Paper 241-2013
Feature Extraction and Rating of a Smartphone Photosharing Application using SAS Sentiment Studio®

Anil Kumar Pantangi, Oklahoma State University, Stillwater, OK, USA.
Siddartha Reddy Mandati, Oklahoma State University, Stillwater, OK, USA.
Sahithi Ravuri, Oklahoma State University, Stillwater, OK, USA.
Dr. Goutam Chakraborty, Oklahoma State University, Stillwater, OK, USA

ABSTRACT

Smartphone users often have to read many reviews available online to know about a particular feature of an application. Online reviews often provide a numeric overall rating using the Likert or semantic scales but these do not fully reveal sentiments of customers about a product's features. For this project we have considered the website Google play, which is a dedicated portal for all Android paid/free applications. Google Play captures both structured and unstructured reviews generic to each application by Helpfulness, Rating and Newest. This leaves the user with no clue about a particular feature and its rating. In this paper we illustrate feature extraction and feature ranking of the Photosharing application using SAS Sentiment Studio®. The reviews are manually categorised as Positive and Non-Postive by multiple experts. Sentiment Studio is used to predict a new review as either Positive or Non-positive. Further to extracting features of the application, we have also built in manual rules in the Rule-Based model. In this data, rule-based model outperform the statistical model or the hybrid model in a test data set. The best model helps us in categorizing each review of the application by its feature along with the rating. Marketing insight companies may deploy this model as a plugin to the social networking sites, review boards and micro blogs to categorize and rank the content.

INTRODUCTION

Consumers often install a free application and try it out to explore its features and then uninstall it if not satisfied. When it comes to paid applications, users want to buy only the application that meets the expectations. To evaluate an application one has to visit many websites and blogs, read the reviews and ratings, which provide generally mixed opinions on any application. The specifications of an application when used to categorize valuable information from the users could help end users and developer communities to assess the quality of an application. This comes in handy for an end user who is searching around for an application that matches his/her interests.

We have considered a popular photo-editing tool for Android to analyze the user reviews and categorize them according to its specifications by using SAS® Sentiment Studio. This helps users distinguish between applications based on categorizing and rating user reviews according to the application's attributes. Overall average ratings of an application will not always help validate the quality of an application, as there may be many factors that contribute to the overall numeric rating.

DATA PREPARATION

Comments posted by 808 users were collected during the period of first and second weeks of October 2012. The collected data is a stratified sample based on the likert ratings of 1 to 5. More than 70% of the comments on a single day are of scale 5.
This indicates more positive polarity of the customer reviews. For the sentiment mining, the comments are manullay rated by the experts as Positive and Non-positive. Experts have rated 380 of the total comments as Non-postive.

Column	Description
Review ID	Unique ID for each Review
Comment	Text Comment from Google Play
Google Rating	Available for each comment on Google Play on the scale of 1 to 5
Expert Rating	Comments rated by Experts

Table 1: Data Description

METHODOLOGY
As part of this research project we have built a Statistical model, Rule-Based model and Hybrid model on the textual data classified by rating – Positive and Non-positive. The best model identified in this analysis is used for testing.

STATISTICAL MODEL
SAS® Sentiment Analysis Studio automatically sets aside a percentage of training documents for validation purposes to adjust the parameters of training results. These adjustments are based on the results obtained from the validation corpora of documents. Based on the precision requirements specified, a user has a choice of building a simple or advanced Statistical model.
For this project we have built a basic Statistical model and configured the settings to derive the best model from the training models provided. Simple model applies four combinations of models with smoothed relative frequency as the text normalization and combined with no feature ranking, Risk ratio, Chi-square and information gain.
Model is built from the training documents by taking term frequencies contributing to the weights of the terms and validation data is used to fine tune the model for increased accuracy. The terms that occur in the training corpora are learned by their contributing weights and applied to the testing data to predict the sentiment of the document as Positive or Non-positive. The Statistical model predicts the sentiment for the overall document but not at the feature or the attribute level. Feature level sentiment prediction can only be accomplished using Rule-Based models.
The Statistical Model is built with 80% of provided documents used for training and 20% for validation. Probability cut off is set to 50%. From the results the best Model is selected to be the Bayes Model.

With text normalization [Relative Frequency] and feature ranking algorithm [Chi Square]
      Overall precision: 71.24%
      Positive precision: 58.02%
      Negative precision: 86.11%

With text normalization [Relative Frequency] and feature ranking algorithm [Information Gain]
      Overall precision: 72.55%
      Positive precision: 60.49%
      Negative precision: 86.11%

Best feature ranking algorithm for text normalization [Relative Frequency] is [Information Gain]

With text normalization [Okapi BM25] and feature ranking algorithm [No Feature Ranking]
      Overall precision: 68.63%
      Positive precision: 62.96%
      Negative precision: 75.00%

With text normalization [Okapi BM25] and feature ranking algorithm [Risk Ratio]
      Overall precision: 65.36%
      Positive precision: 69.14%
      Negative precision: 61.11%

With text normalization [Okapi BM25] and feature ranking algorithm [Chi Square]
      Overall precision: 71.24%
      Positive precision: 55.56%
      Negative precision: 88.89%

With text normalization [Okapi BM25] and feature ranking algorithm [Information Gain]
      Overall precision: 70.59%
      Positive precision: 54.32%
      Negative precision: 88.89%

Figure 1: Training results for the Statistical model

Relative Frequency Information Gain

Figure 2: Graphical results for the Statistical Model

TEST FOR THE STATISTICAL MODEL
A stratified sample of 50 documents was considered for testing the models.
The Precision in scoring came out about 90%. Below are the screenshots of a few scored documents.

Documents correctly predicted as Non-Positive:

I don't know if it's only the Droid x2, but **whenever** I put a **pic** up, it'll keep taking me **back** to the **area** where I **crop** the **pic** and I have to hit 'discard' for it to take me **anywhere**. **Also** after doing that a while, I'll hit the app and it'll take me to my contacts **list**.

| Text Result | Graphical Result |

Overall Document: Test in statistical-based model result is Negative
Probability to be positive is 0.00% with confidence 100.00%

Figure 3: Test results of the Statistical model for a document predicted

Documents correctly predicted as Positive:

A new and **fun spin** on the average **social** networking sites…Puctures take u places that reading a blog or posts never kould…its **easy** to use and I would **definitely recommend** this **site** to others

| Text Result | Graphical Result |

Overall Document: Test in statistical-based model result is Positive
Probability to be positive is 100.00% with confidence 100.00%

Figure 4: Test results of the Statistical Model for a document predicted

| Text Result | Graphical Result |

Sentiment Distribution
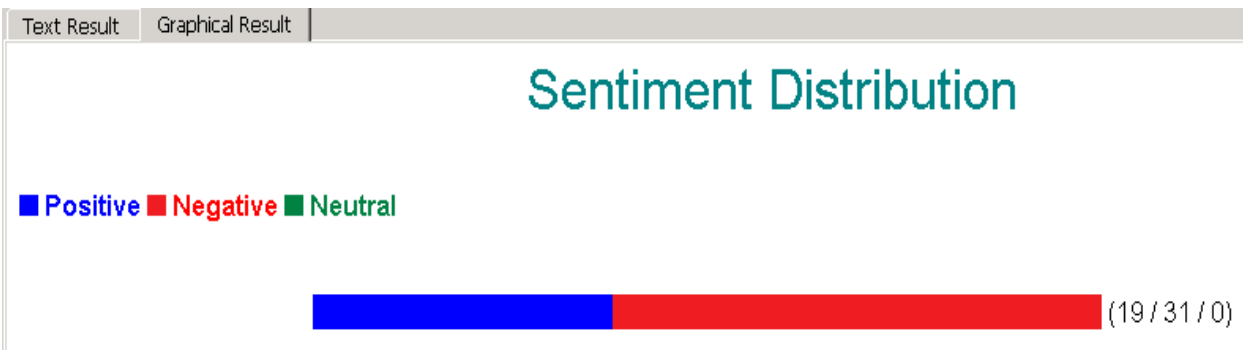
■ Positive ■ Negative ■ Neutral

(19 / 31 / 0)

Figure 5: Overall results of the Statistical Model on Testing Documents

The above results from the Statistical model testing on the scoring data has come out quite reasonable with 19 Positive comments out of 24 are correctly predicted with an accuracy being 79% and 26 Non-positive comments out of 26 are correctly predicted with an accuracy of 100%. However the model has 20.8% of false negative rate (Type || error).

4

RULE-BASED MODEL BUILDING
The Rule-based model is more sophisticated compared to the Statistical model. This model allows us to write custom rules along with the rules learned from the Statistical model. Rules learned from Statistical model are imported into the Tonal Keywords section with Positive and Non-positive keywords. The Rule-Based model provides facility for us to define products and features specific to our project to identify sentiment at the granular level or feature level. All the rules imported into Tonal Keywords are classified as "Classifier" by default. The classifier type of rule consists of a word or a string with a weight assigned. When a word or string matches with the rules in either Positive or Non-positive tonal keyword, the entire word or string is highlighted in the test result to indicate a match. A matching word with positive tonal keyword is indicated in green color and matching word with Non-positive tonal keyword is indicated in red color. There are many other rules like "CONCEPT", "C_CONCEPT", "CONCEPT_RULE", "PREDICATE_RULE" and "REGEX" that can be used depending on the depth of granularity required in the analysis. We have used the basic "Classifier" type of rules for this project. This capability can be taken advantage of when the scoring documents have instances that match the products and features defined.

FEATURES IDENTIFICATION
We have defined PhotoApp as a Product; Editing, Photosize, Updating and General are the features for the product PhotoApp. Different terms that identify the features are defined in the definitions portion of each feature.
Definitions for these features have to be included in the Rule-Based model to identify the terms that capture the attributes of a feature. Terms attributing to a feature are identified from the Tonal Keywords of training documents.

Feature	Attributes that capture the definition of a feature
Effects	Editing, edits, filters, sharpness, framing, picture quality, picture tools
Photosize	Picture size, photo size, crop, size, square, large, small
Updating	Update, upload, refresh, force close, fix, share, problem
General	Follow, beautiful, keep up, simple, addict, great, amazing, fun, easy

Table 2: Feature definitions

TEST FOR RULE-BASED MODEL
Probability threshold determines the overall sentiment of a document processed and a score to that document, which is used for scoring a document as Positive or Non-positive. We used the default setting of 50% which classifies a document as Positive if the score exceeds 50%, else the document would be classified as Non-positive.
The relative weight of positive rules in rule-based model indicates the importance of Positive rules over the Non-positive rules. The default value of 100% is being used for Relative weight of positive rules in Rule-Based model to ensure both Positive and Non-positive rules are treated with equal importance. Testing process is a repetitive task to ensure accuracy in scoring results of Positive and Non-positive terms to be identified correctly. Rules might have to be edited to ensure accuracy.
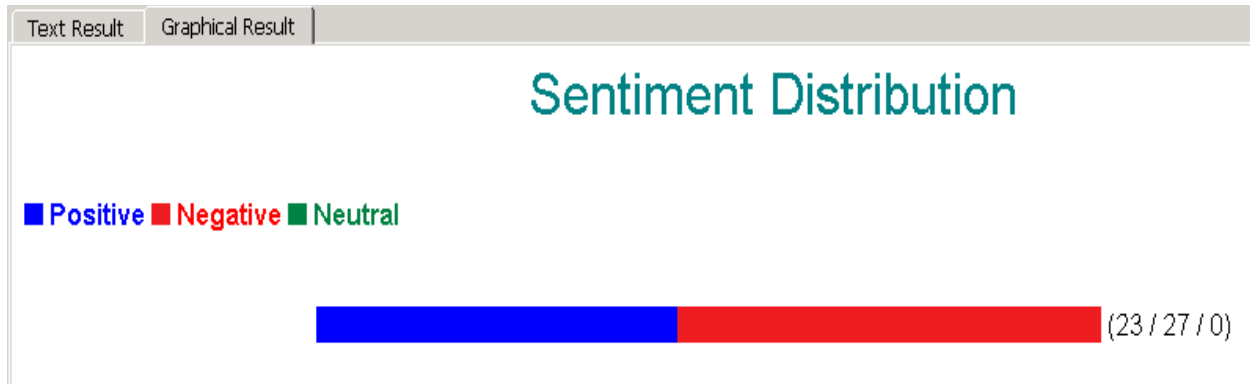
Figure 6: Overall results of Rule Based Model on Testing Documents

Testing the Rule-Based model has also given excellent results with 23 out of 24 Positive comments being correctly predicted with an accuracy of **96%** and 26 out of 26 Non-positive comments being correctly predicted with an accuracy of **100%**. Rule-Based model has a decreased false negative rate (Type || error) of **0.5%.**

HYBRID MODEL
The Hybrid model is a combination of the Rule-Based and Statistical models by applying the rules developed in Rule-Based and the numerical data from the activated Statistical model. This helps in extracting the sentiment at the feature level of the applications. The test configuration for the Hybrid model is set with default settings with probability threshold being 50%, relative weight of positive rules in Rule-Based model being 100% indicating equal importance to Positive and Non-positive rules, and weight of the Statistical model in hybrid model being 70%.

TEST FOR HYBRID MODEL
Testing the Hybrid model has given reasonable results. 19 out of 24 Positive comments are correctly predicted with an accuracy of 79%. Figure 7 illustrates with colors on how a hybrid model captures the rating Positive or Non-Positive along with the feature attributes in a document. Positive term is identified with a green color, Non_positive term is identified with a red color and pre-defined feature is identified with a blue color.

Example 1
Comment which has positive rating about the feature 'Effects'

Its awsome i love sharing pictures..only they should add more effects like to edit :D

Example 2
Comment which has Non-positive rating about the feature 'Cropping'

I dont like the cropping but everything else is easy to use and it has cool filtering effects.

Example 3
Comment which has positive rating about the feature 'Uploading/Effects'

I love this app. ! The only thing i hate about it, is when uploading/editing a picture, you have to crop the picture so it fits in the box. I suggest the PhotoApp techs or whoever allow more cropping area. Allow the crop to go up, down, and sideways more. Other than that, i LOVE the app. ! #PhotoAppFanatic. !!

Example 3
General Comment which has positive rating.

Needs to not make us crop our photos. That's a bit annoying because I don't take a picture of something beautiful to only pick out the good parts. It is all beautiful

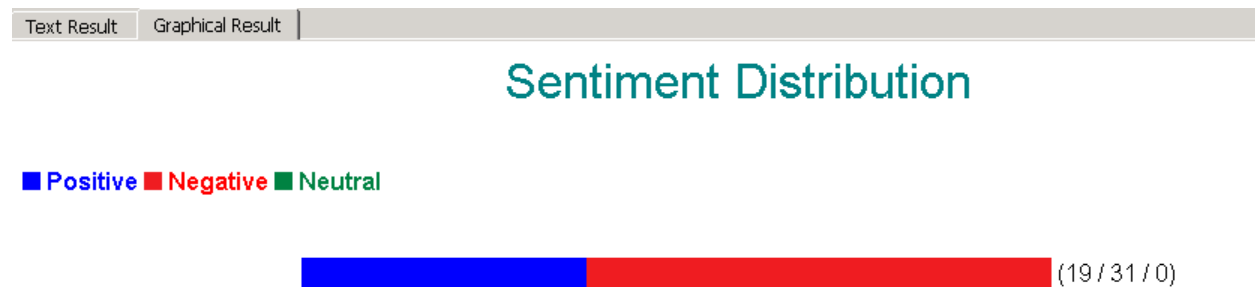**Figure 7: Feature level sentiment detection using Hybrid model**

Text Result   Graphical Result

Sentiment Distribution

■ Positive ■ Negative ■ Neutral

(19 / 31 / 0)

Figure 8: Overall results of the Hybrid model on Testing Documents

MODEL COMPARISON

Rule based model is the best model for this data with a positive precision of 96%. Table 3 shows the positive precision, negative precision and False Negative rate of each of the models employed for this project. The best model can be deployed to find out the customer sentiment and extract the feature from the reviews.

Model	Positive Precision	Negative Precision	False Negative Rate
Statistical	79%	100%	20.8%
Rule-Based	96%	100%	0.5%
Hybrid	79%	100%	20.8%

Table 3: Comparison of the models implemented in SAS® Sentiment Studio

CONCLUSION

Sentiment Mining is a good way to categorize the text reviews data and identify the likely drivers of the ratings. When these identified features of the review data are implemented as the attributes of an application in the reviews filter section, it becomes easier for the end users to pick the smart phone applications based on their interests and needs.

In this paper we are able to evaluate the ratings and extract the features of the application. With fair domain knowledge, the Rule-Based model can further be fine-tuned by incorporating the rules provided by application developers. The best model can be used to identify the features and their ratings. This gives a competitive advantage to the businesses to showcase the sentiments of the users about the smart phone applications. Commercially this model can be used by the marketing insight companies to categorize the content and evaluate the comment when it is posted by the customer.

This SAS® sentiment Studio provides a good interface for an analyst to get a report on the feature based ranking in terms of number of Positive and Non-positive comments (or sentiment expressed) for each feature for the specified product. Monitoring the comments in real time will enable the developer community to be updated and take necessary actions. The user community will benefit in being updated with the current status of the application with respect to a particular feature of interest.

REFERENCES

[1] Pantangi and Chakraborty. 2012. "#104-2012 Classification of Customers' Textual Responses via Application of Topic Mining". Proceedings SAS® Global Forum 2012.

[2] Dobson. 2010. David Dobson, Dobson Analytics Inc., Segmenting Textual Data for Automobile Insurance Claims. Proceedings SAS® Global Forum 2010.

[3] Miller, W.T. 2005. Data and Text Mining-A Business Applications Approach. Pearson Prentice Hall.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Anil Kumar Pantangi, Email: anil.pantangi@okstate.edu

Anil Kumar Pantangi is Master's student in Management Information Systems at Oklahoma State University. He is a BASE SAS® 9 and a certified SAS® predictive modeler using Enterprise Miner 6. He has received student scholarship from SESUG'12, Durham and was selected for Honorable mention from WUSS'12. He received his SAS® and OSU Data Mining Certificate. He has worked with IBM & Wipro Technologies as an ITSM Consultant.

Siddartha Reddy Mandati, Email: siddhartha.mandati@okstate.edu

Siddartha Reddy Mandati is Master's student in Telecommunications Management at Oklahoma State University. He is a SAS Certified BASE programmer for SAS® 9, JMP certified data explorer and SAS® certified predictive modeler using SAS® Enterprise Miner 6. He is currently in the SAS® and OSU Data Mining Certificate program.

Sahithi Ravuri, Email: sahithi.ravuri@okstate.edu

Sahithi Ravuri is Master's student in Management Information Systems at Oklahoma State University. She is a certified JMP data explorer and Oracle certified DBA. She has worked as a session coordinator at SAS Global Forum 2012, Orlando. In May, 2012, she received her SAS® and OSU Data Mining Certificate. She has worked with Capgemini Consulting as a Software Consultant.

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business

Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

TRADEMARKS
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies**.**