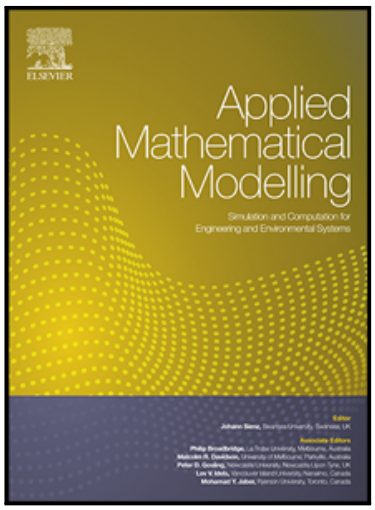

ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models

Fouad Zablith , Ibrahim H. Osman

PII: S0307-904X(19)30117-9
DOI: https://doi.org/10.1016/j.apm.2019.02.032
Reference: APM 12685

To appear in: *Applied Mathematical Modelling*

Received date: 11 July 2018
Revised date: 14 December 2018
Accepted date: 25 February 2019

Please cite this article as: Fouad Zablith , Ibrahim H. Osman , ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models, *Applied Mathematical Modelling* (2019), doi: https://doi.org/10.1016/j.apm.2019.02.032

Highlights

- We propose a novel predictive analytics framework to classify and analyze unstructured reviews
- Evaluation models can potentially train machine learning algorithms for predicting reviews classification and sentiments
- We implement neural networks and logistic regression algorithms tested in the context of e-government service evaluation
- The classification reached a promising F-score of 85.16%, and sentiments correlating 71.44% with a manually validated dataset
- The framework contributes to uncover hidden insights that were not initially captured by closed-ended questionnaires

ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models

Fouad Zablith*, Ibrahim H. Osman

Olayan School of Business,
American University of Beirut
PO Box 11-0236,
Riad El Solh, 1107 2020,
Beirut, Lebanon

* Corresponding author.
E-mail addresses: fouad.zablith@aub.edu.lb (F. Zablith), ibrahim.osman@aub.edu.lb (I. H. Osman).

Abstract

While research interest on product and service evaluation from unstructured text reviews is increasing, investigating the effectiveness of predictive analytical models in this context is still underexplored. With the advancement in machine learning research, an opportunity exists to bridge this gap using a model-based product and service evaluation. We propose in this article *ReviewModus*, a text mining and processing framework that (1) relies on the model structure and its corresponding assessment questions to train a machine learning algorithm to predict the classification of reviews around the model dimensions; (2) predicts the sentiments within the reviews based on external review training datasets; and (3) transforms the extracted measures from the reviews for further analysis. Our approach is evaluated in the context of 11 e-Government services where the performance of the framework is compared to the manual processing of unstructured reviews cross-checked by three independent evaluators. Our study shows promising classification results with a micro-average F-score reaching 85.16%, and a high sentiment prediction correlation (71.44%) with the manually performed sentiment assessment.

Keywords: machine learning; text mining; neural network; logistic regression; e-government

1. Introduction

Traditional evaluation and user satisfaction models have been extensively used to analyze users' feedback provided in structured forms [1–3], however the surge of users' opinions in unstructured forms is still a challenge to such models. Researchers have invested a good amount of effort in developing and testing evaluation models to translate the data collected from users' feedback into meaningful actions. Such models typically take a set of dimensions relevant to the product or service being assessed as input, and feed them into appropriate output dimensions to generate evaluation measures. The input and output variables are usually defined around the product or service characteristics and are often populated by

designing a set of survey questions. Such models play a major role in testing certain hypotheses and ensuring a consistent evaluation process across different products and services. However, they usually require a pre-defined and structured data input which is feasible in controlled data collection settings, achieved through closed-ended questions [1], but is much harder to realize through text-based reviews.

Compared to the traditional challenges associated with prompting users to fill questionnaires [4], today's internet savvy users freely give away their opinions online and through social media platforms, mainly in an unstructured way [5]. Such online platforms are becoming the de-facto channels for reporting users' concerns and product acceptance at an "unprecedented scale in real-time" [6]. As a result, traditional well-defined product and service evaluation processes require more accommodation of the real-time and dynamic aspect of today's opinion sharing channels. While we are witnessing an increased research interest in opinion mining from text [7–10], most of the available approaches do not incorporate the existing structure of well-established product evaluation and user satisfaction models in their methodologies. Our research aims to close this gap, with a focus on the following research question: *how can we accommodate product and service evaluation models in the process of automatically analyzing unstructured users' reviews?*

To answer this research question, we propose in this article "*ReviewModus*", a model-based supervised machine learning framework, to assist with the automatic extraction and analysis of measurable variables from unstructured text in product and service reviews. The novelty of our approach stems from the augmentation of the evaluation process of unstructured users' reviews by using traditional questionnaire evaluation methods as a means for training a predictive machine learning algorithm. The framework learns to predict the classification of reviews into a pre-defined set of evaluation model dimensions, and to predict the degree of sentiments expressed in the reviews. The predicted classifications and sentiment

measures are then processed for further analysis and generating actionable insights. The feasibility of our approach is demonstrated by investigating the use of a neural network-based (NN) algorithm. The NN algorithm is trained on a set of existing survey questions for understanding the pre-defined dimensions of a user satisfaction model. It is coupled with a logistic regression algorithm trained on Amazon.com reviews for predicting and quantifying sentiments expressed in the reviews. The evaluation is conducted in the domain of e-government service assessment. In this context, we employ a user satisfaction model that evaluates e-government services around the Cost, Benefit, Risk and Opportunity dimensions (i.e., COBRA model) [2]. Our evaluation shows promising results. Our tests demonstrate a promising micro average F-score of 85.16% with respect to the multiclass prediction of model dimensions, and a high positive correlation of 71.44% with the assessment of sentiments performed manually by three evaluators.

The remaining parts of the article are structured as follows. Section 2 provides a review of related works in the field. Section 3 presents our model-based supervised machine learning framework. Section 4 focuses on the evaluation procedure in the domain of e-Government services. Section 5 presents our results, and Section 6 concludes with future research directions.

2. Review of User Opinion Analysis: Model and Machine Learning Perspectives

In this section we review existing machine learning and model-based approaches to assess users' opinions on products and services.

Turning information in text into "actionable knowledge" is increasingly getting research attention [11]. This attention is gaining momentum in various domains. For example, Reddick et. al. [12] investigate the impact of the analysis of text in social media on the delivery of public services; while Müller et. al. [13] study how text analytics can help in

better understanding customers' problems and requests for improving customer service. The efforts involved in text processing are pushing for the automation of tasks related to text analysis. Such tasks range from sentiment detection in user generated content on the web [8] and question answering [14,15], to classification and prediction [11,16], to name a few. In this context, machine learning is being more and more involved in performing text analysis functionalities. For example, neural networks were used for classifying text documents [17]; Support Vector Machines (SVM) were heavily employed in pattern recognition [18], in processing customer reviews and product opinions [19], and in feature-based text categorization [20]; statistical and evolutionary algorithms were tested for Part-of-Speech tagging [21]; and a Bayesian approach was used to model customer satisfaction from unstructured text [7]. While those approaches are proving to be effective, some of the challenges remain pertinent to the success or failure of those techniques, including the type of data in focus and required preparation, selecting the right classification approach, and the availability of appropriate training data used for the algorithm [22]. We focus on the challenge involved in providing the supervision and training needed for the success of supervised machine learning algorithms in text-based review analytics.

Parallel to the efforts invested in machine learning and feature-based analysis, another flourishing area of research is studying models to represent and capture various analytical contexts and objectives. For example, researchers have been working on designing models to represent user satisfaction in the context of public services [1,3] and e-Government [2,23,24], while others have focused on modeling usefulness of technology-driven solutions in more generic terms [25,26]. The aim of such related works is to come up with well-defined models to represent the situation being assessed as accurately as possible, for testing certain hypotheses. The developed evaluation models usually involve a set of inter-related dimensions consisting of a set of *inputs*, which are transformed into a set of *outputs*. For

instance, Parasuraman et. al. [1] highlighted the importance of *reliability* and *responsiveness* among other input dimensions to measure the service *quality* output dimension in their SERVQUAL model. For such models to perform well, analysts usually assess the importance of the model dimensions by controlling the collection of data around the dimensions to be evaluated. In the context of users' feedback, such dimensions are often controlled through a set of closed-ended questions to ensure a consistent and measurable assessment of users' perception of the dimensions in focus [27]. To ensure a holistic view of the participants' feedback, service and evaluation processes often provide an option for participants to express their opinions using an open-ended format, which are subsequently analyzed. This option is provided for various reasons including serving as a "safety net" in support for the closed questions of the survey, or seeking further information from the participants on uncovered aspects in the other structured parts of the survey questions [28]. However, the complexity involved in analyzing the open-ended feedback often results in having a substantial amount of untapped text data, which could provide additional insights for service and product improvement.

Hence the question is how could the consistency and robustness of existing models be employed to analyze unstructured users' reviews? While machine learning approaches on mining opinions from text are flourishing, we have seen little efforts on modeling frameworks that incorporate predictive algorithms in model-based approaches to assess text-based opinions. Our aim is to close this gap in the literature.

3. A Framework for Model-Based Supervised Machine Learning

As discussed in Section 2, model-based evaluation approaches provide a solid methodology and well-tested hypothesis for evaluating products and services around key dimensions and variables. We see an opportunity to exploit such model characteristics for automating

unstructured review analysis. Most of such models are proposed and assessed based on questionnaires that involve meticulously crafted closed-ended questions, coupled with Likert scale format answers [27]. For example in SERVQUAL, one of the first models proposed to evaluate services, five dimensions were considered in service evaluation, namely: *Tangibles*, *Reliability*, *Responsiveness*, *Assurance* and *Empathy* [1]. Such dimensions were then tested using a set of Likert scale questions. For instance, *Reliability* was tested based on five questions such as "when these firms promise to do something by a certain time, they should do so." Figure 1 provides an example of a connection between user's input and a model's dimension through a survey question.

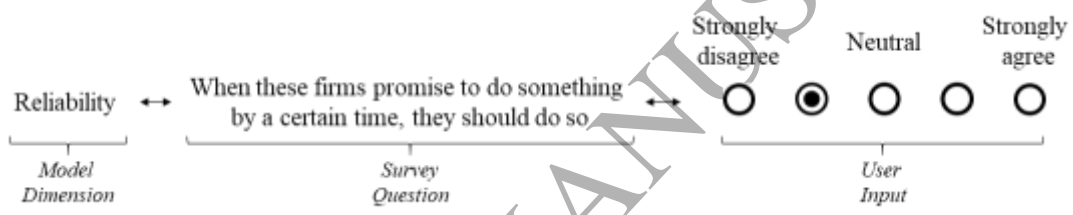

Figure 1 - Example of Collecting User Input on a SERVQUAL Model Dimension through a Closed-Ended Question

In the case of unstructured reviews, we assume that the presence of a model potentially gives an indication of how to interpret the reviews. In other words, the analyst can rely on the model to identify key elements to focus on while processing and coding reviews. The model provides the required semantics and structural mechanisms used for analyzing the service and product in focus. Semantics involve understanding the meaning and classification of content generated from the reviews; while the structural mechanisms involve the dynamics between the content elements. Irrespective of the types of supporting tools, we identify the need for a 3-phase framework to extract meaningful insights from text feedback around a pre-defined model:

- Phase 1: Classify text statements around the product or service variables that are represented in the pre-defined evaluation model.

- Phase 2: Identify the level of agreement and sentiment associated with the mentioned text statement.
- Phase 3: Extract quantifiable entities around text statements for analyzing, interpreting and mining insights.

In this context, to successfully perform Phase 1, the content analyst needs to have a clear understanding of the meaning of the evaluation model variables to be able to consistently classify the review statements. For example, a protocol with an explicit vocabulary for coding and classifying the statements can be developed for the analysts to follow. Concerning Phase 2, a clear methodology for sensing the level of agreement in a review statement is needed. For instance, a dictionary of keywords can be developed, or patterns in the text that indicate a support or disagreement with the stated text. With respect to Phase 3, the analyzed content must be translated through quantifiable measures including for example specified metrics around the evaluation model dimensions, such as the overall agreement level or other measures pertinent to the analytical goals. With the increase in the amount of text to analyze, performing these phases manually is a tedious and challenging task.

We propose in this article a "ReviewModus" framework, which combines model-based assessment and machine learning techniques to achieve new insights that cannot be obtained by either approach separately. The illustrated details of the framework are depicted in Figure 2. At a high level, the framework supports Phase 1 through the Classification Prediction component. Phase 2 is supported by the Sentiment Prediction component, and Phase 3 is realized by the Analytics component. The components are formed of three steps each.

The Classification Prediction component starts with "*Selecting the Algorithm*" that is appropriate to the task and the review data on hand. For example, in the case of classifying reviews along multiple classes, the classification prediction phase could involve the selection of a Neural Network, Support Vector Machine, Ensemble Learning, or k-Nearest-Neighbor algorithm as a potential option. The second step of the classification prediction component is

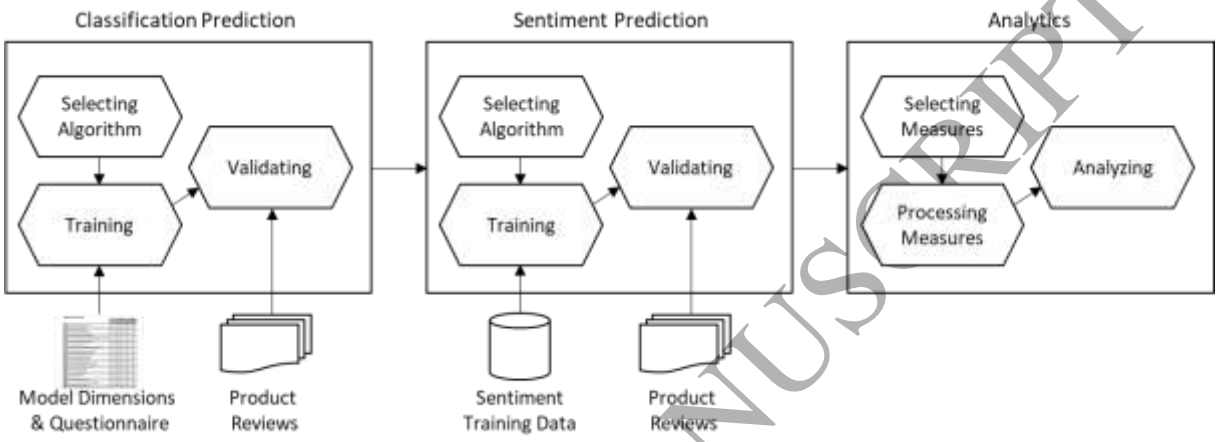

Figure 2 – ReviewModus Framework for Processing Product Reviews using Evaluation Models and Machine Learning

"*Training*" the selected algorithm on appropriate training datasets. For the classification prediction component, a new contribution of the framework lies at the level of using the model dimensions and questionnaire (i.e., survey questions) to supervise the learning process of the algorithm. The reason behind our proposition is that the closed-ended questions provide a solid context around the model dimensions, serving as a potential starting point for classifying segments around the dimensions. For instance, Figure 3 shows the list of questions used for evaluating the *Tangibles* and *Reliability* dimensions (i.e., two of the five dimensions) of services in the SERVQUAL model. As one can see, the questions serve as a potential contextual anchor of the model dimensions. For example in this model, which targets assessing the quality of services, *Reliability* potentially hints to good customer service and adequate records keeping (among others); compared to the reliability of a car which could reflect high mileage to breakdown ratio. Such clarification and distinction are equally important to the person or machine processing the text content. In the absence of such

questions, the analyst must spend time providing adequate explanation and training for understanding the context behind each dimension. This is usually achieved by generating a manually (and usually expensive) annotated dataset used for training algorithms. After training the algorithm, the last step in this component is "*Validating*" the results. This can be achieved by comparing the predictive abilities of the algorithm to a subset of pre-annotated data unused in the training step. If this is not feasible due to the unavailability of such pre-annotated data, the results can be manually validated through the creation of a gold standard used as a reference point to compare the algorithm's prediction output.

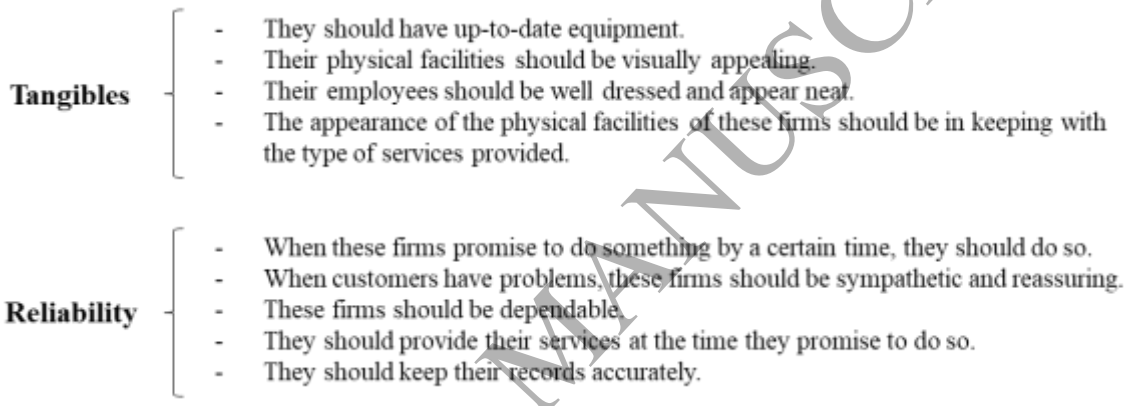

Figure 3 - Questions Used to Assess the Tangibles and Reliability Dimensions in SERVQUAL

The following phase in the framework is supported by the Sentiment Prediction component. Similar to the classification prediction component, the sentiment prediction component starts with "*Selecting the Algorithm*" appropriate to the sentiment detection approach to follow. For example, if a user is commenting on the reliability dimension of the service, it is expected to sense the opinion behind such a statement. If the approach involves a binary sentiment prediction (i.e., positive versus negative) then a binary logistic regression would be a potential algorithm to use. However, if the model imposes a detection of the degree of agreement with the stated review text, then the ordinal logistic regression would be a better fit for the task. After selecting the algorithm, the component involves "*Training*" the selected algorithm. Being context-dependent, the performance of sentiment detection depends on the selection of training data and review text characteristics [8]. For that the analyst should

carefully choose a training dataset that is contextually aligned with the reviews in focus. For instance, if the evaluation and reviews are around hotel services, a training dataset in this context (or close to this context) is advised to be used as input. Similar to the classification prediction component, the sentiment classification component ends with "*Validating*" the results through a new pre-annotated dataset or manually checked for accuracy.

Once the predictions are validated, the data is passed to the Analytics component. At this level, the analysis starts with "*Selecting the Measures*" from the processed text. This step is tightly related to the analysis objectives. For example, if the objective is to check the highest or lowest performing products or services, one can choose to select the sentiment measures at the level of products. After selecting the measures, a certain level of "*Processing*" and transformation of the algorithms' output measures might be needed to align with the adopted measures of the model. If a comparison along the model dimensions is required, then the classification measures can be added to the analysis. For example, one objective might be to identify from the reviews how a service *reliability* compares across different domains (e.g., banking, repair and maintenance) [1]. In such cases the opinions can potentially be aggregated and processed through averaging the quantified level of sentiments extracted from the text comments and compared across the different service domains. The last step in the analytics component is "*Analyzing*" the extracted measures. This step is key to make sense and generate insights out of the processed reviews. The choice of the analytical reports and visuals is tightly linked to the number of measures and objectives of the analysis. For instance, in the case of single dimension comparison (e.g., identifying the highest overall performing service), simple reporting and data rendering would fit. However in the case of a comparative analysis across multiple dimensions and measures, interactive reporting and visuals can be of great value for the task.

4. Evaluation Procedure: The Case of e-Government Service Assessment

We test the feasibility of our framework in the context of an e-government service evaluation process. The success of e-Government services has been increasingly playing a major role in today's governance objectives that call for decreasing the digital divide among its citizens. For this reason, we have been witnessing an increased interest in measuring the quality and degree of adoption of governments' electronic services. As a result, there has been a surge in studies that focus on designing and testing service evaluation models in this context [2,23,24]. However, we observe that existing approaches in this field either opt for a quantitative analysis of the model dimensions through a controlled Likert scale based closed ended questions (e.g., [2,23]), or opt for a qualitative and manual analysis of data generated from semi-structured interviews (e.g., [24]). To our knowledge, our approach provides a first attempt to employ the elements of an existing model to supervise and train machine learning algorithms to automate the analysis of unstructured text data available from users' feedback.

4.1. Evaluation Context and Data

We evaluate the framework on users' feedback collected on 11 e-Government services deployed in three countries, one European, another from the Gulf region and the third from the Middle East. As part of the evaluation process, we put to the test the COBRA model, used to assess users' satisfaction with e-Government services around the Cost, Benefit, Risk and Opportunity dimensions [2]. Figure 4 shows the structure of the COBRA model.

Figure 4 - COBRA Model for e-Government Service Evaluation and Satisfaction

The four COBRA categories are further broken down into two sub-categories each: Cost includes tangible and intangible costs; Benefit includes tangible and intangible benefits; Risk includes personal and financial risks; and Opportunity includes service and technological opportunities. A questionnaire was designed to collect users' feedback on the e-services in this context. The questionnaire includes 46 questions linked to the sub-categories of the COBRA model, coupled with a five-point Likert scale answer format. The questions are listed in Table 1. In addition to the closed-ended questions, the users were asked an open-ended question to provide feedback in a comment textbox limited to 256 characters, with the objective to capture additional elements not covered in the questionnaire. The question is formulated as follows: "Do you have any further comments on this e-service? Please feel free to do so below." A total of 3,119 responses were collected from the three countries. We cleaned the comments (e.g., removed the ones that contained less than 3 words and duplicated statements), translated the non-English comments to English, segmented the comments into stand-alone concise statements to reflect the key idea behind the comments, and glossed them for a consistent representation. We ended up with a dataset of 1,492 text statements of e-Government reviews that we used in this study.

4.2. Performing Phase 1: Predicting the Classification of Reviews around the COBRA Model Dimensions

In our proposed framework, we investigate the feasibility of training the machine learning algorithm using the questions related to the COBRA dimensions. We propose using the survey questions as background knowledge for explaining the dimensions, as they can potentially be used by analysts (and algorithms) to cognitively make sense of the analytical model dimensions to further analyze the collected data. While many machine learning methods for text classification techniques exist (e.g., k-Nearest-Neighbor [29], Support Vector Machine [30], Ensemble Learning [31]), our objective here is to test and assess one approach that can be applied as a proof of concept for testing the feasibility of our framework components with the aim to answer our initial research question. This can reflect the potential of using the structural nature of evaluation models, for performing automated machine learning-based text reviews classification. We selected and tested a neural network-based algorithm, which we trained on the e-government survey questions, to predict the classification of the text segments.

Question	**Model Dimension**
Using the e-service saved me time	Tangible Cost
Using the e-service saved me money	Tangible Cost
The service removes any potential under table cost to get the service	Tangible Cost
The service reduces the bureaucratic process	Tangible Cost
The password and renewal costs of service are reasonable	Tangible Cost
The internet subscription cost is reasonable	Tangible Cost
The service reduces my travel costs to get the service	Tangible Cost
It takes a long time to arrange an access to the service	Intangible Cost
It takes a long time to load the service homepage	Intangible Cost
It takes a long time to find my needed information	Intangible Cost
It takes a long time to download and fill the service application	Intangible Cost
It takes several attempts to complete the service due to system breakdowns	Intangible Cost
It takes a long time to acknowledge the completion of service	Intangible Cost
The service is easy to find	Tangible Benefit
The service is easy to navigate	Tangible Benefit
The description of each link is provided	Tangible Benefit
The service information is easy to read (font size, color)	Tangible Benefit
The service is accomplished quickly	Tangible Benefit
The service requires no technical knowledge	Tangible Benefit
The instructions are easy to understand	Tangible Benefit
The service information is well organized	Tangible Benefit
The drop-down menu facilitates completion of the service	Intangible Benefit

New updates on the service are highlighted	Intangible Benefit
The requested information is uploaded quickly	Intangible Benefit
The service information covers a wide range of topics	Intangible Benefit
The service information is accurate	Intangible Benefit
The service operations are well integrated	Intangible Benefit
The service information is up-to-date	Intangible Benefit
The referral links provided are useful	Intangible Benefit
I am afraid my personal data may be used for other purposes	Personal Risk
Using the service leads to fewer interactions with people	Personal Risk
The service obliges me to keep a record of documents in case of future audit	Financial Risk
The service may lead to a wrong payment that needs further correction	Financial Risk
I worry about conducting transactions online requiring personal financial information	Financial Risk
The Frequently Asked Questions (FAQs) are relevant	Service Opportunity
The provided multimedia services facilitate contact with service staff	Service Opportunity
I can share my experiences with other service users	Service Opportunity
The service can be accessed anytime	Service Opportunity
The service can be reached from anywhere	Service Opportunity
The information needed for using the service is accessible	Service Opportunity
The service points me to errors during a transaction	Technology Opportunity
The service allows me to update my records online	Technology Opportunity
The service offers tools for users with special needs (touch screen)	Technology Opportunity
The information is provided in different languages (Arabic, English)	Technology Opportunity
The service provides a summary report on completion	Technology Opportunity
There is a strong incentive for using e-services	Technology Opportunity

Table 1 - List of Survey Questions used around the COBRA Model Dimensions

We briefly describe here our implementation of the classification prediction component. First, the e-Government reviews' text segments are pre-processed by changing all words to lower case. Second, the segments were stemmed to merge the different words variations. Third, we removed duplicate words from the e-Government reviews dataset. We rely on the traditional bag-of-words model that doesn't consider the order of words in the statements [22]. The bag-of-words model is used to transform the text statements into a set of matrices, which are passed as input to the neural network model. Armed with the universality theorem stating that a single hidden neural network can potentially solve any continuous function with a certain degree of approximation [32], we adopted in our evaluation a two-layer neural network with one hidden layer. This can be extended in the future to investigate other configurations of neural network algorithms such as adding further hidden layers. Our code builds on the Python implementation provided by Trask [33,34] and Kassabgi [35] to

perform our initial tests and analysis[1]. One advantage of using their neural network implementation is the ability to trace and fine-tune the code elements whenever needed. Our training data from the COBRA model related questions generated a vocabulary of 162 unique words. We show in Figure 5 the structure of the neural network, having 162 neurons as input layer (i.e., based on the 1x162 matrix representing the 162 words extracted from the training dataset), 20 neurons set at the hidden layer, and 8 neurons at the output layer reflecting the 8 COBRA dimensions. We relied on the sigmoid activation function $f(x) = \frac{1}{1+e^{-x}}$ to perform a forward propagation of weights $w_{i(a,b)}$ along the synapses on layers i (0 and 1) from neuron a to b. We performed the training of the algorithm using the 46 survey questions from the questionnaire, and validated the results on the 1,492 e-Government reviews' text segments that were checked by three analysts who evaluated the algorithm's output independently.

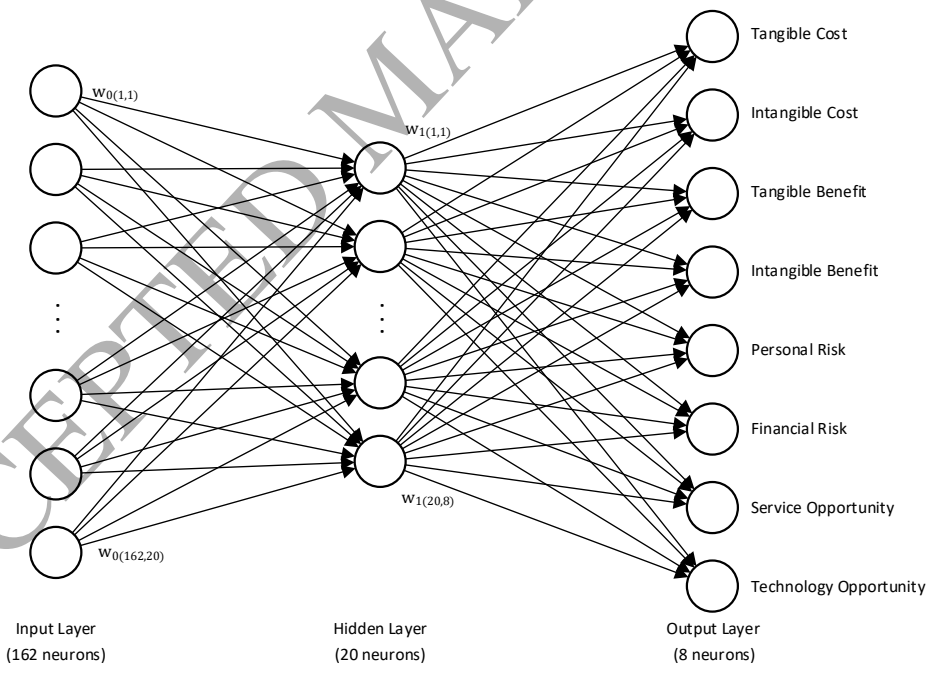

Figure 5 – Two-Layer Neural Network Diagram Used in our Evaluation

4.3.Performing Phase 2: Predicting the Level of Sentiments in the Reviews

[1] The code can be downloaded from the following link:
http://fouad.zablith.org/code/reviewmodus_algorithms_code.zip

The second phase after text classification is to predict the level of sentiment in the reviews. To align this task with the Likert scale agreement followed in the questionnaire-based data collection around the COBRA model, we need to replicate this task and automatically generate the *degree* of polarity in the review. In other words, our objective is not to have a binary classification of positive or negative sentiment in the statement, but rather the degree of positivity from a scale of one to five, where one is on the negative side (i.e., strongly disagree) and five is on the positive side (i.e., strongly agree).

To fulfill this objective, we surveyed various existing machine learning algorithms used to evaluate sentiments from unstructured text. Sentiment analysis is a major task in text mining, and various machine learning techniques have been proposed to handle it. Vector space models are found to be one of the most common approaches tested in the field to provide promising results. For example Pang and Lee [36] proposed the use of Support Vector Machine (SVM) that captures bag-of-words to detect sentiments tested on the movies review dataset. Vector based models were also improved when complemented with term frequencies weighting techniques that reflect the importance of a term relative to the other terms in the vector space [37]. Such models were used for training and validating different classification techniques (e.g., logistic regression [38]) that showed promising results when applied to sentiment analysis [39]. Inspired by the potentials behind those techniques, we decided to put them to the test in our context.

We implemented a logistic regression-based algorithm to train a classifier to predict sentiments in our experimental e-Government dataset. In short, we broke the e-Government text reviews into bag-of-words, transformed them into a vector-based representation, extracted the features from the text, enhanced the classification by introducing a term frequency-inverse document frequency (tf-idf) term weighting [22], and employed an *n*-gram [40] model to take the co-occurrence of words into account (e.g., not happy versus happy).

We rely on the scikit-learn packages [41] and implement our code in Python[2]. We follow a similar logic proposed in the implementation of Li [42]. Given the ordinal type (i.e., 5-level Likert scale) of sentiment followed in the COBRA model, we put to the test an *ordinal-regression* classification model to predict the Likert scale value y (i.e., $y \in \{1 \rightarrow 5\}$), for a given statement x using the following cumulative probability formula: $P(y \leq j|x) = \frac{1}{1+e^{(wx-\theta_j)}}$ [43], where j is the number of classes (i.e., 5 in our case), thresholds θ_j and coefficients w are determined after training the classifier on the data.

We also tested a *binary-mapped* logistic regression with a 5-scale mapping classifier to predict the classification of positive versus negative statements. We used the regular logistic probability formula in this case ($P(x) = \frac{1}{1+e^{-(wx+\theta)}}$), reflecting the probability of a statement x being a positive statement. However, in order to align the algorithm's probability value with the needed 5-scale agreement levels, we did an indirect translation of the generated probability from the algorithm into one of the five equal intervals of measures: a 0%-20% positive probability interval was given an agreement level of 1 (i.e., strongly disagree); a 21%-40% interval was given an agreement level of 2 (i.e., disagree); a 41%-60% interval was given an agreement level of 3 (i.e., neutral); a 61%-80% interval was given an agreement level of 4 (i.e., agree); and 81%-100% interval was given an agreement level of 5 (i.e., strongly agree).

One of the challenges we faced in our scenario is the choice of an appropriate training dataset for sentiment analysis. Our e-Government reviews dataset exhibits the following properties. First, the e-Government review statements are short with a very limited context. Second, the e-Government service domain is a relatively narrow domain compared to discussions pertaining to social media platforms. We initially thought of training our

[2] The code can be downloaded from the following link:
http://fouad.zablith.org/code/reviewmodus_algorithms_code.zip

algorithm on the widely used Twitter datasets for sentiment analysis. But then we anticipated that this would introduce a substantial amount of noise, due to the informal nature of discussions on Twitter, which is not the case in our data. Then we decided that one source of training data that is close to our context is the Amazon.com reviews dataset. We got hold of the data prepared and cleaned by the Stanford Network Analysis Platform (SNAP) team [44]. One of the advantages of the collected Amazon.com reviews is that they were collected based on product categories. One of the closest categories to our domain is the software domain, in which reviewers might comment on features that reflect similar functionalities to online government services (e.g., performance, interface, etc.) The software products' related reviews contained 95,084 Amazon.com reviews. The dataset included different fields such as users' ratings from 1 to 5, product ID, full reviews, summary reviews, and others. Figure 6 shows an example of an Amazon.com review highlighting the summary versus the full review texts. We first trained our classifiers on the full Amazon.com reviews, followed by a training on the Amazon.com reviews' summary. We then compared their performance to the manual sentiment classification performed by three independent analysts.

Figure 6 - Amazon.com Review Example Showing the Difference between the Review Summary and the Full Review

4.4. Performing Phase 3: Extracting Quantifiable Measures around Text Statements for Analyzing and Generating Insights

The last phase in our framework is to make sense of the measures extracted from the text segments. In our scenario we consider the following analytical objective: how the review-based measures can potentially help e-service providers in each of the three countries by identifying potential service improvements based on a cross-service comparative analysis.

To achieve this objective, one might be interested in measuring the *average sentiment* expressed in the reviews, computed around the *COBRA dimensions*, over the *11 e-services* across the *3 countries*. We customized our algorithm to generate the processed data extracted from the e-Government reviews into a comma separated file for an easy import into visual and analytical tools (e.g., Excel, Tableau or others). Figure 7 shows an example of a dynamic report using the Pivot Table feature in Excel, extracted from the e-Government reviews dataset. This report computes the average polarity measure detected from the unstructured comments, with respect to the 11 e-Government services in the 3 countries and the COBRA model dimensions. At a country level, one can see that Country 1 had a higher overall sentiment of 4.39 detected from the service reviews compared to the other countries. At a service level, this report shows that, Service 5 had the lowest overall average sentiment score of 3.22 expressed in the unstructured reviews compared to the other services.

Average of Polarity	Column Labels									
Service	Tangible Cost	Intangible Cost	Tangible Benefit	Intangible Benefit	Financial Risk	Personal Risk	Service Opportunity	Technology Opportunity	No Category	Grand Total
Country 1	**4.60**	**4.00**	**4.59**	**4.38**			**2.50**	**3.00**	**3.00**	**4.39**
Service 1	4.45		4.50	4.00				3.00	2.00	4.22
Service 2	4.71		4.48	4.33			2.50		3.33	4.37
Service 3	4.43	4.00	5.00	5.00						4.69
Country 2	**4.22**	**2.64**	**4.24**	**3.94**	**2.33**	**1.00**	**2.69**	**3.00**	**4.10**	**3.84**
Service 4	3.61	1.71	4.09	3.61	2.33	1.00	2.47	2.44	4.10	3.44
Service 5	4.50	3.00	3.43	2.88			2.80	3.40	3.00	3.22
Service 6	3.50	5.00	4.67	4.63			5.00	4.00		4.50
Service 7	4.54	3.00	4.57	4.39			2.75	3.63	4.83	4.31
Service 8		5.00	5.00	4.50						4.80
Country 3	**4.03**	**3.23**	**4.42**	**3.75**	**3.50**	**4.00**	**2.82**	**2.89**	**2.81**	**3.97**
Service 9	3.67	3.50	4.41	3.80	2.00	4.00	3.14	2.78	2.71	3.93
Service 10	4.25	3.29	4.32	3.55	5.00		2.73	2.10	3.00	3.78
Service 11	4.50	2.50	4.51	3.95			2.50	4.29	2.75	4.22
Grand Total	**4.28**	**2.96**	**4.38**	**3.90**	**2.80**	**2.50**	**2.72**	**2.94**	**3.40**	**3.96**

Figure 7 – Table Visualization Example of the Average Sentiments Predicted by the Algorithm from Review Statements around the COBRA Model Dimensions and the 11 e-Government Services in the 3 Countries

With the presence of such dynamic reports, one can dig deeper to identify the statements behind the numbers generated in the report. Figure 8 shows a subset of statements

that generated an average predicted polarity of 4.71 for Service 2 with respect to Tangible Cost.

Reference	Country Labe	Service Labe	Review Statement	Classificatio	Polarity
789	Country 1	Service 2	The citizen could get his driving license and residence permit through the service	Tangible Cost	3
760	Country 1	Service 2	The ISP is charging high fees for the internet	Tangible Cost	4
716	Country 1	Service 2	The service simplifies the method of obtaining the information for the citizen and it saves his time which makes him satisfied	Tangible Cost	5
688	Country 1	Service 2	the service saves time, effort and reduces traffic	Tangible Cost	5
685	Country 1	Service 2	The service saves time, effort and money	Tangible Cost	4
676	Country 1	Service 2	The service saves time since it is done online which makes the citizen satisfied	Tangible Cost	5
667	Country 1	Service 2	the service saves time and simplifies the procedure that should be done	Tangible Cost	4
652	Country 1	Service 2	The service saves time and money and it educates citizens in using electronic services	Tangible Cost	5
648	Country 1	Service 2	The service saves time and meets all the citizen's main needs	Tangible Cost	5
644	Country 1	Service 2	The service saves time and executes the request fast thus the citizen is satisfied	Tangible Cost	5
641	Country 1	Service 2	The service saves time and effort when submitting documents and applications.	Tangible Cost	5
640	Country 1	Service 2	The service saves time and effort thus the citizen is satisfied.	Tangible Cost	5
638	Country 1	Service 2	The service saves time and effort of going to the government offices which makes the citizen satisfied.	Tangible Cost	5
633	Country 1	Service 2	the service saves time and effort and simplifies finishing the paperwork.	Tangible Cost	5
631	Country 1	Service 2	The service saves time and effort	Tangible Cost	5
627	Country 1	Service 2	The service saves time	Tangible Cost	5
626	Country 1	Service 2	The service saves the transportation cost of going to the office thus the citizen is satisfied	Tangible Cost	5
619	Country 1	Service 2	The service saves the citizen time of waiting in a queue	Tangible Cost	4
601	Country 1	Service 2	The service saves money and time	Tangible Cost	4
598	Country 1	Service 2	The service saves a lot of time thus the citizen is satisfied	Tangible Cost	5
596	Country 1	Service 2	the service saves a lot of time and effort thus the citizen is satisfied	Tangible Cost	5
26	Country 1	Service 2	The citizen can save his documents through the service and present them at any time so this saves his time therefore he is satisfied	Tangible Cost	5

Figure 8 - Sample of Statements Behind the Predicted Polarity of Service 2 with Respect to the Tangible Cost Dimension

It is worth noting how the trained neural network algorithm predicted that "saving time and effort (e.g., Ref. 631 in Figure 8)" is highly positive, without having explicit terms and adjectives reflecting positivity in the statements. We further discuss the performance of the classification and sentiment detection algorithms in the next section. Figure 9 shows another set of statements behind the low polarity of Service 4 around the Technology Opportunity dimensions. This granular view of text statements can potentially help in suggesting recommendations for improving this e-service. As reflected in the list, there are statements (e.g., Ref. 161 in Figure 9) where the algorithm missed predicting the right polarity. In this example it seems that the algorithm picked on the terms "better offers" and "speeds" to infer a high polarity, while it missed detecting that such positive features are lacking in the service as the citizen is "waiting" for them to happen. Another example of misclassification happened for example with statement Ref. 144 in Figure 9. This is a similar case where "new website" mentioned in the statement seems to have increased the predicted polarity, while in fact it was not part of the service as the user was "asking for this" to

happen. Further investigations should happen at the level of better training the algorithm to detect cases like this. We discuss further potential improvements of our approach in the Discussion and Conclusion section.

Reference	Country Labe	Service Labe	Review Statement	Classification	Polarity
1	Country 2	Service 4	A mobile version of the website should be created for the loading to be faster on mobile devices	Technology Opportunity	1
883	Country 2	Service 4	The service allows the citizen to know his bill in advance thus making the payment method faster	Technology Opportunity	5
863	Country 2	Service 4	The citizen wants the speed of the internet and the capacity allowed for downloading to increase.	Technology Opportunity	2
534	Country 2	Service 4	The service only uses the website to check for his next bill but there is always delay in showing the bills	Technology Opportunity	1
6	Country 2	Service 4	even though the citizen has a direct debit phone line, the line is disconnecting because the payment is not being reached to the targeted party.	Technology Opportunity	3
275	Country 2	Service 4	The service doesn't allow the citizens to pay online	Technology Opportunity	1
274	Country 2	Service 4	The service doesn't allow the citizen to update or cancel	Technology Opportunity	1
192	Country 2	Service 4	The ISP provides limited download and upload capacity	Technology Opportunity	1
175	Country 2	Service 4	The citizen wants to be able to pay online by a visa card	Technology Opportunity	1
161	Country 2	Service 4	The citizen is waiting for better offers and speeds in using and requesting the service	Technology Opportunity	5
13	Country 2	Service 4	Other services provide more options	Technology Opportunity	3
144	Country 2	Service 4	The citizen is asking for a new website	Technology Opportunity	4
34	Country 2	Service 4	The citizen cannot see information related to his bill during the updating period	Technology Opportunity	3
16	Country 2	Service 4	The billing system should be updated everyday.	Technology Opportunity	3

Figure 9 - Sample of Statements Behind the Polarity of Service 4 with Respect to the Technology Opportunity Dimensions

5. Results and Findings

In this section, we present the results and findings on model classification and sentiment predictions in our e-Government scenario.

5.1. Model Dimension Classification Performance

One of the challenges we faced in our work is that, to our knowledge, similar data settings used to classify open-ended text around an evaluation model are not available. Hence, it was not feasible to find external approaches and testing performance results to benchmark our approach to. For that we rely on the manual intervention from human evaluators to assess the quality of the classification using Precision, Recall and F-score that are widely adopted in machine learning-based text classification tasks [45]. The performance results of our applied neural network-based classification are validated by having three assessors who manually and independently evaluated the classification results of the 1,492 e-Government reviews' text statements. The evaluators were given the COBRA-based survey questions to get familiar with its eight classification dimensions. The evaluators were asked to mark independently each classification predicted by the NN algorithm as True or False. In the case where a

statement was marked false, the evaluators were required to propose the right classification. Having three evaluators independently evaluate the text statements, has enabled us to identify the level of agreement between the evaluators. This helped in generating a cross-checked and validated dataset that will serve as a base-line to compare the performance of the NN algorithm.

Out of the 1,492 statements, 754 were agreed by the three evaluators as correctly classified, 176 were agreed to be wrongly classified, resulting in 62% agreement level between the three evaluators. We focus our analysis on the 930 statements that were agreed by the three evaluators. For the 176 wrongly classified statements, 29 statements had an agreed alternative category, hence we assumed that the remaining 147 statements did not belong to an appropriate COBRA dimension. These 147 statements are of great value to our scenario, as such unclassified statements potentially represent some concerns or satisfaction with service elements that are not captured by the COBRA model. Figure 10 shows a sample of statements that were not classified. We identify three scenarios that could occur from those unclassified cases. First, some unclassified statements might reflect a missing question or category in the evaluation model. For example, the statement that mentions “the service lacks a lot of options in order to upgrade or customize the service in a way that benefits the citizen not the government” is not reflected in the questionnaire questions. Such statements are a potential source for further understanding users’ perception in the reviews, which were not captured by the initial study. Those examples are valuable for improving the design of questionnaire and model structure through either adding new questions or proposing new categories to the evaluation model. In cases where the set of unclassified examples is substantial, automating the process of proposing new categories can be performed through for example applying cluster analysis techniques to the unclassified text statements. Second, some unclassified statements might be irrelevant to the study being performed. For example,

the review highlighting that “the citizen finds that some of the taxes should be abolished” is beyond the scope of evaluating the online aspect of e-Government services. Such cases can be ignored as they are not aligned with the objectives of the study and the models involved. Third, missing classifications might reflect a misalignment between the semantics of the text reviews and the evaluation model questions. For example, the statement mentioning that “the service doesn’t send confirmations” is semantically close to the already existing question in the questionnaire “the service provides a summary report on completion”. This might require the use of deeper semantic analysis using for example dictionaries to enable the classification of such statements.

Review Statement
The service lacks a lot of options in order to upgrade or customize the service in a way that benefits the citizen not the government
The service doesn't send confirmations
The service does not show all the payment transactions regularly
The service confuses the citizen when making payments
The service allows the citizen to get what he wants from one location
The citizen would like to have alternative ways of using the service not only online means
The citizen was not able to change details online
The citizen finds that some of the taxes should be abolished

Figure 10 - Sample of Unclassified Statements

Figure 11 shows the frequencies of statements classified and validated by the evaluators around the COBRA model dimensions. We can see that the tangible benefit and cost categories dominate the other intangible dimensions in the comments collected from the users. However the risk-related statements were relatively absent from the processed data. A potential explanation for this is that there were no comments on financial and personal risks as they were adequately captured in the survey questions, which did not prompt users to engage and further discuss them in the open-ended questions.

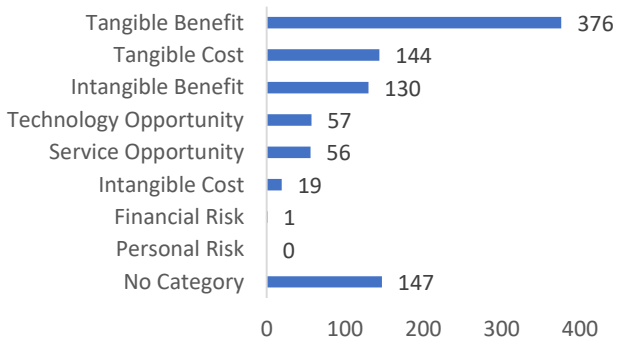

Figure 11 - Frequency of Statements Validated by the Three Evaluators around the COBRA Model Dimensions

To get further insights behind the performance of the classification task, we compute the Precision, Recall and F-score for each model dimension classification, coupled with the micro-average scores across all the model dimensions [45]. We present in Table 2 the formulas used in this context. We calculate Precision (P) using Formula (a), considering the ratio of statements that have been correctly classified (i.e., True Positive *tp*), to the sum of True Positive and False Positive (*fp*). Recall (R) in Formula (b) is based on the ratio of True Positive, to the sum of True Positive and False Negative (*fn*). The F_score (F) in Formula (c) combines the precision and recall by assigning in our case an equal weight to both. For an overall Precision, Recall and F_score across the multiclass classification we adopt the micro-average scores in Formula (d), (e) and (f) respectively.

Formula for Individual Class Classification	**Formula for Multiclass Classification**
(a) $Precision\ (P) = \frac{tp}{tp+fp}$	(d) $Precision_{\mu}(P_{\mu}) = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$
(b) $Recall\ (R) = \frac{tp}{tp+fn}$	(e) $Recall_{\mu}(R_{\mu}) = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$
(c) $F\_score\ (F) = \frac{2*P*R}{P+R}$	(f) $F\_score_{\mu}(F_{\mu}) = \frac{2*P_{\mu}*R_{\mu}}{P_{\mu}+R_{\mu}}$

Table 2 – Formula Used for Computing the Classification Precision, Recall and F-score

With the predicted classification, the NN algorithm provides a probability of the statement belonging to certain category. As part of the performance evaluation, we had to determine the right threshold above which the proposed statement classification is accepted,

and below which the statement is considered not classified. For example if a statement is predicted to be classified as Tangible Cost with probability 0.3, a threshold of 0.4 will dismiss this prediction and mark the statement as unclassified. To determine the optimal threshold, we implemented an optimization model in which we maximized the micro-average multiclass F-score (Formula f), by changing the threshold *t* (i.e., decision variable), and subject to the constraints limiting the threshold to be between zero and one as follows:

$$\boldsymbol{Max}: F_{\mu} = \frac{2 * P_{\mu} * R_{\mu}}{P_{\mu} + R_{\mu}}$$

By changing threshold *t*

Subject to:

t <= 1

t >= 0

We solved the model using an evolutionary solving method to get the appropriate threshold that maximized (F_{μ}). We present in Table 3 the performance results of the automatic classification around the COBRA dimensions. We see that the algorithm did not perform equally on the different COBRA categories. The major difference is in the F-score of detecting personal and financial risks from the data. These two categories had much lower F-scores compared to the other classifications. One potential explanation for this could be the fact that, as presented in Figure 11, these two risk-related categories were not prominently present in our validation dataset. Our results show that our algorithm suffered in terms of recall on identifying the statements that were not agreed by the three evaluators to belong to a certain COBRA dimension (i.e., not classified). In other words, the algorithm was less sensitive to identifying non-classified segments (c.f. Figure 10 for a sample of non-classified statements). However, the classification around all the other COBRA categories performed well in terms of precision and recall. The overall micro-average F-score for our scenario turned out to be a promising 85.16%.

	Tangible Cost	Intangible Cost	Tangible Benefit	Intangible Benefit	Personal Risk	Financial Risk	Technology Opportunity	Service Opportunity	No Classification
Precision	0.91	0.98	0.91	0.73	-	0.20	0.79	0.92	0.88
Recall	0.88	1.00	0.98	1.00	-	1.00	0.92	1.00	0.27
F-score	0.89	0.81	0.94	0.84	-	0.33	0.85	0.96	0.42

Table 3 - Precision, Recall and F-score of the Classification Prediction of Text Statements around the COBRA Dimensions using a two-Layer Neural Network

5.2. Sentiment Detection Performance

To assess the performance of the sentiment and polarity detection from text, we asked three additional evaluators to independently and manually assess the sentiment for each text statement by giving it a score ranging from -10 for highly negative, to +10 for highly positive statements. We opt for this wide range of sentiment assessment to give more flexibility to the evaluators to assess the degree of sentiments. We used the three scores given by the three evaluators to calculate the average polarity for each statement. We then used this manually processed data to compare it to the ordinal-regression and binary-mapped regression algorithms. To obtain the estimates of the weights and threshold of the regression models, the algorithms were trained first on the Amazon.com full reviews text (i.e., the full version of the reviews left by users as shown in Figure 6), and second on the summary of reviews text (i.e., the summarized header of the reviews left by users).

The ordinal-logistic regression algorithm, when trained on the Amazon.com full-text reviews, had a low 36.49% correlation with the manually processed data. However, when trained on the summary review text of Amazon.com, the correlation improved to 65%. One possible interpretation of this improvement is that, given the nature of our validation dataset that mainly includes short text statements, the short summary version of the Amazon.com reviews was more effective in training the algorithm and providing results that better correlated with the manually processed data. Concerning the binary-mapped logistic regression with 5-scale mapping, when trained on the Amazon.com full-text reviews, it correlated 63.02% with the manually processed data. However, when the algorithm was

trained on the summary reviews on Amazon.com, the correlation has increased to 71.44%, reflecting a high positive correlation. Table 4 summarizes the correlation results.

Algorithm	**Full-Text Reviews**	**Summary Reviews**
Ordinal-Regression	39.49%	65%
Binary-Mapped Regression	63.02%	71.44%

Table 4 - Comparison of Sentiment Prediction Algorithms' Correlation with the Manually Processed Data

One potential explanation behind the better performance of the binary-mapped regression, is that we removed the neutrally ranked statements (i.e., ranked 3 in the scale 1 - 5) from the Amazon.com training dataset, in order to create stronger classification classes being either positive or negative only. The removal of borderline sentiment cases from the training dataset seems to have improved the prediction sentiment performance of the binary-mapped regression. However, in the case of the ordinal-regression algorithm, we had to train the algorithm on all the sentiment rating classes, including the neutral ones. This might also reflect that when writing open-ended reviews, people might express in the text highly negative or positive expressions, but still rank the product moderately. Further work will be required to support this assumption.

6. Discussion and Conclusion

We presented in this article our approach for automating the analysis of unstructured text reviews around product and service evaluation models. We proposed ReviewModus, a framework that (1) predicts the classification of unstructured text product and service reviews around existing model dimensions using machine learning algorithms trained on the closed-ended survey questions; (2) predicts the sentiments from text using an algorithm trained on external review sources; and (3) converts the entities from text into quantifiable variables used as input for further analysis and insights generation. The evaluation of the framework shows promising results, reflecting the potential use of machine learning algorithms to bridge

the gap between the loose nature of unstructured text review analysis around the well-tested product and service evaluation models.

Our approach can benefit from further improvements at different levels. First, while the performance of the classification algorithm was promising, such results might fluctuate. This is largely due to the initial randomization of weights applied on the synapses of the neural network, coupled with the small training dataset used in our scenario. Our work can be extended to test and compare the performance of different machine learning techniques including for example Support Vector Machine [30], Ensemble Learning [31], and further neural network-based configurations such as deeper neural networks, Long Short Term Memory (LSTM) [46] or Character Level Convolutional Networks [47], and the possibility of using transfer learning [48] from different product evaluation and user satisfaction model scenarios. Second, while this work was tested in the context of e-Government services, the framework can benefit from further tests to perform in other domains. One interesting aspect is to study how the product and service context might impact the performance of the review analysis tasks. For example, in some contexts the designed questionnaires and related model might be high level, compared to reviews that capture more granular feedback, making the classification task more challenging. Third, one challenge related to the use of machine learning is the inability to investigate how results have been generated. A potential extension to our work is to complement this machine learning based approach with external background knowledge sources to provide a degree of reasoning behind the classification and sentiment analysis tasks. Fourth, further investigation is required at the level of improving the performance of our tested algorithms. As shown in the previous statement samples (e.g., Figure 9), the algorithm mis-classified some review statements. We anticipate that this can largely be due to the training phase of the algorithm used. One potential way to address this limitation is to investigate in the future the option of implementing a feedback loop during

the training step of the algorithm, coupled with a monitoring process of the improvement of the fitness of the algorithm above the current acceptable correlation level of 71.44%.

In addition to further improving our approach, we are planning as part of our future work to check how the results extracted from the reviews correlate with the quantitative analysis performed through the survey questions around the COBRA dimensions. This presents a good research opportunity, given our access to both structured and unstructured data emanating from the same users assessing the same services. It would be interesting to investigate how these two approaches complement each other, and ultimately see to what degree can unstructured review analytics lift the burden imposed by conducting traditional survey methods.

To conclude, we see our proposed approach as a mean for augmenting the analyst's ability to make sense of the increasingly available unstructured user feedback guided by product evaluation and satisfaction models. One of the major contributions of our work is that it can possibly help uncovering hidden insights that were not initially captured by closed-ended questionnaires and pre-designed models. Furthermore, our proposed approach can potentially help improving and refining models that are more aligned with users' expectations from products and services.

Acknowledgements

We would like to thank the editors and anonymous reviewers for their suggestions and invaluable feedback for improving our article. This work was supported by NPRP grant # [NPRP 09 - 1023 - 5 – 158] from the Qatar National Research Fund (a member of Qatar Foundation), and the University Research Board (URB) grant at the American University of Beirut.

References

[1] A. Parasuraman, V.A. Zeithaml, L.L. Berry, SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality, Journal of Retailing; Greenwich. 64 (1988) 12–40.
[2] I.H. Osman, A.L. Anouze, Z. Irani, B. Al-Ayoubi, H. Lee, A. Balcı, T.D. Medeni, V. Weerakkody, COBRA framework to evaluate e-government services: A citizen-centric perspective, Government Information Quarterly. 31 (2014) 243–256. doi:10.1016/j.giq.2013.10.009.
[3] Alessandro Ancarani, Towards quality e-service in the public sector: The evolution of web sites in the local public service sector, Managing Service Quality. 15 (2005) 6–23. doi:10.1108/09604520510575236.
[4] S.D. Crawford, M.P. Couper, M.J. Lamias, Web Surveys: Perceptions of Burden, Social Science Computer Review. 19 (2001) 146–162. doi:10.1177/089443930101900202.
[5] V. Dhar, Data science and prediction, Communications of the ACM. 56 (2013) 64–73.
[6] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter? — An empirical investigation of panel data, Decision Support Systems. 45 (2008) 1007–1016. doi:10.1016/j.dss.2008.04.001.
[7] M. Farhadloo, R.A. Patterson, E. Rolland, Modeling customer satisfaction from unstructured data using a Bayesian approach, Decision Support Systems. 90 (2016) 1–11. doi:10.1016/j.dss.2016.06.010.
[8] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval. 2 (2008) 1–135. doi:10.1561/1500000011.
[9] Atika Qazi, Ram Raj, Glenn Hardaker, Craig Standing, A systematic literature review on opinion types and sentiment analysis techniques: tasks and challenges, Internet Research. 27 (2017) 608–630. doi:10.1108/IntR-04-2016-0086.
[10] S. Moghaddam, M. Ester, Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2010: pp. 1825–1828. doi:10.1145/1871437.1871739.
[11] R. Gopal, J.R. Marsden, J. Vanthienen, Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining, Decision Support Systems. 51 (2011) 727–731. doi:10.1016/j.dss.2011.01.008.
[12] C.G. Reddick, A.T. Chatfield, A. Ojo, A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use, Government Information Quarterly. 34 (2017) 110–125.
[13] O. Müller, S. Debortoli, I. Junglas, J. vom Brocke, Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data, MIS Quarterly Executive. 15 (2016) 243–258.
[14] S.P. Conlon, B.J. Reithel, M.W. Aiken, A.I. Shirani, A natural language processing based group decision support system, Decision Support Systems. 12 (1994) 181–188. doi:10.1016/0167-9236(94)90002-7.
[15] D. Ravichandran, E. Hovy, Learning surface text patterns for a question answering system, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002: pp. 41–47. http://dl.acm.org/citation.cfm?id=1073092 (accessed November 14, 2016).
[16] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR). 34 (2002) 1–47.
[17] B. Yu, Z. Xu, C. Li, Latent semantic analysis for text categorization using neural network, Knowledge-Based Systems. 21 (2008) 900–904. doi:10.1016/j.knosys.2008.03.045.

[18] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery. 2 (1998) 121–167.
[19] K.-W. Cheung, J.T. Kwok, M.H. Law, K.-C. Tsui, Mining customer product ratings for personalized marketing, Decision Support Systems. 35 (2003) 231–243. doi:10.1016/S0167-9236(02)00108-2.
[20] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning (ECML), Springer Berlin/Heidelberg, 1998: pp. 137–142. http://link.springer.com/chapter/10.1007/BFb0026683 (accessed November 8, 2016).
[21] R. Forsati, M. Shamsfard, Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches, Applied Mathematical Modelling. 38 (2014) 3193–3211.
[22] D.M. Christopher, R. Prabhakar, S. Hinrich, Introduction to information retrieval, 2008.
[23] D. Gouscos, M. Kalikakis, M. Legal, S. Papadopoulou, A general model of performance and quality for one-stop e-Government service offerings, Government Information Quarterly. 24 (2007) 860–885. doi:10.1016/j.giq.2006.07.016.
[24] S. Rotchanakitumnuai, Measuring e-government service value with the E-GOVSQUAL-RISK model, Business Process Management Journal. 14 (2008) 724–737. doi:10.1108/14637150810903075.
[25] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly. 27 (2003) 425–478.
[26] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly. 13 (1989) 319–340.
[27] Nitin Seth, S.G. Deshmukh, Prem Vrat, Service quality models: a review, Int J Qual & Reliability Mgmt. 22 (2005) 913–949. doi:10.1108/02656710510625211.
[28] A. O'Cathain, K.J. Thomas, Any other comments? Open questions on questionnaires–a bane or a bonus to research, BMC Med Res Methodol. 4 (2004) 25.
[29] P. Soucy, G.W. Mineau, A simple KNN algorithm for text categorization, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 2001: pp. 647–648.
[30] T. Joachims, Learning to classify text using support vector machines: Methods, theory and algorithms, Kluwer Academic Publishers Norwell, 2002.
[31] L. Shi, X. Ma, L. Xi, Q. Duan, J. Zhao, Rough set and ensemble learning based semi-supervised algorithm for text classification, Expert Systems with Applications. 38 (2011) 6300–6306.
[32] M.A. Nielsen, Neural networks and deep learning, Determination press USA, 2015.
[33] A. Trask, A Neural Network in 11 lines of Python (Part 1) - i am trask, Https://Iamtrask.Github.Io. (2015). https://iamtrask.github.io//2015/07/12/basic-python-network/ (accessed May 20, 2018).
[34] A. Trask, A Neural Network in 13 lines of Python (Part 2 - Gradient Descent) - i am trask, Https://Iamtrask.Github.Io. (2015). https://iamtrask.github.io//2015/07/27/python-network-part2/ (accessed May 31, 2018).
[35] G. Kassabgi, Text Classification using Neural Networks, Machine Learnings. (2017). https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6 (accessed March 21, 2018).
[36] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004: p. 271. http://dl.acm.org/citation.cfm?id=1218990 (accessed November 28, 2014).
[37] J. Martineau, T. Finin, Delta TFIDF: An Improved Feature Space for Sentiment Analysis., Icwsm. 9 (2009) 106.

[38] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research. 9 (2008) 1871–1874.
[39] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Association for Computational Linguistics, 2011: pp. 142–150.
[40] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based N-gram Models of Natural Language, Comput. Linguist. 18 (1992) 467–479.
[41] scikit-learn, scikit-learn: machine learning in Python, (2010). http://scikit-learn.org/stable/ (accessed April 4, 2018).
[42] S. Li, Scikit-Learn for Text Analysis of Amazon Fine Food Reviews, DataScience+. (2017). https://datascienceplus.com/scikit-learn-for-text-analysis-of-amazon-fine-food-reviews/ (accessed April 21, 2018).
[43] F. Pedregosa-Izquierdo, Feature extraction and supervised learning on fMRI: from practice to theory, PhD Thesis, Université Pierre et Marie Curie-Paris VI, 2015.
[44] J. McAuley, J. Leskovec, Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2013: pp. 165–172. doi:10.1145/2507157.2507163.
[45] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management. 45 (2009) 427–437.
[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation. 9 (1997) 1735–1780.
[47] X. Zhang, J. Zhao, Y. LeCun, Character-level Convolutional Networks for Text Classification, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA, 2015: pp. 649–657. http://dl.acm.org/citation.cfm?id=2969239.2969312 (accessed July 10, 2018).
[48] S.J. Pan, Q. Yang, A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering. 22 (2010) 1345–1359. doi:10.1109/TKDE.2009.191.