PROTOTYPING A WORDSTAT/QDA MINER AUTOMATIC DOCUMENT CLASSIFICATION MODEL

DECEMBER 10, 2018

DATANALYTICS
Email: DatAnalytics@iinet.com.au
Webpage: https://dat-analytics.net/
PROTOTYPING A WORDSTAT/QDA MINER AUTOMATIC DOCUMENT CLASSIFICATION MODEL FOR PRODUCT REVIEW

KEYWORDS

Sentiment Analysis; WordStat/QDA Miner; Machine Learning; Automatic Document Classification Model; k-Nearest Neighbour, Similarity Index; Co-occurrence; R

AIMS

To prototype a WordStat/QDA Miner Automatic Document Classification Model for Sentiment Analysis.

INTRODUCTION

Automatic Document Classification is a form of supervised Machine Learning through Natural Language Processing (NLP). Its key aim is to assign a document to one or more classes or categories. Automatic Document Classification is widely used these days to catalogue emails, news articles, and in auto-tagging customer queries and product reviews.

Machine Learning Document Classification employs a variety of classifiers such as Support Vector Machine (SVM), k-Nearest Neighbours (k-NN) and Naïve Bayes. The last two methods are employed in WordStat/QDA Miner.

METHODS

Dataset

Customer ratings /sentiments reported in the Amazon-reviews-unlocked-mobile-phones dataset (by PromptCloud (2016) was studied. Data comprised 400 thousand reviews and 5 ratings /sentiments (namely terrible, bad, neutral, good and excellent), which reflected the users’ experience with the unlocked mobile phones.
WordStat/QDA Miner

We developed the following methodology for data pre-processing:

- Wrote a script in R programming language to:
  - Generate a scalable and automatable method to generate a representative, random sampling method that ensured the proportions of ratings was maintained.
  - Recode the original five ratings (terrible, bad, neutral, good and excellent) into three ratings (terrible, neutral and excellent) by collapsing the ‘terrible’ and ‘bad’ into a ‘terrible’ rating and by collapsing the ‘good’ and ‘excellent’ into an ‘excellent’ rating.
  - Extracted two separate and representative, 10% samples:
    - Machine Learning and validation (or testing) sample; and
    - Unknown sample
- Undertook Machine Training/Learning and validation in QDA Miner and WordStat, as previously reported by Nyakuengama (2018):
  - Created a classification model (*.wclas) in WordStat:
    - Pre-processing parameters:
      - Lemmatization: English
      - Stemming: 4-grams
    - Post-processing options in WordStat:
      - Add words with frequency (30)
      - Keep a maximum of 3,000 items based on TF*IDF
    - Classification method in WordStat:
      - k-Nearest Neighbour (k-NN) with
        - Leave-one-out
    - Obtained classification performance statistics in WordStat:
      - Precision
      - Accuracy
      - Recall.
  - Applied the WordStat classification model on the an unknown sample in QDA Miner and saved the encoded ‘Unknown’ sample.
  - Imported the encoded ‘Unknown’ sample back into WordStat, applied the sentiment model (WordStat Sentiment.wmodel) on top and saved the classification codes.
  - Undertook some crosstab analyses (i.e. clustering and heatmaps and correspondence plots) and co-occurrence analyses (i.e. dendrogram, mapping and proximity plots).
RESULTS

WordStat/QDA Miner Machine

We experimented with different model settings – method, model parameters, number of features, selection and testing methods, as shown below:
This image shows that performance metrics of this classification model which was developed using the k-NN method were excellent in terms of accuracy, precision and recall. This case-occurrence classification model was saved and used to subsequently classify new cases in QDA Miner.
The QDA Miner image below shows the coding frequency of the test document, after applying the WordStat classification model developed above. It suggests that of the 3,096 codes in the document, 545 (17.6%) had been encoded as ‘terrible’; 32 cases (1%) were encoded as ‘neutral’ and 2,519 (81.4%) were encoded as ‘excellent’. Other key text statistics, the percent cases and words, are shown.

The image below is an example of a case encoded as ‘terrible’.
The image below is an example of a case encoded as ‘neutral’.
The image below is an example of a case encoded as ‘excellent’.

```plaintext
I bought this refurbished, and you could never tell that it wasn’t brand new. Saved a couple bucks, and made the kid happy for his birthday! He’s had so much fun with the Apple Watch, that I went the next weekend and bought myself one... hence all the reviews about the watch band(s) I have posted.
```
Cross-checking against a sentiment dictionary

The QDA Miner image below shows coding co-occurrences, with the WordStat sentiment model (Sentiment.wmodel) applied. The first three codes from the left (i.e. 1 - Terrible, 3 - Neutral and 5 - Excellent) were developed in our current classification model. The rest of the codes were from the WordStat Sentiment.wmodel.

Overall, we find excellent agreement between our classification model and the Sentiment.wmodel. In particular, see that:

- 1 - Terrible correlated well with negations, negative words and real_bad;
- 3 - Neutral correlated equally with negations, negative words, not bad and not good; and
- 5 - Excellent correlated well with positive words and real_good.
The image below is a 2D graphical representation of co-occurrences, portraying the above noted results.
The image below shows coding co-occurrences also echoes the co-occurrence results:

Proximity plots
Below are three proximity plots for 1-Terrible, 3-Neutral and 5-Excellent encoded cases. The also agree with the above results.
DISCUSSION

- This study successfully designed, implemented and applied a WordStat/QDA Miner automatic document classifier for use in the unblocked mobile phone data. It focused on 3 key ratings and demonstrated that this package can be used reliably.
- During the study we found that:
  - Model building was moderately fast but required considerable computer processing power. This suggests the need for good strategic planning in data preparation and presentation for analysis in WordStat/QDA Miner.
  - Performance measures of our automatic document classification model were excellent (i.e. accuracy, precision and recall).
  - Most importantly, predicted results from the model were very reliable when measured against a known sentiment model during the following analyses:
    - Code co-occurrences;
    - Proximity plots; and
    - Occurrence Similarity Index (Jaccard’s coefficient).
  - As expected, the developed classification model was not totally perfect. We observed a few false results - such as real_bad cases that were encoded as 5-Excellent and not bad cases that were encoded as 1-Terrible.

CONCLUSION

- Our WordStat/QDA Miner automatic document classification results were pleasingly very reliable, when measured against a published sentiment model.
- The next step would be to scale-up the WordStat/QDA Miner classification model with the full, Amazon Unlocked Phone dataset.
  - This would require developing a case-specific spelling correction dictionary and a word replacement dictionary.
- Another avenue worth pursuing would be to build an automatic sentiment classification model for each specific mobile phone type.
BIBLIOGRAPHY

Dr Normand Peladeau’s webinars on QDA Miner and WordStat:
- *Supervised and Unsupervised Machine Learning Features*
- *Webinar on the New Features of WordStat 8 - Content*


WordStat / QDA Miner Users Manuals: [https://provalisresearch.com/](https://provalisresearch.com/)

ACKNOWLEDGEMENTS

We thank **PromptCloud** for the Amazon Reviews Dataset: **PromptCloud**: [https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones](https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones)

We thank **Dr Normand Peladeau** for QDA Miner and WordStat and his webinars:
- *Supervised and Unsupervised Machine Learning Features*
- *Webinar on the New Features of WordStat 8 - Content*

We thank **Survey Design and Analysis Services Pty Ltd**, vendor of QDA Miner, WordStat and Stata.

We thank the **R Studio** owners.