Feature Extraction and Rating of a Smartphone Photosharing Application using SAS Sentiment Studio®

Siddhartha Mandati, Anil Pantangi, Sahithi Ravuri, Dr.Goutam Chakraborty

Oklahoma State University, Stillwater,OK,USA

ABSTRACT

Online reviews of smartphone applications 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. Google Play captures both structured and unstructured reviews generic to each application. 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®. Sentiment Studio is used to predict a new review as either Positive or Non-positive. We 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.

INTRODUCTION

Consumers often consider various factors and buy only the application that meets the expectations. 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.

DATA PREPARATION

A stratified sample of 808 comments was collected in October 2012 based on the likert ratings of 1 to experts as Positive and Non-positive.

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

We have built a Statistical model, Rule-Based model and Hybrid model on the textual data classified by rating – Positive and Non-positive. 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.

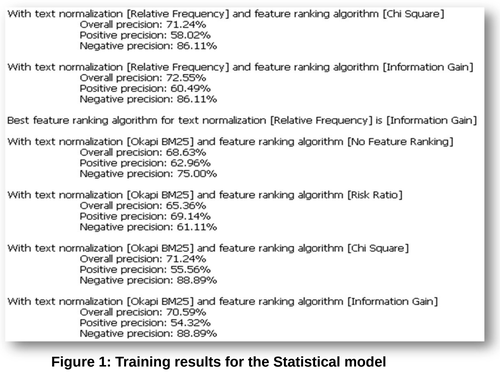

Figure 1: Training results for the Statistical model

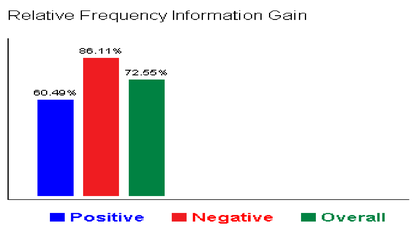

Figure 2: Graphical results for the Statistical Model

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:

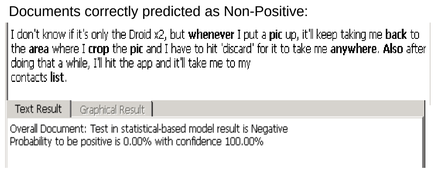

Figure 3: Test results of the Statistical model for a document predicted

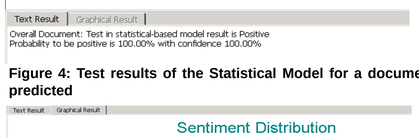

Figure 4: Test results of the Statistical Model for a document predicted

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

Statistical model test results on the scoring data 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).

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. 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.

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.

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.

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

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.

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.

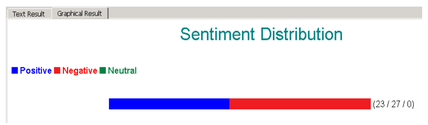

Figure 6: Overall results of Rule Based Model on Testing Documents

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.

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 2

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

Example 3

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

Figure 7: Feature level sentiment detection using Hybrid model

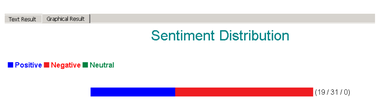

Figure 8: Overall results of the Hybrid model on Testing Documents

MODEL COMPARISON

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

Conclusions

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.

Acknowledgments

The authors would like to thank Dr. Goutam Chakraborty for his invaluable supervision and support throughout this research

Contact Information

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

Name: Siddhartha Reddy Mandati
Company: Oklahoma State University
E-mail: siddhartha.mandati@gmail.com

Name: Anil Kumar Pantangi
Company: Oklahoma State University
Email: anil.pantangi@okstate.edu

Name: Sahithi Ravuri
Company: Oklahoma State University
Email: sahithi.ravuri@okstate.edu

Name: Goutam Chakraborty
Company: Oklahoma State University
Email: goutam.chakraborty@okstate.edu