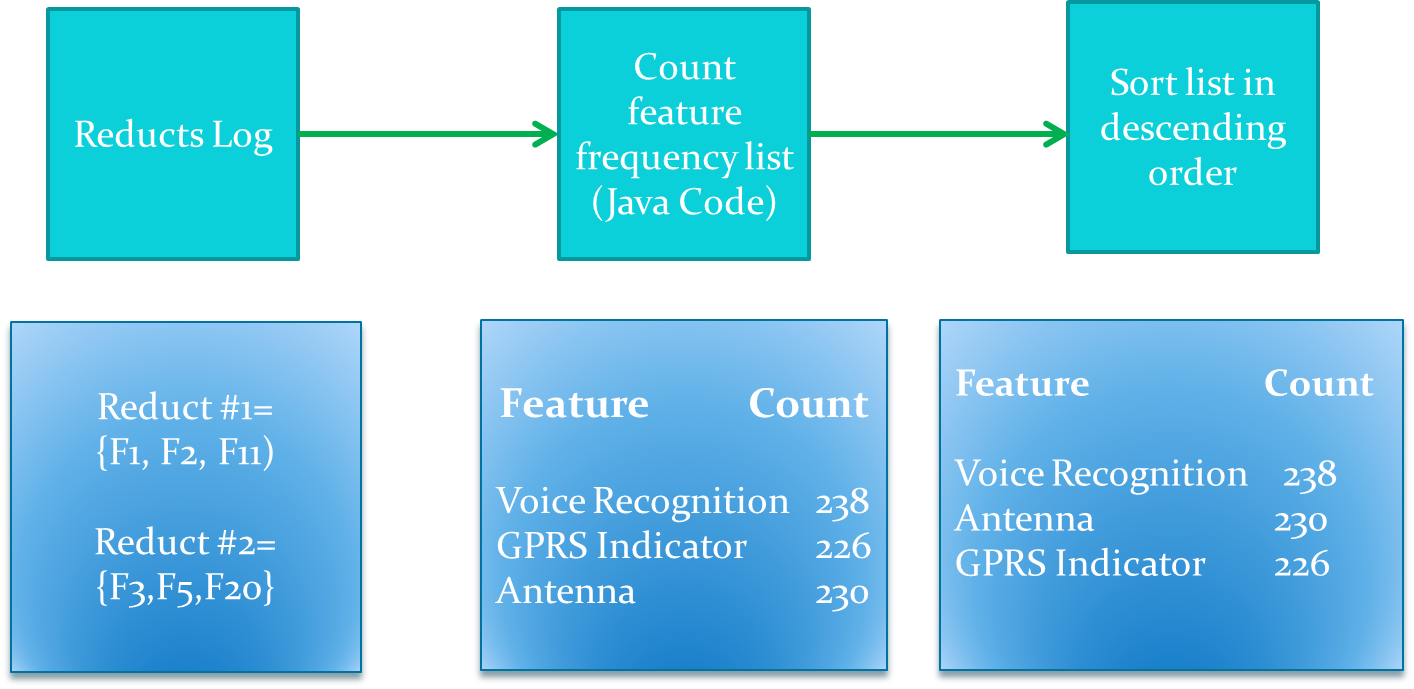
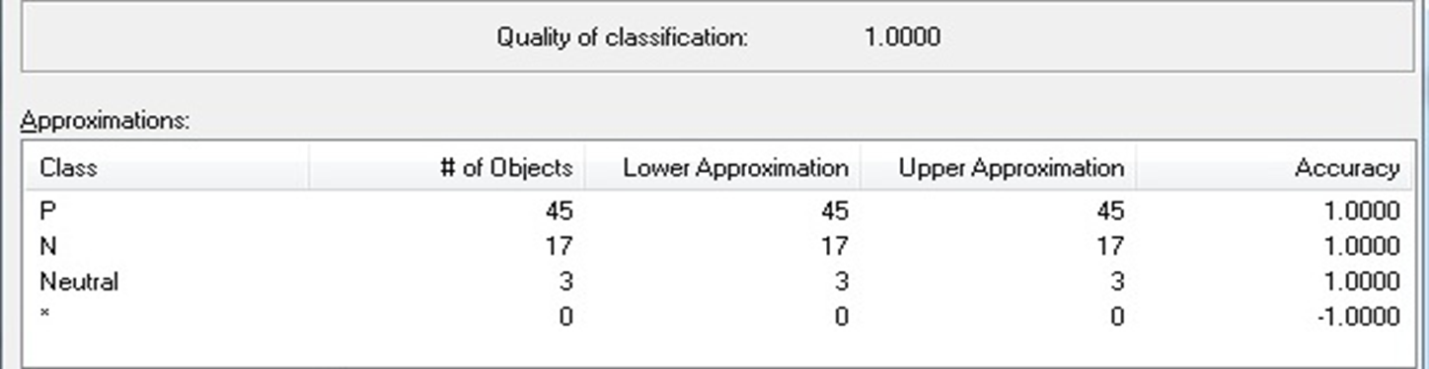
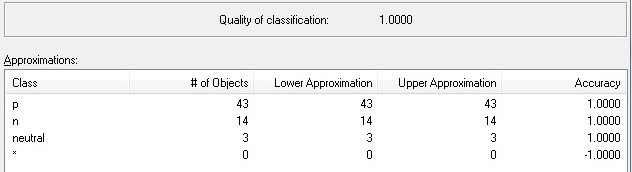


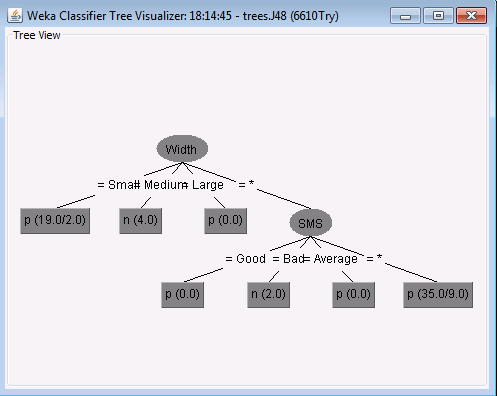
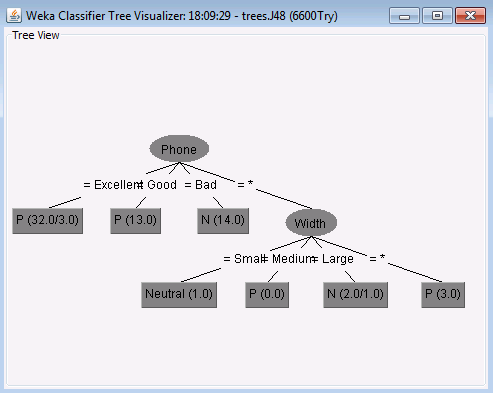
Fig 12: Processing reducts

1. **Results** 
   1. **Approximations using ROSE2**

  
Fig 13: Approximations calculation for Nokia 6610 (2004 Data)

  
Fig 14: Approximations calculation for Nokia 6600 (2009 Data)

* 1. **Decision Tree** **using J48 – WEKA**

Fig 15: Decision Tree for Nokia 6600 Fig 16: Decision Tree for Nokia 6610

* 1. **Feature Frequency in Reducts using ROSE2**

Total no of reducts generated for Nokia 6610 = 301

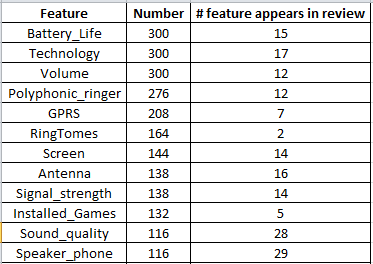
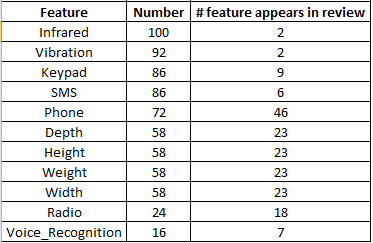
 

Fig 17: Count for Nokia 6610 - #Reducts, # Feature appears in review

Total no of reducts generated for Nokia 6600 = 239

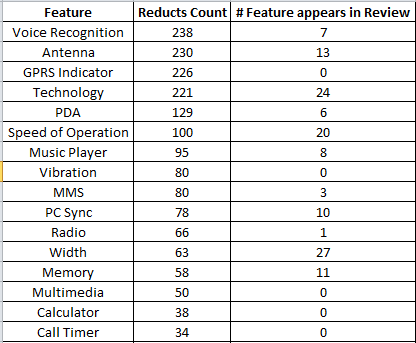
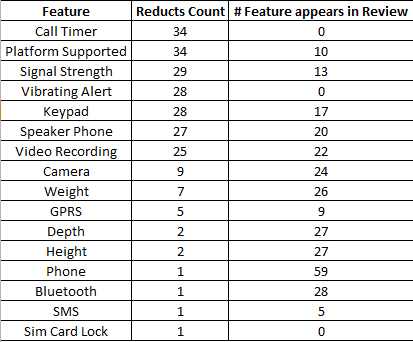
 

Fig 18: Count for Nokia 6600 - #Reducts, # Feature appears in review

1. **Conclusion**
   1. **Nokia 6610 – 2004 data**

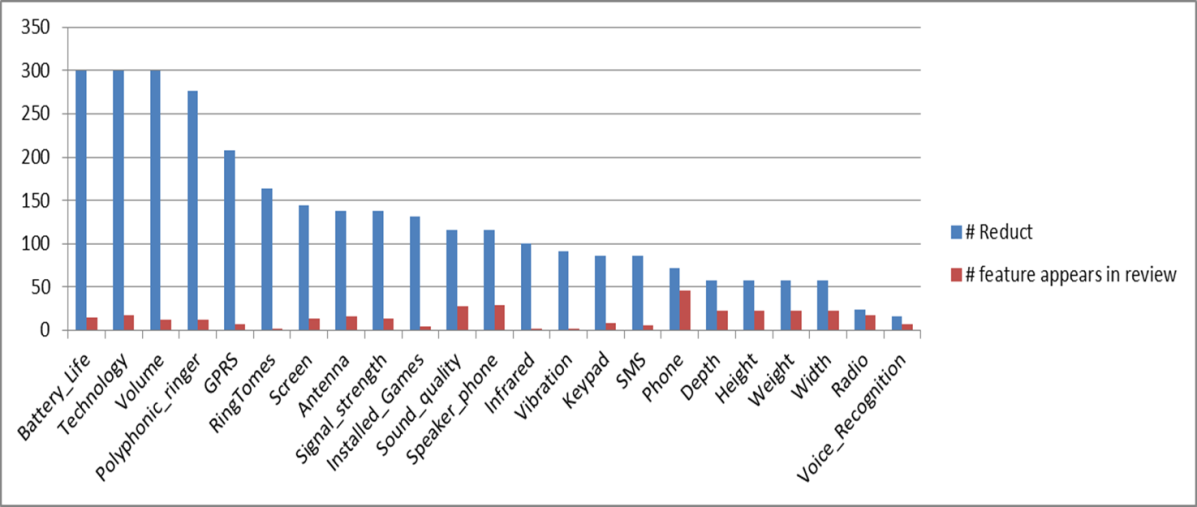


Fig 19: Comparison of the No. of reducts and No. of times a feature appears in a review

* 1. **Nokia 6600 – 2009 data**

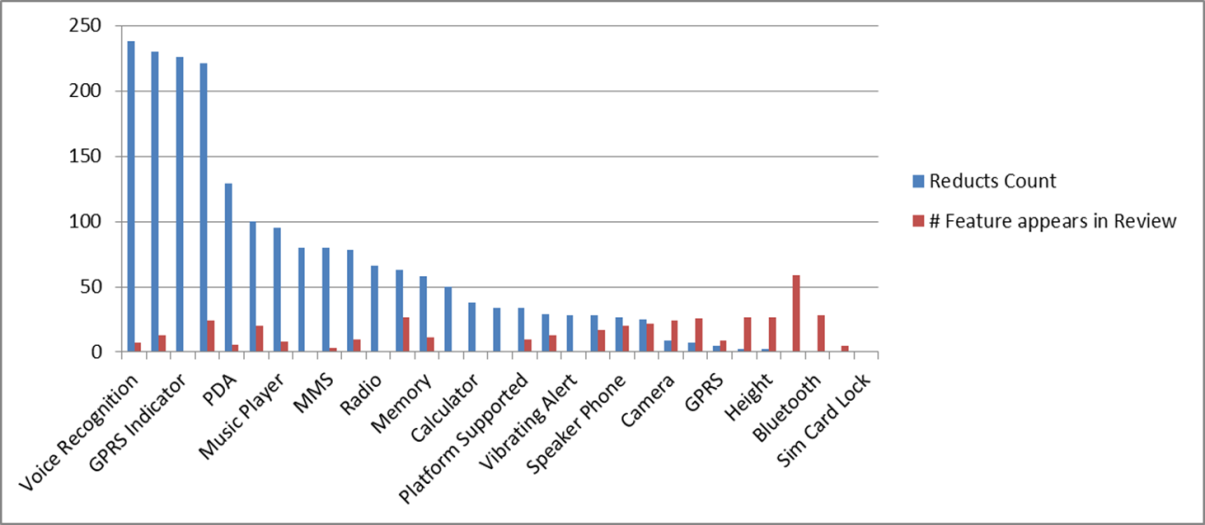
****

Fig 20: Comparison of the No. of reducts and No. of times a feature appears in a review

* 1. **Other conclusions**
* WEKA had identified the following features:

2004 Data – NOKIA 6610 – Width, SMS  
2009 Data – NOKIA 6600 – Phone, Width

* The top few features that ROSE2 identified were:

2004 Data – NOKIA 6610 – Battery Life, Technology  
2009 Data – NOKIA 6600 – Voice Recognition, Antenna

* WEKA produced the Detailed accuracy by class for Nokia 6600 as:

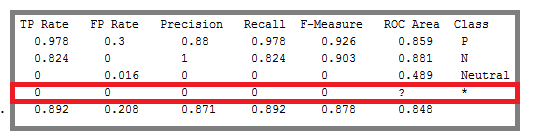


Fig 21: Detailed accuracy by class from WEKA

Hence, we can conclude that WEKA cannot handle missing values for discrete inputs [7]. For discrete values, the missing value becomes a new value. Hence, this operation is not reported to WEKA and the no. of values of the variable is incremented. [7] For continuous values, missing values are replaced with computed average on the data. [7] Since this project had discrete data, the missing values could not be accounted for.

1. **Future work**

This method can be applied to various products so as to see the important features of that product according to the customers.



Fig 22: Future work can be applied to various other products

We have also applied this work to another product – Cannon G3. The table below summarizes all the parameters that were used and also the results from ROSE 2. This can also be experimented and computed using the WEKA tool and comparing the same.

# Reviews collected = 70  
# Features considered = 18  
# Reviews after preprocessing = 45

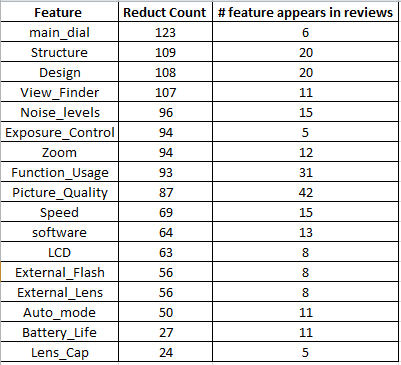


Fig 23: Reduct Count and the no. of times a feature appears in a review

1. **References**

[1] Amazon database [www.amazon.com](http://www.amazon.com), [www.datamob.org](http://www.datamob.org)

[2] ROSE2 tutorial <http://idss.cs.put.poznan.pl/site/rose.html>

[3] WEKA homepage <http://www.cs.waikato.ac.nz/ml/weka/>

[4] KDD process <http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html>

[5] Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In Proceedings of the 19th national conference on Artifical intelligence (AAAI'04), Anthony G. Cohn (Ed.). AAAI Press 755-760.

[6] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 168-177. DOI=10.1145/1014052.1014073 <http://doi.acm.org/10.1145/1014052.1014073>

[7] <http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Handle_WEKA_File.pdf>