Using BPM Frameworks for Identifying Customer Feedback About Process Performance

Sanam Ahmad, Syed Irtaza Muzaffar, Khurram Shahzad(✉), and Kamran Malik

Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan
{sanam.ahmad, irtaza, khurram, kamran.malik}@pucit.edu.pk

Abstract. Every organization has business processes, however, there are numerous organizations in which execution logs of processes are not available. Consequently, these organizations do not have the opportunity to exploit the potential of execution logs for analyzing the performance of their processes. As a first step towards facilitating these organizations, in this paper, we argue that customer feedback is a valuable source of information that can provide important insights about process performance. However, a key challenge to this approach is that the feedback includes a significant amount of comments that are not related to process performance. Therefore, utilizing the complete feedback without omitting the irrelevant comments may generate misleading results. To that end, firstly, we have generated a customer feedback corpus of 3356 comments. Secondly, we have used two well-established BPM frameworks, Devil's Quadrangle and Business Process Redesign Implementation framework, to manually classify the comments as relevant and irrelevant to process performance. Finally, we have used five supervised learning techniques to evaluate the effectiveness of the two frameworks for their ability to automatically identify performance relevant comments. The results show that Devil's Quadrangle is more suitable framework than Business Process Redesign Implementation framework.

Keywords: Business data analytics · Customer reviews
Process performance analysis · Text analytics · Supervised learning techniques

1 Introduction

Business processes are everywhere [1] and they are widely pronounced as the basic unit of work for every organization [2, 3]. Recognizing the pivotal role of processes, growing number of organizations are automating their processes [4] and utilizing their execution logs for the performance analysis [5]. However, presently, there are numerous organizations that are yet to automate their processes. Consequently, these organizations cannot exploit the potential of execution logs for analysing processes' performance.

R. Matulevičius and R. Dijkman (Eds.): CAiSE 2018 Workshops, LNBIP 316, pp. 55–69, 2018.
https://doi.org/10.1007/978-3-319-92898-2_5

To facilitate these organizations, a possible alternate is to collect the customer feedback about the business process under consideration, and use the collected feedback to gain insights about the process performance. Such an approach is particularly useful for service-oriented companies, such as insurance companies and restaurants, where customer satisfaction is of higher significance [6]. In addition to service-oriented companies, the effective utilization of customer feedback has the potential to offer manifold benefits to every organization [7]. These benefits include, but not limited to, introducing new products or services, evaluating customer satisfaction, identifying customer preferences, sustaining existing features and introducing new features [8, 9]. However, customer feedback includes the comments that are not related to process performance. Hence, any insights obtained by processing the entire collection of comments, that is, without segregating irrelevant comments, may be misleading. This arises the question how to distinguish between performance relevant and irrelevant comments? The answer to this question essentially requires a clear understanding of the notion of performance in the context of business processes. To this end, in this paper we have used two well-established BPM frameworks to evaluate the effectiveness of the two frameworks for their ability to distinguish between relevant and irrelevant comments. Specifically, we have made the following three main contributions:

- *Feedback Corpus:* We have generated a corpus of over 3356 comments by collecting feedback from two sources, social media and survey.
- *Benchmark Annotations:* We have generated two datasets by manually annotating each comment as relevant or irrelevant, using two different criteria. The criteria stem from the constituents of two well-established conceptual frameworks: Devil's Quadrangle framework [10] and Business Process Redesign [11] framework.
- *Suitability Evaluation:* We have thoroughly evaluated the effectiveness of the two frameworks, using the generated datasets as their proxies, for their abilities to distinguish between relevant and irrelevant comments. For the evaluation, we have performed experiments using five established supervised learning techniques to automatically classify the comments in both datasets.

1.1 Problem Illustration

To illustrate the problem that all the comments in the customer feedback are not related to process performance which may mislead process analysts; consider an excerpt version of admission process of an institute. The process starts when an applicant collects an application form. Each form has a unique ID that is used to track an application throughout the admission cycle. The application form comprises of several subsections including biography, academic background, experience and an entry test slip. Each candidate is required to fill the form and deposit entry test fee. There are two modes of fee payment, online payment and payment through bank. If a candidate desires to pay through bank, he/she must use a part of the admission form as an invoice. Once the payment is deposited, the completed form along with all the documents is submitted to the institute.

Presently, neither any part of the admission process is automated nor the specification of the process is documented in the form of a process model. However, for a

better comprehension of the example, we have presented an excerpt of the admission process model in Fig. 1.

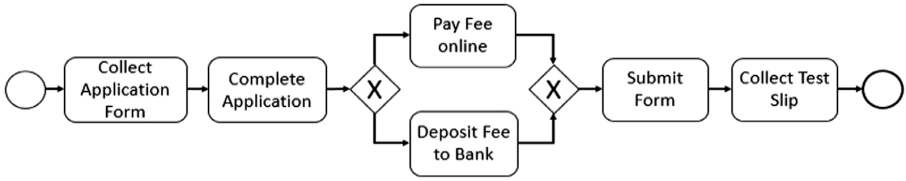

Fig. 1. An excerpt version of the admission process model.

Table 1 contains six example comments about the admission process to illustrate the classification problem. From the table it can be observed that some of the comments are about process performance whereas others are irrelevant to process performance. In the rest of the paper, the former is referred to as relevant comments and the latter are referred to as irrelevant comments. For instance, the comment 'the application form was so lengthy that it took me one hour to complete it' is clearly about the time consumed in completing the application form. Therefore, it is declared as a relevant comment. Similarly, the comments about the ease of use, delays and longer queues are also related to process performance. In contrast, the two comments, 'the campus is very far from my place' and 'my friend helped me to prepare for the test' are not relevant to process performance. Therefore, these comments are declared as irrelevant comments.

Table 1. Example of relevant and irrelevant comments.

Example comments	Relevance
1. The application form was so lengthy that it took me one hour to complete it	Yes
2. The application portal was easy to use	Yes
3. I had to wait for one hour to get access to computer when I went to campus for applying	Yes
4. There were longer queues at the bank	Yes
5. I am happy that my friend helped me in test preparation	No
6. The campus is very far from my place	No

Consider that the institute's administration is interested in knowing, how often applicants talk negatively about admission process? Generating the answer to this question requires classifying comments as positive or negative and then counting the number of negative comments. However, if the complete set of comments are used, without excluding irrelevant comments, misleading results may be generated. For instance, the comment 'I am happy that my friend helped me in test preparation' is a positive comment. However, from the process analysis perspective it is an irrelevant comment that should not be counted in generating the answer to the posed question. Similarly, 'the campus is very far from my place' is a negative comment but the

institute may not like to consider this comment due to its irrelevance with process performance. Therefore, this comment should not be used in generating the answer of the posed question. However, if the two comments are used in answering the question, it may mislead the administration.

Based on the illustration it can be concluded that it is necessary to first identify the comments that are related to process performance, before they can be used to gain insights about process performance. Else, if the complete set of feedback is used, the irrelevant comments may skew the results and mislead analysts. To address this problem, in this paper, we aim to use two well established BPM frameworks for identification of performance relevant comments.

2 The BPM Frameworks

Development of a comprehensive and adequately crisp criteria for the classification of customer feedback is a challenging task, due to the involved intricacies. Our initial attempt to invent classification criteria from scratch, resulted in a long list of heuristics and their prolonged descriptions, which hindered the development of a common understanding of the criteria. Therefore, we rely on two well-established and widely used conceptual frameworks for the development of relevance criteria. The frameworks are, Devil's Quadrangle framework and Business Process Redesign Implementation framework. The key reason for choosing these frameworks is their strong association with business processes. That is, DQ framework describes the performance dimensions that must be taken into consideration for analyzing process performance, whereas, BPRI framework describes the elements that must be considered in improving the design of a process. A brief overview of each frameworks is as follows:

Devil's Quadrangle (DQ) Framework. The DQ framework is composed of four dimensions that were introduced to evaluate the impact of each best practice on business process [10]. The framework is widely pronounced as an *ideal* framework for the performance analysis of a process [10, 11]. The four performance dimensions are, time, cost, quality and flexibility. In the framework, time dimension refers to the amount of time consumed or delayed in executing a process P. Cost refers to the effort, resources or revenue consumed during the execution of P. Quality refers to the satisfaction with the specification and execution of P, and flexibility refers to the ability of process to respond to a change.

Business Process Redesign Implementation (BPRI) Framework. The framework was developed with the intent to help process designers in delivering a design that is superior than the existing design, by identifying the elements that should be considered and relationships between these elements [11, 12]. Furthermore, the framework has also been used to think and reason about the most important manifestations of redesign [13]. It consists of seven elements, customers, products, business process (operation and behavior view), participants, information, technology, and environment. Customer, the first element of the framework, refers to the internal or external customers of the process that benefit from the process. Product refers to the items or services generated or consumed by the process. Business process refers to the set of activities as well as

dependencies between activities. The element, participants in the framework, refers to the individuals or roles that execute the activities. Information refers to the data produced or generated by the process. Technology refers to the methods or techniques used in the process, and environment refers to the external conditions or surroundings in which the process executes.

3 Customer Feedback Corpus

In this section, we outline the corpus generation procedure and the classification criteria corresponding to each framework. Subsequently, the procedure for generating the two datasets is presented.

3.1 Corpus Generation

For the study, we collected student feedback about the admission process of an academic institute. Every year, the institute receives several thousand applications for admission to its various programs. The admission process starts with announcement of the admissions schedule and ends with the announcement of admissions decisions. Due to the space limitations, we only present key activities of the admission process. These are, announce admissions, collect application form, complete application form, choose preferred program, choose campus, submit application form, collect fee voucher, pay fee through bank, verify academic record, generate entry test slip, appear in the admission test, rank students, and announce admission decisions.

For this study, we collected student feedback from two sources, social media and a survey. To collect student feedback from social media, we scrapped the Facebook page of the institute to extract over 1000 student posts and comments on these posts. To further extend the corpus, we conducted an unstructured survey with applicants. The survey was composed of a brief introduction to the study, few open-ended questions and a few example answers. We opted to use open-ended questions due to two reasons, (a) to give respondents the complete freedom to share their feelings or experiences, and (b) to avoid emphasizing any fragment of the process for feedback. The participants were given three weeks to fill the survey with the freedom to save and updated their comments.

At first, we compiled a corpus of 3510 comments from the two sources. However, after omitting the incomplete comments, non-English, and trivial comments, the corpus size was reduced to 3356. Subsequently, the corpus was pre-processed by correcting the spellings and replacing the abbreviations with complete words. For spelling correction, we used a two-step semi-automated approach. In the first step, a python script tokenized each comment and searched each token in WordNet (an online English dictionary), to identify the tokens that were not available in the dictionary. In the second step, a researcher reviewed each unverified token and corrected it. The corpus generation procedure is presented below in Fig. 2.

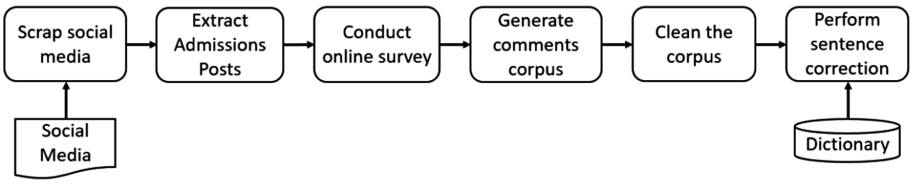

Fig. 2. Corpus generation procedure.

3.2 Generating Relevance Criteria

Once the feedback corpus was generated, the next step was to generate the benchmark datasets by manually classifying the comments as relevant or irrelevant. However, as discussed earlier, declaring a comment relevant or irrelevant is a challenging task due to the involved intricacies. Hence, it is essential to develop a common understanding about which type of comment should be declared as relevant and which type of comment should be declared as irrelevant. To that end, we rely on two well-established and widely used frameworks, as discussed in the preceding section. While both the frameworks are widely pronounced as precious artifacts in their respective context [10–14], the mere description of performance dimensions (in DQ framework) or key constituents of process redesign (in BPRI framework) are not adequate for the classification of feedback. An accurate classification rather requires a scale, rule, or principle for evaluating whether a given comment is relevant or irrelevant. For this study we generated two separate criteria based on the two frameworks. An excerpt version of each criteria is presented below in Tables 2 and 3.

Table 2. An excerpt of the DQ criteria.

Relevance criteria	Remarks
Is the comment related to robustness of the process?	A comment related to delay, queue or waiting time is relevant
Is the comment related to cost incurred by the candidate for the process?	A comment about the characteristic of the product is irrelevant
Is the comment related to quality of the process?	Any feeling or suggestion about the execution of the process is relevant
Is the comment related to flexibility of the process?	Any suggestion about the changes in process is relevant

For generating the first criteria, hereafter DQ criteria, we defined at least one question corresponding to each performance dimensions of the DQ framework. Similarly, for generating the second criteria, hereafter BPRI criteria, we included at least one question corresponding to each element of the BPRI framework. While the development of DQ criteria was a straightforward task the development of BPRI criteria was found to be challenging. This was due to the peculiar nature of some elements of the framework. For each such element, we defined a candidate question and iteratively tweaked the question by improving its formulation and adding remarks

Table 3. An excerpt of the BPRI criteria.

Relevance criteria	Irrelevance/Remarks
Is the comment or suggestion related to the use of an object in an activity?	A comment on the object used in the activity. For instance, the prospective quality was good
Is the comment or feeling related to the technology used for performing an activity?	A comment on the technology used in an activity is irrelevant
Is the comment related to the manual use of the data generated/produced by an activity?	A comment related to the data/information used/generated/produced during the activity
Is the comment related to the response or behaviors of the participant?	A comment on the participant that has no effect on the perform an activity
Is the related to the quality of service delivered by third party or environment?	A comment related to the third part or the environment

against the question. Each iteration was performed by two researchers and each tweak was performed based on the assessment of the question on 10 relevant and 10 irrelevant comments. Accordingly, we finalized a question corresponding to each element of the framework. For instance, *information* is an element of the BPRI framework that cannot be easily used for developing relevance criteria for process performance. As an outcome of the two iterations, we developed the following question regarding this element. Is the comment related to the manual use of the data generated or produced by an activity? In addition to the question, an example comment was also added i.e. 'So, disappointing that the spellings of my name in the list were incorrect'.

3.3 Generating Datasets for Each Framework

Once the two criteria are defined, the next step in generating the benchmark datasets for each framework is to use the criteria for manually classifying the 3356 comments in the feedback corpus. For that, at first, we randomly collected 1000 comments from the corpus and asked two researchers, R_1 and R_2, to independently classify all the collected comments using the DQ criteria. Both the researchers are PhD candidates and they had taken at least two courses on natural language processing and business process management. Furthermore, both the researchers are familiar with the concept of annotation procedures, annotation guidelines and inter-annotator agreement. The results generated by the two researchers were compared and their inter-annotator agreement was computed. From the comparison it was observed that out of the 1000 comments, there were 895 agreements, 105 disagreements and a Kappa statistic of 0.762. The detailed specifications are presented in the Table 4.

After three weeks, two researchers were asked to classify the same collection of 1000 comments using the BPRI criteria. The time gap between the two cycles was maintained to ensure that the researchers cannot reuse the knowledge of applying the DQ criteria [15]. Similar to the DQ criteria based classification, the two researchers classified comments using the BPRI criteria, and their inter-annotator agreement was computed. From the comparison of two annotated datasets it was observed that the use of the BPRI criteria resulted in 778 agreements, 222 disagreements, and a Kappa statistic of 0.49.

Table 4. Specification of the annotations.

	DQ Framework	BPR Framework
Total number of comments	3356	3356
Size of random sample size	1000	1000
Identical markings	895	778
Different markings	105	222
Kappa statistics	0.762	0.49
Remaining annotations	1	2

We have the following observations from the application of BPRI and DQ criteria:

- Both the researchers expressed that the task of classifying comments using BPRI criteria was harder than that of DQ criteria. It is because, the BPRI criteria involves several intricacies which increases the cognitive effort required to apply the BPRI criteria. For instance, the researchers found it hard to distinguish between the comments *about* the technology from the comments related to the *use* of the technology.
- The impact of the preceding observation can also be noted in the results. That is, the number of agreements for DQ criteria are greater than BPRI criteria (895 > 778). Similarly, the inter annotator agreement for DQ criteria is higher than BPRI criteria (0.762 > 0.49).

Due to the higher inter-annotator agreement in the first cycle of using DQ criteria, the remaining 2356 comments were classified by a single researcher. However, prior to that, each disagreement was discussed by the two annotators and a common understanding of the criteria was developed. The dataset generated by this approach is referred to as DQ dataset in the rest of the paper. In contrast to the DQ criteria, the inter annotator agreement was low when BPRI criteria was used. Therefore, the remaining 2356 comments were classified by both researchers, and subsequently conflicts were resolved by discussion and mutual consent. The dataset generated by this approach, is referred to as BPRI dataset. Table 5 shows a confusion matrix of the two frameworks.

Table 5. Confusion matrix of two frameworks.

		DQ framework		
		Relevant	Irrelevant	Total
BPRI Framework	Relevant	998	35	1033
	Irrelevant	1618	705	2323
	Total	2616	740	3356

To gain further insights about the classified comments, we compared the classified comments to reveal the following:

- The 2616 comments classified as relevant by using the DQ criteria includes a significant share of the comments (998 out of 1033) that are declared relevant by the BPRI criteria.

– The 2323 comments classified as irrelevant by using the BPRI criteria includes a significant share of the comments (705 out of 740) that are declared irrelevant by the DQ criteria.

The two results represent that the relevant comments in the DQ dataset subsumes the relevant comments in the BPRI dataset. Similarly, the irrelevant comments in the BPRI dataset subsumes a significant percentage of the irrelevant comments in the DQ dataset. More specifically, the first observation represents that the use of DQ criteria enables identification of a large share of comments that are also declared as relevant by BPRI framework. Furthermore, the DQ criteria enables identification of 1618 additional relevant comments that were declared as irrelevant by the BPRI framework. This number is so large that it can skew the answer of virtually every question and may also generate entirely different perception about the performance of the process.

4 Automatic Classification of Customer Feedback

We have performed experiments using two datasets and five supervised learning techniques to evaluate the effectiveness of the two frameworks for distinguishing between relevant and irrelevant comments. In case all the supervised techniques achieve higher accuracy for one dataset, it conclusively represents that the framework used for generating the dataset is more effective for classifying comments. It is because, all five techniques rely on a set of feature values for learning and predicting the relevance of a comment, and the presence of similar and non-conflicting feature values results in boosting the effectiveness of supervised learning techniques and vice versa. These similar or non-conflicting feature values represent that majority of the comments in the dataset, that are placed in one class, have identical or similar feature values.

The following subsections provide an overview of the five supervised learning techniques and our evaluation setup. Subsequently, we present a detailed analysis of the results.

4.1 Supervised Learning Techniques

We have used five widely used supervised learning techniques for experimentation, Logistic Regression (LR) [16], Support Vector Machine (SVM) [17], Decision Tree (DT) [18], Random Forest (RF) [19] and K Nearest Neighbors (KNN) [20].

Support Vector Machines (SVM). The basic idea behind the training process in SVM is to find the hyperplane which optimally separates data points of different classes. The optimal hyperplane is the one which yields maximum margin. Margin is the distance between hyperplane and closest data point of other classes. In our domain two possible classes of comments to be classified are relevant and irrelevant. Each comment to be classified is denoted by document d. The SVM model is defined as

$$h_w(d) = \{1 \quad if \quad w^T d \geq 1 \qquad 0 \quad if \quad w^T d \leq -1\}$$

Logistic Regression (LR). Logistic regression is one of the most widely used and powerful algorithms for classification problems. In logistic regression, the selected hypothesis function always predicts output values between 0 and 1.

$$0 \leq \mathrm{h}_w\left(w^{\mathrm{T}}d\right) \leq 1$$

The hypothesis function is represented by the sigmoid function as follows:

$$\mathrm{h}_w\left(w^{\mathrm{T}}d\right) = \frac{1}{\left(1+\mathrm{e}^{-w^{\mathrm{T}}d}\right)}$$

Where $\mathrm{h_w}(\mathrm{w^T d})$ is the hypothesis function for logistic regression, parameterized by w, and d is the input variable or feature which is in our case comment.

K Nearest Neighbor (KNN). KNN is an instance based classification technique where comment is classified either relevant or irrelevant by comparing its similarity to the comments in training data that are already labelled.

Each comment is treated as document. Let U is set of unlabeled documents and L is set of labelled documents. A given document d $\in U$, Let NN_K^L (d) is set of top K documents in L that are most similar to the input document d using some similarity function. We label the document d as the label of K most similar documents to the document d.

Decision Tree (DT). Decision tree is non-parametric classification methodology. Decision tree model predicts the value of class for a given data point by learning the decision rules inferred from labelled data set.

Random Forest (RF). Random Forest technique is used to overcome the problem of being over fitted to the training data set in decision trees. Random Forest uses random feature selection for individual decision tree development. Random forest also uses bagging method. The trees in random forest are tested using out-of-bag sample and predictions of these trees is either averaged or voted for final prediction calculation.

4.2 Experimentation

For the experimentation, we have used DQ and BPRI datasets. Recall, the DQ dataset includes 2616 relevant comments and 740 irrelevant comments. In contrast to that, BPRI dataset includes 740 relevant comments and 2323 irrelevant comments. Evaluation is carried out using three widely used measures Precision, Recall and F_1 score [21]. Precision is the fraction of correctly classified comments among the classified comments. Recall is the fraction of correctly classified comments among the comments that should have been classified correctly. F_1 score is the harmonic mean of Precision and Recall.

As discussed above, we have performed experiments using five supervised learning techniques. For the experiments we have used Scikit-learn library in Jupyter notebook. The input to each technique is a set of numeric values called feature values. In our case, for both the datasets, we have performed separate experiments using unigram, bigram

and trigram feature matrices. Generating each feature matrix involves the following steps (i) tokenize the dataset, (ii) preprocessing dataset by omitting stop words and stemming each token using Stanford parser [22], (iii) generating a set (unique) of word tokens of length N, called N grams (unigrams, bigrams or trigrams), and (iv) generating feature matrix. In a matrix, columns represent the set of N grams generated from the third step of the above procedure and rows represent the comments. A cell in the matrix corresponding to Row J (say, R_J) and column K (say, C_K) contains a binary score of 1 or 0. The value 1 in (R_J, C_K) represents that the comment in R_J contains the word token C_K, whereas, the value 0 in (R_J, C_K) represents that the comment in R_J does not contain the word token C_K.

For each experiment, we have used a training and testing ratio of 65:35. The results are calculated by using 10-fold cross validation to rationalize the bias that may be induced due to the choice of training and testing samples. The results presented in the subsequent section are the average scores of the 10-fold cross validation. Additionally, experiments are also performed by using all possible combinations of preprocessing, removing punctuations, removing stop words and stemming, to choose the most appropriate combination.

4.3 Results and Analysis

Table 6 summarizes the results of 10-fold cross validation for both datasets. From the table it can be observed that for DQ dataset, LR technique achieved very high F_1 score ($F_1 = 0.95$) using unigram feature matrix. Also, the precision and recall scores are comparable with the F_1 score (P = 0.93 and R = 0.96). From the table, it can also be observed that RF achieved a very low F_1 score using trigram feature matrix ($F_1 = 0.56$). In this case, the precision score is still higher (P = 0.96), however, the Recall is very low (R = 0.39). These results represent that most of the comments declared relevant by RF techniques are also relevant in the benchmark dataset. However, majority of the relevant comments in the gold standard are declared irrelevant by RF technique.

For the BPRI dataset, overall LR and DT achieved a high F1 score ($F_1 = 0.77$) using unigram feature matrix. Also, both Precision and Recall scores are comparable, i.e. for LR and DT techniques, P = 0.83 and 0.78, respectively; and R = 0.73 and 0.77, respectively. Below, we present some key observations about the results.

Most Appropriate Feature. Figures 3 and 4 shows a comparison of N-gram feature matrices (Unigram, Bigram and Trigram) in supervised learning techniques. From the figures it can be observed that the unigram is the most appropriate feature for both datasets. Furthermore, it can be observed from Fig. 3, all supervised learning techniques are equally effective for Unigram (i.e. N = 1). However, as the value of N increases the difference in performance becomes more visible. From Fig. 4 it can be observed that all the techniques are not equally effectively for BPRI dataset. These observations represent that the feature values in DQ dataset are similar and non-conflicts, hence, more suitable for learning. In contrast, the feature values in BPRI dataset are diverse and conflicting, hence, not suitable for learning.

Table 6. Summary results of the experiments.

		DQ dataset			BPRI dataset		
Feature	Algorithm	P	R	F1	P	R	F1
Unigram	KNN	0.94	0.88	0.91	0.73	0.44	0.54
	SVM	0.95	0.94	0.94	0.78	0.74	0.76
	LR	0.94	0.96	0.95	0.83	0.73	0.77
	RF	0.93	0.96	0.94	0.84	0.64	0.71
	DT	0.94	0.93	0.93	0.78	0.77	0.77
Bigram	KNN	0.79	0.97	0.87	0.75	0.15	0.24
	SVM	0.95	0.74	0.83	0.8	0.62	0.69
	LR	0.93	0.9	0.92	0.87	0.53	0.66
	RF	0.94	0.77	0.84	0.87	0.51	0.65
	DT	0.94	0.76	0.84	0.74	0.65	0.69
Trigram	KNN	0.78	0.99	0.87	0.72	0.08	0.14
	SVM	0.96	0.57	0.71	0.86	0.29	0.43
	LR	0.78	0.99	0.87	0.91	0.21	0.34
	RF	0.96	0.39	0.56	0.88	0.22	0.37
	DT	0.96	0.43	0.59	0.78	0.37	0.49

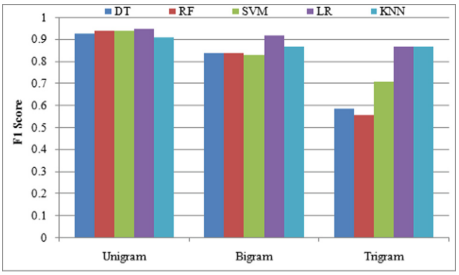

Fig. 3. Feature selection for DQ dataset

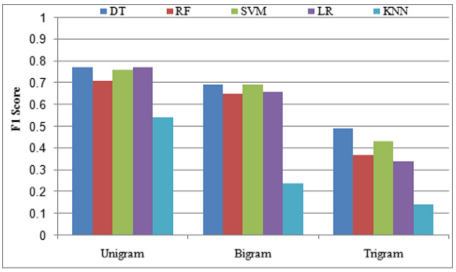

Fig. 4. Feature selection for BPRI dataset

Performance Variation Across Datasets. Figures 5, 6 and 7 shows a comparison of five techniques for the two datasets. From the figures it can be observed that using unigram feature matrix (i.e. the most discriminating feature), the performance scores of all the techniques for the DQ dataset are higher than BPRI datasets. Similar trends can be observed for the bigram and trigram features. These higher performance scores of all techniques and across all features represent that, the DQ dataset contains similar and non-conflicting feature values. These results represent that the comments having identical or similar feature values belong to the same class. In contrast, the BPRI dataset contains diverse and conflicting feature values, representing that the dataset includes several comments having similar feature values but they are placed in different classes. These results, together with the expression of the researchers (that the use of BPRI criteria for classifying comments is harder than DQ criteria), are abundantly conclusive to declare that DQ framework is more effective than BPRI framework.

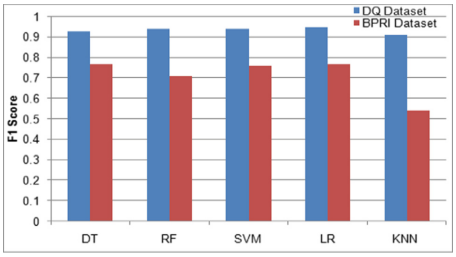

Fig. 5. Comparison of both frameworks using unigram feature

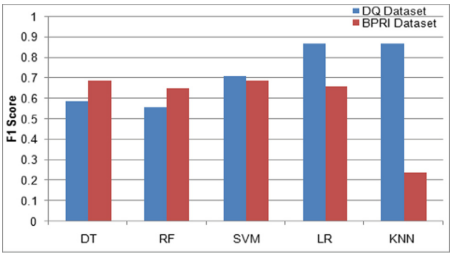

Fig. 6. Comparison of both frameworks using bigram feature

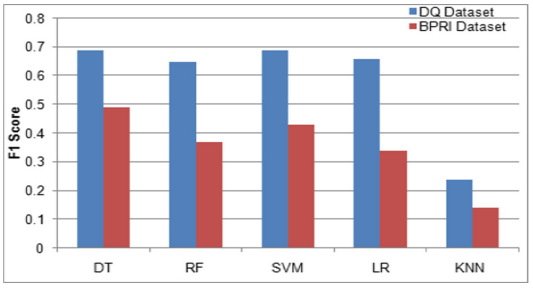

Fig. 7. Comparison of frameworks using trigram feature

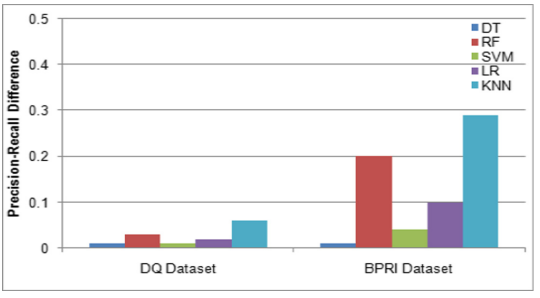

Fig. 8. Difference between Precision and Recall.

Variation Between Precision and Recall Across Dataset. Figure 8 shows the absolute difference between Precision and Recall scores across the two datasets using unigram feature matrix. The plotted values are computed by taking modulus of the difference between Precision and Recall scores. From the figure it can be observed that for DQ dataset the difference between Precision and Recall is very small compared to that of BPRI dataset. These results further affirm the suitable of the DQ framework.

5 Discussion and Conclusions

In this paper we propose an alternate to the traditional process performance analysis approaches that essentially requires the execution log or event log of the business process, whose performance analysis is desired. Our proposed alternate involves collecting and utilizing unstructured customer feedback and using it for the performance analysis of a business process. However, a key challenge to such an approach is that, the feedback includes several comments that are not related to process performance. Therefore, utilizing the complete feedback may generate misleading results. This arises the question, how to identify the comments that are related to process performance? To answer this question, in this paper we have used two well-established BPM frameworks to evaluate their suitable for identifying process performance related comments. The frameworks are, Devils Quadrangle and Business Process Redesign Implementation. For that, we have first generated a feedback corpus that includes 3356 comments.

Secondly, we have generated two criteria, based on the two frameworks, and used them for manually classifying relevant and irrelevant comments. During the classification it was observed the use of BPRI framework based criteria (BPRI criteria) is harder than that of DQ framework based criteria (DQ Criteria). The impact of that can also be observed in the results, that is, the number of agreements in applying the DQ criteria are significantly more than BPRI criteria. An analysis of the two datasets revealed that a large majority of the comments declared relevant by BPRI criteria are also declared relevant by the DQ criteria. Furthermore, the use of DQ criteria leads to identification of additional relevant comments, in addition to the relevant compared identified by the BPRI criteria.

Thirdly, we have compared the effectiveness of the two frameworks by using the two datasets generated in the preceding step. The results reveal that, (a) all five techniques generate achieve higher accuracy for the DQ dataset as compared the BPRI dataset, (b) unigram is the most discriminating feature for classification, (c) the absolute difference between precision and recall for DQ dataset is negligible for all the techniques, whereas the same difference is significant for the BPRI datasets.

The summarized results represent that DQ framework is more suitable because, it not only identifies a large set of process performance related comments, but also classifies the comments in the same class that has similar feature set. In contrast, the cognitive effort required to use BPRI framework is higher due to intricacies in the criteria and its use in supervised learning techniques also impedes the performance of supervised learning techniques.

Given that all organizations have business processes, in this study we argue that there is a need to engage the organizations that are yet to embrace BPM. For that, we have taken an initial step towards proposing an innovative solution in which such organizations can get a sense of their business process performance without going through the complete BPM lifecycle. The solution involves, application of data analytics on the customer feedback to gain insights about the process performance. In the future we plan to utilize the classified comments for business process redesign.

References

1. van der Aalst, W.M.P., Pesic, M., Song, M.: Beyond process mining: from the past to present and future. In: Pernici, B. (ed.) CAiSE 2010. LNCS, vol. 6051, pp. 38–52. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13094-6_5
2. Zhang, Y., Liang, R., Shi, Z., Ma, H.: The design and implementation of a process-driven higher education administrative system. IERI Procedia **2**(5), 176–182 (2012)
3. Krajewski, L., Ritzman, L., Malhotra, M.: Operations Management: Processes and Value Chains, 8th edn. Prentice Hall, Upper Saddle River (2006)
4. Jorge, M.: Conformance Checking and Diagnosis in Process Mining: Comparing Observed and Modeled Processes. LNBIP. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-49451-7
5. van der Aalst, W.M.P., Ardiansyah, A., Dongen, B.: Replay history on process model for conformance checking and performance analysis. WIREs Data Min. Knowl. Discov. **2**(2), 182–192 (2012)
6. Danaher, P.J., Mattsson, J.: Customer satisfaction during the service delivery process. Eur. J. Mark. **28**(5), 5–16 (1994)
7. Thomke, S., von Hippel, E.: Customers as innovators: a new way to create value. Harvard Bus. Rev. **80**(4), 74–85 (2002)
8. Hauser, J.R.: How Puritan-Bennett used the house of quality. Sloan Manag. Rev. **34**(3), 61–71 (1993)
9. Kujala, S.: User involvement: a review of the benefits and challenges. Behav. Inf. Technol. **1** (22), 1–16 (2003)
10. Reijers, H.A., Mansar, S.L.: Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics. Omega **33**(4), 283–306 (2005)
11. Mansar, S.L., Reijers, H.A.: Best practices in business process redesign: validation of a redesign framework. Comput. Ind. **56**(5), 457–471 (2005)
12. Dumas, M., Rosa, M.L., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33143-5
13. Dumas, F., Aalst, W., Hofstede, A.H.M.: Process Aware Information Systems. Wiley, Hoboken (2005)
14. Jansen-Vullers, M.H., Kleingeld, P.A.M., Netjes, M.: Quantifying the performance of workflows. Inf. Syst. Manag. **25**(4), 332–343 (2008)
15. Cornal, K., Schuff, D., Louis, R.D.S.: The impact of alternative diagrams on the accuracy of recall: a comparison of star schema and entity relationship diagrams. Decis. Support Syst. **42** (1), 450–468 (2006)
16. Kutner, M.H., Nachtsheim, C., Neter, J.: Applied Linear Regression Models. McGraw-Hill/Irwin, New York (2004)
17. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2001)
18. Rokach, L., Maimon, O.: Data Mining with Decision Trees. Theory and Applications, 2nd edn. World Scientific, River Edge (2014)
19. Zhang, C., Mai, Y.: Ensemble Machine Learning Methods and Application. Springer, New York (2012). https://doi.org/10.1007/978-1-4419-9326-7. p. 157
20. Kirk, M.: Thoughtful Machine Learning a Test-Driven Approach. O'REILLY, Sebastopol (2015)
21. Yates, R.B., Neto, B.R.: Modern Information Retrieval. ACM Press, New York (1999)
22. Version S.P.1.6: SPSS Inc., ChicagoIII (2008). https://nlp.stanford.edu/software/lex-parser.shtml