An International Journal of Advanced Computer Technology

IMPACT OF TRANSFORMED FEATURES IN AUTOMATED SURVEY CODING

Steven Edward[1]

[1]University of Dodoma, P. O. Box 490 Dodoma, Tanzania

Abstract: Survey coding is a process of transforming respondents' responses or description into a code in the process of data analysis. This is an expensive task and this is the reason for social scientists or other professionals in charge of designing and administering surveys tend to avoid the inclusion of many open-ended questions in their surveys. They tend to rely more on the less expensive multiple-choice questions, which by definition do not require a coding phase. However multiple-choice questions strictly limit the respondents' possible answers. This study aims at automating the survey coding process using transformed features. Five intelligent coders were developed using k Nearest Neighbor algorithm, Support Vector Machine with linear function, Support Vector Machine with RBF function and Support Vector Machine with polynomial function. Different response features were applied to improve the coding performance. Techniques that were applied to origin response features include: Relative Frequency, Power Transformation, Relative Frequency Power Transformation and Term Frequency Weighted by Inverse Document Frequency. Furthermore the study proposed new features including: Normalized Relative Frequency, Normalized Relative Frequency with Power Transformation and Normalized Relative Frequency with Term Frequency Weighted by Inverse Document Frequency. The micro-averaged F-measure was used to evaluate the performance of each automated coder. Among all machine learning techniques used Support Vector Machine polynomial was the best when implemented with transformed features.

Keywords: Open-ended Survey Coding; Features; Classifiers;

I. INTRODUCTION

Open-ended questions are important way of obtaining informative data in surveys. This is so for a variety of discipline including market research, customer relationship management, enterprise relationship management, and opinion research in the social and political sciences [1], [2].

Closed-ended questions generate data that are certainly more manageable, but suffer from several shortcomings since they straitjacket the respondent into conveying her thoughts and opinions into categories that the questionnaire designer has developed a priori [3]. As a result, a lot of information that the respondent might potentially provide can be lost.

Sebastian [3] justify that, asking an open-ended question tells the respondent that her opinions are seriously taken into account and her needs cared about. The same cannot be said of closed-ended questions, since these may instead convey the impression that interviewers are interested only in orthodox responses and orthodox respondents.

A. The Drawback of manual coding

According to Andersson & Lyberg [4], manual coding has a number of problems. First it is a source of error in survey. As with most other survey operations, coding is susceptible to errors. The errors occur because the coding function is not always properly applied by the coder and because either the coding function itself or the code is improper.

The function is the one that ensures each element is coded with respect to a specific variable by means of verbal descriptions. In fact, in some statistical studies coding is the most error prone operation next to data collection. For some variables error frequencies at the I0% level are not unusual [4].

Another problem is that coding is difficult to control. Accurate coding requires a lot of judgment on the part of the coder, and it can be extremely hard to decide upon the correct code number. Even experienced coders display a great deal of variation in their coding. Thus there are problems in finding efficient designs for controlling the coding operation.

A third problem is that many coding operations are difficult to administer. Coding has a tendency to become time-consuming and costly. In many countries, coders in large-scale operations must be hired on a temporary basis and the consequences for maintaining good quality are obvious.

B. Statement of the problem

To avoid manual coding challenges (human involvement) some researchers like Giorgetti & Sebastiani [5] had tried to automate the coding task using untransformed feature vectors of questionnaires responses. One limitations of this technique is the dependency on text length leading into lower coding performance. This is because text length may differ within the same class hence lower separability because it depends on a length of text [6]. Furthermore untransformed features can have ill-formed sample distribution leading to more errors.

This study proposed the use of machine learning techniques with transformed features to automate survey coding task so as to increase the text separability, coding performance and speed.

C. Related Works

This study involved a number of features including Absolute Word Frequency (AF), Relative Absolute Frequency(RF), Normalized Relative Frequency (NRF), Absolute Frequency Power Transformation (AFPT), Relative Frequency with Power Transformation (RFPT), Normalized Relative Frequency with Power Transformation (NRFPT), Term Frequency Weighted by Inverse Document Frequency (TFIDF), Relative Frequency with Term Frequency Weighted by Inverse Document Frequency (RFTFIDF) and Normalized Relative Frequency with Term Frequency Weighted by Inverse Document Frequency (NRFTFIDF).

In the literature there are some works that used the absolute frequency features in ASC [7]. However, it is notably that this study and that of Giorgetti [7] are not identical. First, they used untransformed vector while in this study different transformation techniques are used to improve the coding effectiveness.

Another study conducted by Roessingh & Bethlehem [8], in the family expenditure using an trigram coding method and the study by Giorgetti & Sebastiani [5] using Naive Bayes and Multiclass Support vector Machine (MCSVM) differ with this study in terms of methods used. The use of transformation techniques and machine learning techniques in ASC makes this study unique.

II. METHODOLOGY

A. Data for Experiment and Features Extraction

Open-ended questionnaires data were collected at the University of Dodoma in three colleges (CIVE, CHSS and CoED) then coded by professionals. Prepossessing was done and feature vectors were generated. Features vectors generated were used in training and automatic coding.

A total of 1560 questionnaire responses were obtained, they were grouped into twelve categories that is 130 response per category were used. In order to retain independence of data, a cross-validation technique was used by grouping responses of each category into three groups.
One of the groups became the evaluation data and the remaining two groups were used as training data. That is to say by alternating the groups the same data could be used 3 times (two times for training, one time for evaluation).

The average values of the experimental results were used to evaluate classification techniques. The features extracted can be represented in form of feature vector X which can be denoted as:

$$X = (x_1, x_{2,} \dots, x_n)^T, \qquad (1)$$

Whereby n is dimensionality (size of lexicon), x_i is the frequency value of i^{th} word and T refers to the transpose of a vector. In this study the size of dictionary was 486 and the original value of n was 1788.

B. The Procedural Model

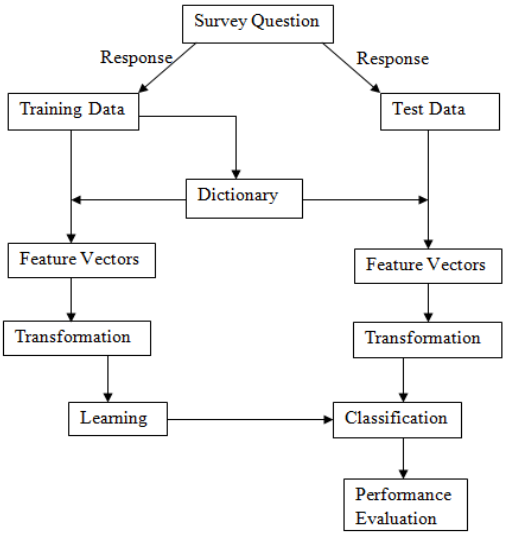

Figure 1: The Procedural Model of this study

The study was guided by the procedural model illustrated in Figure 1. One group of responses from open-ended questions was taken as a training data, the group was used to generate a dictionary (lexicon).

Both training data and test data were converted to feature vectors using the Lexicon and transformed. Machine learns from transformed training data. The learning system can be used to develop the model which can help in coding unseen responses. The performance evaluation was done using F-measure.

C. Convectional Features

2.1. Relative Frequency (RF)

One of the most popular feature extraction techniques is absolute Word frequency. The problem of this technique is the dependency in text length hence lowers separability of feature space [6]. Let y_i be relative frequency which is calculated by

$$y_i = \frac{x_i}{\sum_{j-1}^{n} x_j}, \qquad (2)$$

Whereby x_i is the AF of word i and n is the number of different words.

2.2. Absolute Frequency Power Transformation (AFPT)

In power transformation the coding rate is improved by expressing the absolute frequency of the feature vector as shown in equation

$$z_i = x_i^v \ \langle 0 < v < 1 \rangle. \qquad (3)$$

$$z_i = x_i^v \ \langle 0 < v < 1 \rangle. \qquad (3)$$

This variable transformation improves the symmetric of the distribution of the frequency $x_i \geq 0$ [6]. In this study the value of v was set to 0.5 since the use of this value, normalized the length of RFPT to 1 as shown in equation (4):

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_j} \right) = 1. \qquad (4)$$

2.3. Relative Frequency with Power Transformation (RFPT)

In this technique the output of the relative frequency was taken as the input of the power transformation function as shown in equation (5):

$$m_i = \left(\frac{x_i}{\sum_{j=1}^{n} x_j} \right)^v \quad 0 < v < 1. \qquad (5)$$

This technique of transformation basically aimed at achieving a desirable Gaussian distribution since the distribution leads to an optimal decision boundary [9].

2.4. Term Frequency Weighted by Inverse Document Frequency (TFIDF)

In this technique the feature vector of AF was transformed using the expression shown in equation (6). This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus [10]. Let w_i be the TFIDF:

$$w_i = \frac{\log_2(x_i + 1) + \log_2(N/df)}{\sqrt{\sum_{i=1}^{n} x_i}}, \qquad (6)$$

where df is the number of frequency in which term i occur and *N* is the total number of questionnaires' responses. The study used this feature because the importance of word in questionnaire increases as the number of times a word appears in the questionnaire.

D. Proposed Features

2.5. Normalized Relative Frequency (NRF)

This is a feature technique which finds the relative frequency for normalized feature vector to the unit length. Let h_i be NRF, the equation (7) shows how NRF can be calculated.

$$h_i = \frac{x_i}{\sqrt{\sum_{j=1}^{n} x_j}}, \qquad (7)$$

Whereby x_i is the AF of word i and n is the number of different words. Although this exist in statics literature it has not been applied in ASC.

2.6. Normalized Relative Frequency with Power Transformation (NRFPT)

This technique used the result of the NRF obtained in equation (3.7) in a power transformation function found in equation (3.3). Let q_i be NRFPT q_i can be expressed in the equation (8).

$$q_i = \left(\frac{x_i}{\sqrt{\sum_{j=1}^{n} x_j}} \right)^v \quad 0 < v < 1. \qquad (8)$$

2.7. Relative Frequency with Term Frequency Weighted by Inverse Document Frequency (RFTFIDF)

The result obtained from equation (2) was passed into the TFIDF feature technique as shown in equation (9); Let L_i be RFTFDIF.

$$L_i = \frac{\log_2(y_j + 1) + \log_2(N/df)}{\sqrt{\sum_{j=1}^{n} y_j}}. \qquad (9)$$

2.8. Normalized Relative Frequency with Term Frequency Weighted by Inverse Document Frequency (NRFTFIDF)

The result of the NRF from equation (7) was also passed to TFIDF technique. Let K_i be NRFTFIDF

Table 1: Micro-averaged Performance of Classifiers in different Features in %.

Coder	AWF			RF			NRF		
	AWF	PT	TFIDF	RF	RFPT	RFTFIDF	NRF	NRFPT	NRFTFIDF
KNN	93.60	94.04	93.82	93.48	94.04	93.60	93.48	94.04	93.48
SVM Linear	93.26	93.48	94.16	94.16	93.82	94.16	93.82	93.82	94.16
SVM RBF	93.26	93.37	94.16	94.16	93.82	94.16	94.04	93.71	94.16
SVM Polynomial	93.15	93.26	94.49	94.16	93.60	94.16	93.71	93.82	94.27

$$K_i = \frac{\log_2(h_j + 1) + \log_2(N/df)}{\sqrt{\sum_{j=1}^{n} h_j}}. \qquad (10)$$

E. Feature Reduction

Dimensionalities of training and validation feature vectors were reduced using Principal Component Analysis (PCA). PCA was used to reduce the dimensionality of the feature vectors from a long length to a length more manageable by the machine learning techniques. It does this by projecting orthogonal the features across all feature vectors. It maximizes the total variance, and then removing those that contribute least to the variation.

III. Result and Discussion

A. Performance comparisons of Coders

During the comparison of the coders' algorithms, it was found, that the algorithm with SVM Polynomial performs better than kNN, SVM RBF and SVM linear. To illustrate this, the performance of each algorithm in each transformation technique was measured and compared.

The comparison of the performance of each coders' algorithm was done using the Micro measure F-measure. The results are presented on table 1.

Table 1 shows that the maximum micro-averaged was 94.49 from the algorithm developed by SVM polynomial. The TFIDF and NRFTFIDF features techniques have higher performance when used with any coder.

The kNN algorithm is not the best algorithm for automatic survey coding compared to other three algorithms, because it has bad performance even when the transformation techniques was applied as it was clearly depicted on figure 4.10.

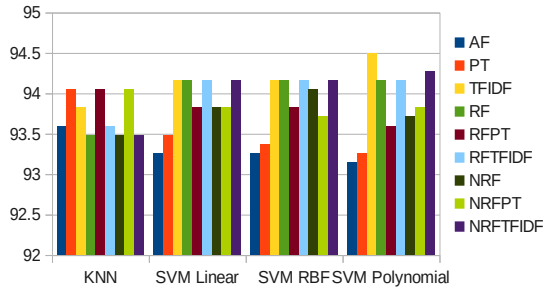

Figure 2: Comparison of coders' Micro-averaged in %

B. Future Work

Future experiments for Automated Survey Coding includes the following areas;

1. *Extensive experimental evaluation using more responses remains as future study. This is due to the fact that the sample size was relatively small.*
2. *Combination of coders algorithm by means of adaptive classifier selection rule (ACS) may be adopted in future work because Sebastiani, (2002) point out that the technique improve the classification performance.*
3. *The use of another dimensionality reduction technique likes canonical discriminant analysis (CDA), or both PCA and CDA may be used in future work. This is because PCA ignores code specific information.*

IV. References

[1] S. Presser and H. Schuman, "The open and closed question," *American Sociological Review*, no. 44, 5, pp. 692–712, 1979.

[2] U. Reja, K. L. Manfreda, V. Hlebec, and V. V., "Open-ended vs close-ended questions in web questionnaires," *Developments in Applied Statistics*, pp. 159–177, 2003.

[3] F. Sebastian, A. Esuli, and T. Fangi, "Machines that learn how to Code Open Ended Survey Data," *University of Consiglio Nazionale delle Ricerche, Pisa, Italy*, 2009.

[4] R. Andersson and L. Lyberg, "Automated coding at statistics Sweden," *Proceedings of the Survey Research Methods Section*, no. American Statistical Association, pp. 41–50, 1983.

[5] D. Giorgetti and F. Sebastiani, "Automating Survey Coding by Multiclass Text Categorization Techniques," *University of Consiglio Nazionale delle Ricerche Pisa, Italy*, 2003.

[6] L. S. P. Busagala, W. Ohyama, T. Wakabayashi, and F. Kimura, "Machine Learning with Transformed features in Automatic Text Classification," in *Proceedings of ECML/PKDD-05 workshop on Sub-symbolic Paradigms for Learning in Structured Domains (Relational Machine Learning)*, 2005, pp. 11–20.

[7] D. Giorgetti, I. Prodanof, and F. Sebastian, "Automated Coding of Open-ended Surveys: Technical and Ethical Issues," *Instituto di Linguistice Computazionale, CNR, Pisa Italy*, 2004.

[8] M. Roessingh and J. Bethlehem, "Trigram coding in the family expenditure survey in statistics," *Netherlands Central Bureau of Statistics*, 1983.

[9] A. Malero and L. S. . Busagala, "The Impact of Transformed features in Automatic Spam Filtering," *Journal of Informatics and Virtual Education, Tanzania*, vol. 1, pp. 55–58, 2011.

[10] D. . Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," *Cambridge University Press*, 2008.

[11] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.