ISSN 1870-4069

Gathering, Classifying and Visualizing Results from Social Surveys using OCR and Machine Learning Techniques

David Céspedes-Hernández, Juan Manuel González-Calleros,
Josefina Guerrero-García, Liliana Rodríguez-Vizzuett

Benemérita Universidad Autónoma de Puebla, Puebla, Mexico

{dcespedesh, juan.gonzalez}@cs.buap.mx,
{jguerrero, lilianarv}@cs.buap.mx

Abstract. Statistical representation of data as graphics is useful for making its interpretation easier, and for helping on decision making processes. When it comes to information that is obtained as result of social surveys, the problem is not only on presenting it, but on gathering the results by different means, and classifying them into categories, considering that for open questions, this is a process that is commonly done in a manual way. This paper is aimed on presenting a software tool for gathering data from social surveys relying on OCR technology or through a web application, classifying it using Machine Learning techniques, and presenting the resulting information in a mostly automatic way. For illustrative purposes, a case of study of a survey that was applied in Puebla, Mexico, in 2013 is considered. Quantitative results were obtained and reported for the classification process, and general user interfaces of the developed application are presented and described.

Keywords. Machine learning, information visualization, social survey, web development, optical character recognition.

1 Introduction

When gathering data for research purposes, surveys are one of the most commonly used methods [15]. Several authors have addressed the creation and application of surveys for obtaining data about specific subjects from social, psychological and statistical points of view to mention a few. Also, it is remarkable that to a number of domains such as politics, economics and marketing, surveying is relevant [2, 10, 17].

Usually, it is not possible to survey all of the members within a population and in order to have representative results from surveys, it is necessary to carefully pick samples that reflect tendencies of how a determined population thinks or acts. In this sense, if it is of interest to understand the opinion of a complete town towards some concept, it will be necessary to determine a sample size [3], to survey the members of the sample (traditionally but not limited to paper based questionnaires), to gather the

pp. 81–95; rec. 2017-09-10; acc. 2017-10-20

results, and to obtain and present information from them through a diversity of statistical methods.

However, for some studies such as census, or surveys on social needs applied on remote communities, because of the sample size and its geographic distribution, gathering the results, organizing them, and giving an interpretation is not always a straightforward process.

In addition, to process surveys that were applied to large samples can be a time consuming, expensive task, and human errors may happen either when gathering the data, when capturing it (if needed), when categorizing the resulting information, or when generating statistics for analysis.

It is evident that multiple domains are affected by this problem, and that its solution through a software tool would help in minimizing and quantifying the errors, reducing the invested time and the cost of the whole process, and presenting statistical information in an opportune way.

The general objective of this work, is to develop an application for gathering results either from paper based surveys or from digitally stored questionnaires, classifying such results into domain specific categories, and finally giving visual representation to that information for its further analysis. As it can be inferred, this objective can be decomposed into a set of specific goals to be accomplished:

1) Support data gathering from paper based surveys.
2) Support data gathering from digitally stored questionnaires.
3) Classify the gathered data into a set of defined, specific to the domain categories.
4) Present the classified information in a graphic way for its analysis.

The remainder of this paper is structured into sections. Section two contains the state of the art regarding techniques and tools that are useful for the development of this work: Optical Character Recognition tools, Machine Learning techniques applied to text classification, and suites or Application Programming Interfaces (APIs) for information visualization. The third section is aimed on describing the development of an application for meeting the objectives that were defined in previous lines. Later, results of the implementation of the developed application are reported in Section 4.

Next, Section 5 consists on a discussion about the before presented results, comparing them to the usage of different algorithms for text classification on the proposal context. Finally, in Section 6, conclusions and future work are provided in terms of how the defined goals were accomplished and of further research to be done.

2 State of the Art

Considering the specific objectives that were described in the Introduction, it is possible to identify the requirement of exploring techniques and tools that allow to address them. This section is dedicated to describe a state of the art on such concepts. First of all, taking into account that it is relevant for this work to have a mechanism for automatically processing paper based questionnaires, Optical Character Recognition (OCR) is presented. Next, as after having a set of answers either from paper based or digital surveys it will be needed to classify such information, some Machine Learning (ML) algorithms for text classification are described. And Finally, in order to meet the requirement of presenting the resulting data in a graphic way, suites and tools for Information Visualization (IV) are studied.

2.1 Optical Character Recognition

OCR is not a new problem in the field of pattern recognition. Moreover, this has been addressed by several authors and from different perspectives for decades. The definition of the term and the first works on the matter date back to 1930's [9, 20] and specifically involving computers, to the decade of 1950's [16].

The earlier attempts of OCR algorithms reported a precision of around 60% correct recognitions, while modern proposals are accurate in a 97 to 100% [4], and what is pursued nowadays, is to improve the processing time, the availability of recognition services, and other add-ons such as variety of formats of the outcome, APIs definition, and real-time translation services.

As providing a complete overview on OCR tools and algorithms is not an objective of this paper, a summarized comparison is reported in Table 1, taking only into account minimum characteristics such as proposal name, whether if an SDK/API is provided for it or not, the format of its outcome, and if multilingual support is given.

Table 1. Summarized comparison of OCR tools.

\	**SDK/API provided?**	**Outcome format**	**Multi lingual support**
Google Vision[1]	Yes	JSON structure.	Yes
Omnipage[2]	Yes	XML and text files.	Yes
Online OCR[3]	Yes	Text and Office suite formats.	Yes

For this particular work, the proposal by Google Cloud Platform is especially interesting as it provides an API that allows its inclusion to web applications, it provides high precision detection (even for handwritten texts), it allows to perform up to 1,000 free requests monthly, and it counts on active support and documentation by Google, and by its developers' community.

2.2 Machine Learning Techniques Applied to Text Classification

By definition, Text Classification (TC) is the assignment of Boolean values to each pair (di, ci) ϵ D X C, where D represents the documents to be classified, and C is the set of classes in which the documents will be categorized [19].

[1] https://cloud.google.com/vision/

[2] https://www.nuance.com/

[3] https://www.onlineocr.net/

Since the early 60's, TC has become an activity of interest for its application in various contexts that require information retrieval such as news filtering and organization, opinion mining, and email classification (spam filtering), among others. On the first efforts about TC, a set of rules was manually defined for encoding expert knowledge on how document classification was supposed to be done. Then, those rules were used in order to classify incoming data [11]. The main disadvantage then, was the dependence towards experts for the creation of constrains, and that the accuracy was directly proportional to the extension and completeness of the set of rules.

During the decade of 1990, the Knowledge Engineering community worked on the application of ML techniques for providing a more automatic and more accurate approach, following the principle of making this process less reliant on the expert fed.

A number of ML techniques have been designed or adapted for accomplishing TC. If well most methods are created for being used on the quantitative data related to text, such approaches are usually applied depending on the characteristics of the documents to be classified, delivering different results in terms of precision, computational cost, and usage or implementation complexity.

The most commonly used proposals for TC in the literature, are as follows: Decision trees, Pattern-based classifiers, SVM classifiers, Neural network classifiers, and Bayesian classifiers [12].

In general, probabilistic classifiers, also called generative classifiers, are designed for using a mixture model for generation of the underlying documents. This mixture model typically assumes that each class is a component of the mixture. Each mixture component is essentially a generative model, which provides the probability of sampling a particular term for that component or class [1].

According to the nature of the gathered data, to be described in Section 3.1, Probabilistic Bayesian methods are considered for the elaboration of this proposal. Bayesian classifiers, are probabilistic techniques, based on modeling the underlying word features in different classes. Their operation is based on classifying text considering the posterior probability of the documents belonging to different classes, taking into account the presence of the word that was modeled in the documents [13].

2.3 Suites and APIs for Information Visualization

IV is the process of transforming abstract information into more easily understandable, graphical forms [18]. Typically, before applying statistical graphics or other representation techniques, it is necessary to preprocess data i.e. converting non-numerical data into a numerical form, and organizing or arranging information, for easier understanding and analysis.

Nowadays, along with the development of systems that allow to work with large amounts of information, and to deliver results or reports for analysis or for supporting decision making processes, the need of tools for visualizing such information is imminent [6]. With that necessity, came the proposal by the research community on the matter, of providing a diversity of tools, techniques and applications.

For this work, it is of especial interest those proposals having an application on web environments, providing the options of using different representations for numerical data, counting on support by developers' community, and allowing the creation of custom graphs. In this sense, Google Charts API4, Chart.js [5], JFreeChart [8], and InfoViz5 were considered and contrasted. It was finally decided to use InfoViz library because of the variety of charts that it provides, and that they may be useful for the purposes of this proposal, and for the future work to be described in Section 6.

3 Development of a Software Tool for Gathering, Classifying, and Visualizing Results from Social Surveys

Taking into account the objectives that were defined in the introduction and applying what was reported on the state of the art towards their accomplishment, it was decided to create a software tool for gathering results from social surveys, that allows to classify such results in a semiautomatic way, and that reports information about the done classification in a graphic manner.

It was also noticed, that due to the nature of the surveys, the gathering process should consider that data may come either from scanned paper based questionnaires, or from digital versions of such questionnaires.

In order to expose the development that was done, a case of study is considered. This experiment consisted on taking data from 25,014 answered surveys that were applied in the city of Puebla, Mexico during the year 2013. The objective of that research work was to understand the population opinion regarding the government performance, and to identify needs in terms of *security*, *health*, *employment*, *economy*, and *education*.

The inquiry form consisted on 40 questions divided on 3 blocks. The first set of questions was defined for getting relevant information about the context of the respondent. From this point, it was possible to get sociodemographic data.

The second block contained only closed questions, and was aimed on getting quantitative data representing the evaluation from the population to the government. Finally, the third block was designed for obtaining qualitative information regarding what the population identifies as their specific needs.

The methodology for this study consisted on surveying people from different locations within the city, as an attempt to have a representative sample, using a paper based version of the questionnaire as the one depicted in Figure 1.

Every day, the answered surveys were sent to a capture department in order to get the results digitalized. At the same time, data capturers had the assignment of classifying the answers from block three into the five previously mentioned categories.

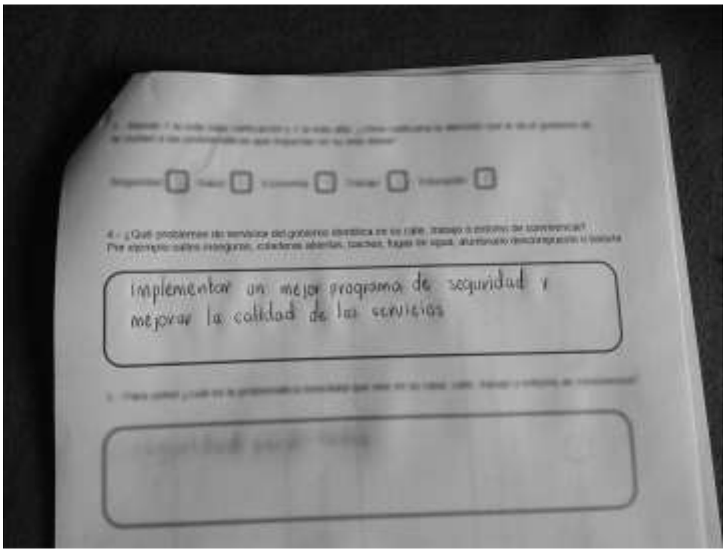
Once the determined period of time for applying the surveys was ended, and all the answers were computed, a tailored web application was implemented for counting, organizing, and presenting the resulting information in a graphic way.

From this experience, some insights were taken:

[4] https://google-developers.appspot.com/chart/

[5] http://infoviz.org/

a) It was necessary to involve a big team for surveying people, and another big team for capturing and categorizing the data, making this an expensive experiment.
b) Processing a survey, took around 25 man-minutes each (for the 25,014 questionnaires it took 10,400 man-hours).
c) During the data capture and manual classification process, human errors happened.
d) Even though the capture department completed the classification process and therefor acquired experience, if it was necessary to add more surveys to the study i.e. include more results, or to start a new study with same parameters, the time for accomplishing the tasks would not significantly decrease.



4.- ¿Qué problemas de servicios del gobierno identifica en su calle, trabajo o entorno de convivencia? Por ejemplo calles inseguras, coladeras abiertas, baches, fugas de agua, alumbrado descompuesto o basura

implementar un mejor programa de seguridad y mejorar la calidad de los servicios

Fig. 1. Extract of a paper based questionnaire used on the case of study.

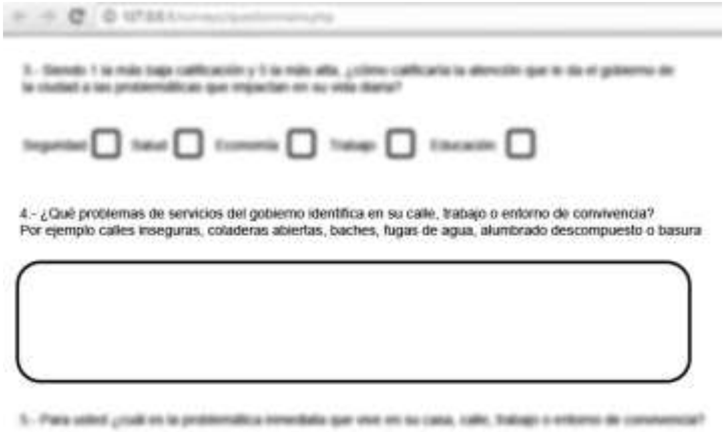
Next sections are dedicated to describe the development of a software web-based application for reaching the objectives that were defined for this work, instantiating it in terms of the introduced case of study.

3.1 Data Gathering

In the presented case of study, paper based surveys were used. This caused the necessity of capturing the answers in a separate process.

Aligned with the objectives of this work, our proposal for gathering data in a more efficient way, consists on either developing a simple web based application, and to use it for directly entering the information in an electronic format, or to scan the questionnaires and use OCR tools for its digitalization.

Regarding the provision of a system for enabling the application of digital surveys, Figure 2 shows an extract of its User Interface (UI). The use of this tailored application is pretty straightforward, as each question is listed along with a component for inputting the response, and after completing each of the questionnaire blocks, it is possible to save changes into a local database.



3.- Siendo 1 la más baja calificación y 5 la más alta, ¿cómo calificaría la atención que le da el gobierno de la ciudad a las problemáticas que impactan en su vida diaria?

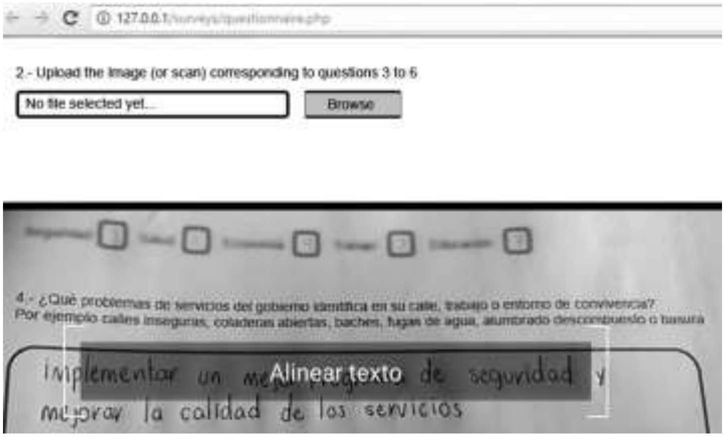
Seguridad □ Salud □ Economía □ Trabajo □ Educación □

4.- ¿Qué problemas de servicios del gobierno identifica en su calle, trabajo o entorno de convivencia?
Por ejemplo calles inseguras, coladeras abiertas, baches, fugas de agua, alumbrado descompuesto o basura

5.- Para usted ¿cuál es la problemática inmediata que vive en su casa, calle, trabajo o entorno de convivencia?

Fig. 2. User interface of the developed survey system.

An interesting feature of this software tool, is that it allows to apply the surveys off-line, and later, when Internet connection is available, to perform the information load to a data base stored in a proprietary server.



127.0.0.1/surveys/questionnaire.php

2.- Upload the Image (or scan) corresponding to questions 3 to 6

No file selected yet... | Browse

Seguridad 1 Salud 1 Economía 4 Trabajo 2 Educación 3

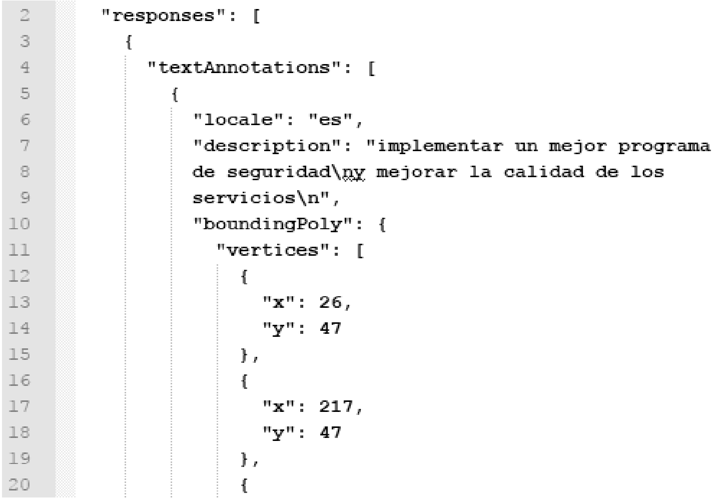
4.- ¿Qué problemas de servicios del gobierno identifica en su calle, trabajo o entorno de convivencia?
Por ejemplo calles inseguras, coladeras abiertas, baches, fugas de agua, alumbrado descompuesto o basura

Implementar un me[illegible] de seguridad y
mejorar la calidad de los servicios

Fig. 3. User interface of the developed OCR application.

The second alternative given for data gathering, consists on providing a software tool for allowing document scanning, so later, by using OCR, digitalize the responses. For this purpose, it was necessary to develop another tool, which UIs are shown in Figure 3.

The interaction with this software, consists on following the instructions in the screen, for uploading/scanning the paper based surveys in a specific order. The outcome of Google Vision OCR, presented in Figure 4, is a JSON format structure with the read text and vectors containing the position of each word on the image.



```
  "responses": [
    {
      "textAnnotations": [
        {
          "locale": "es",
          "description": "implementar un mejor programa
          de seguridad\ny mejorar la calidad de los
          servicios\n",
          "boundingPoly": {
            "vertices": [
              {
                "x": 26,
                "y": 47
              },
              {
                "x": 217,
                "y": 47
              },
              {
```

Fig. 4. JSON outcome from Google's OCR.

As in the other option for gathering information, part of the process (image capture or scanning) can be done off-line, and then, the file upload should be performed when having Internet connection.

The above described systems address the survey and data capturing processes. The next step in the methodology is then the classification of such data into the already mentioned five categories of interest. In the following section, a ML approach for that purpose is presented.

3.2 Data Classification

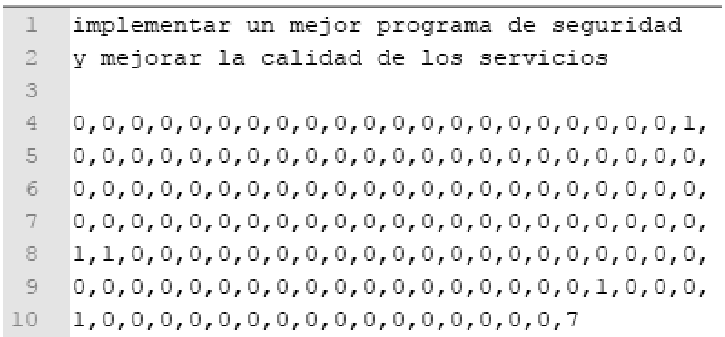
The manual classification that is described in the case of study, followed two main principles: 1) All responses must fit in a category, and 2) Every answer can only be labeled with one class, even if two or more are applicable. Following those same principles, and considering that manual classification has already been done, it was possible to instantiate a probabilistic Bayesian classifier.

Before implementing the classifier, it was necessary to preprocess information and to prepare its parameters. First of all, the structure of the information was analyzed, and it was noticed that most responses were not longer than 10 words, and that there were even some of them containing only one. Taking this into account, documents were normalized, i.e. accents and punctuation were removed, and all text was converted to lowercase.

After normalization, documents were separated according to the category they belong to, in order to calculate the frequency of meaningful words for each class. As it can be inferred, in this counting closed words were not considered.

The outcome of this operation consisted on five lists of words, one for each category, along with their appearance in the set of documents. The 50 most frequent words were selected from each one and then combined, removing repeated utterances, and resulting on a 149 words dictionary organized in an array of 149 Strings (categories).

Using that dictionary, each document was translated to numerical comma separated data, considering if whether a word from the dictionary appeared in the document or not, and how many times it was found in there, respecting the already given classification (class), as it can be seen in Figure 5.



```
implementar un mejor programa de seguridad
y mejorar la calidad de los servicios

0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,7
```

Fig. 5. Example of document translation to numerical comma separated form, using the created dictionary.

With such defined categories and classes, the Bayesian classifier implementation was possible. The execution of this algorithm is transparent to the user, but for research purposes metrics can be applied for evaluating its precision and performance. The results obtained from the training and testing of this approach are presented in Section 4.

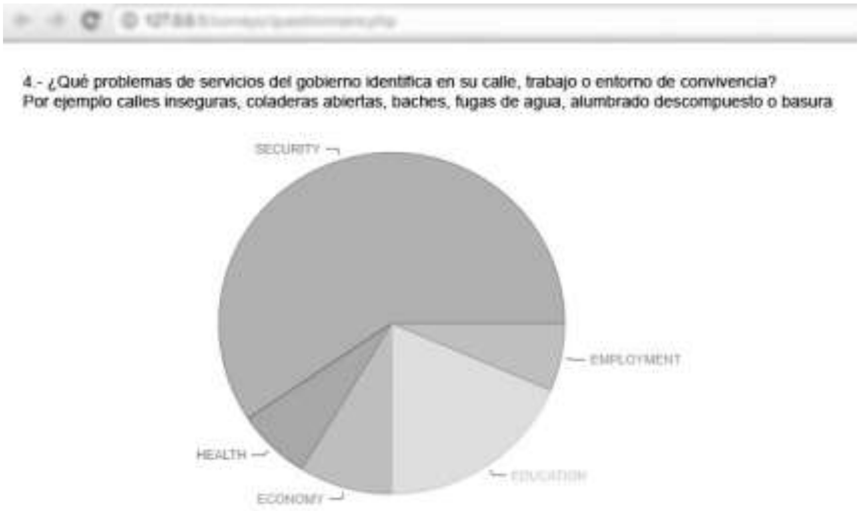
Once training of the classifier is complete, and the organization of the testing set is done, it is possible to automatically classify new coming information and to move on to the next stage of the methodology, enable visualization of the obtained information. The development of the IV module is briefly described in the next section.

3.3 Information Visualization

According to the methodology followed in the case of study, and the objectives that were established in the introduction, IV is the last phase of the process. The idea that is pursued with linking the information presentation to the data classification, is to allow users to have real-time interpretation of the gathered data.

The development of this module, consisted on creating a generic application, that given a set of classes, a chart type, and a data source, prepares a UI for displaying graphics corresponding to the fed information using InfoViz.

One of the resulting UI for the case of study is presented in Figure 6. For this specific example, the used classes are: *security*, *health*, *economy*, *education*, and *employment*; the chart type is pie chart; and the data source is the outcome of the classification process.



4.- ¿Qué problemas de servicios del gobierno identifica en su calle, trabajo o entorno de convivencia?
Por ejemplo calles inseguras, coladeras abiertas, baches, fugas de agua, alumbrado descompuesto o basura

Fig. 6. User interface of the information visualization module.

This graphic presentation, combined with the ones of other questions, supports the interpretation of the study, and a decision making process as it was the objective of the research experiment. The following section is dedicated to detail the results that were obtained during the development of this work.

4 Results

The present work had results in different terms. During its development, a web based application was created for allowing users to survey, gather data from surveys, classify information, and present it in a graphic way.

For the implementation of such software, along with web development languages and tools such as HTML, PHP, JavaScript, jQuery, MySQL, Google's OCR API, a Bayesian classifier, and InfoViz Library were used. Regarding the Bayesian classifier that was used, it was possible to gather statistic information about its performance. Table 2 contains the confusion matrix that was from it obtained, after using 65% of data for training and 35% as testing set.

From the 25,014 documents that were available, 21,293 were successfully classified, representing a total error of 14.88%. If well this number provides information regarding the behavior of the classifier, it is necessary to carry out a deeper analysis to understand its limitations.

There were 13,236 documents classified into the security class. In the implementation of the classifier, 95.58% (12,652/13,236) of such items were correctly organized. In a similar way, the precision for the other classes were 72.97% (1,461/2,002), 58.40% (1,856/3,178), 81.40% (3,961/4,866), and 78.69% (1,363/1,732).

The class that got the lowest precision is Economy, and when reviewing its wrongly classified documents, it was noticed that most of them contained the words "security", "employment", and "work", and thus, were incorrectly categorized in the Security and Employment classes.

Table 2. Confusion matrix for the Bayesian net classifier.

\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Security	12652	199	144	169	72
Health	424	1461	39	52	26
Economy	714	146	1856	403	59
Employment	494	81	280	3961	50
Education	195	55	46	73	1363

On the other hand, modifying the methodology that was expressed in the case of study, in order to use the developed software, would represent a diminution of 70% - 100% of the time employed for data capture, considering that either capture would be done while surveying or, that scanning a paper based questionnaire would take around 7 minutes.

In terms of total precision and complexity of the proposal, as Google's OCR precision and complexity order is not reported in the literature, it is only possible to have an estimation based on assumptions. Considering that in the best cases, OCR algorithms in the literature report having up to a 98% of precision in handwritten texts [22] with a time complexity order of O(n2) [21], while Bayesian net classifiers have reported an order of O(td) for training, and O(cd) for testing where t is the number of training examples, c is the number of classes, and d is the number of attributes [7, 14]. In the worst of the cases, the order for the classifier tends to O(n2), and then the total order for the process is O(2n2).

Finally, considering the error probability on the OCR process of 2%, and an error probability of 14.88% on the classification, the total error probability is given by $P_{TOTAL} = P_{OCR} \text{ U } P_{BN} = 0.02 + 0.1488 - (0.2 \text{ x } 0.1488) = 16.58\%$, where P_{OCR} is the probability of having an error on the OCR process, and P_{BN} is the probability of error on the Bayesian classifier. Next section presents a discussion over the development and the obtained results.

5 Discussion

When analyzing the results that were reported in the previous section, it was possible to highlight some features of the proposal. The first ones corresponding to the change in the work methodology that is required. As it was commented in Section three, at some point of time, the interviewer will need to have Internet access or to be provided with a device for the specific purpose of surveying, representing an additional investment to be done, and leading to ask what happens if information is lost before the interviewer is able to upload it to the data base? And, how often does the upload process needs to be performed?

The answer to these questionings is directly related to the study in which it is to be applied, depends on the number of interviewers, the amount of surveys to be applied, and the availability of Internet services for the interviewers.

On the other hand, the main advantage that this proposal has against the as-is process is on time consumption. As it was commented in the results section, for the case of study, a saving of a 70% of the invested time was possible. This means that a research study that took 10,400 man-hours in the capture process, may be reduced to 3,120.

Table 3. Confusion matrix for the Bayesian net with 10 folds cross-validation, naïve Bayes with 65% of data on training set, and naïve Bayes with 10 folds cross-validation classifiers.

<table>
<tr><th colspan="6">Bayes net with 10 folds cross-validation</th></tr>
<tr><th>\</th><th>Security</th><th>Health</th><th>Economy</th><th>Employment</th><th>Education</th></tr>
<tr><th>Security</th><td>12625</td><td>209</td><td>150</td><td>179</td><td>73</td></tr>
<tr><th>Health</th><td>430</td><td>1452</td><td>43</td><td>51</td><td>26</td></tr>
<tr><th>Economy</th><td>723</td><td>149</td><td>1847</td><td>400</td><td>59</td></tr>
<tr><th>Employment</th><td>493</td><td>88</td><td>279</td><td>3956</td><td>50</td></tr>
<tr><th>Education</th><td>199</td><td>56</td><td>50</td><td>80</td><td>1347</td></tr>
<tr><th colspan="6">Naïve Bayes with 65% of data as training set</th></tr>
<tr><th>\</th><th>Security</th><th>Health</th><th>Economy</th><th>Employment</th><th>Education</th></tr>
<tr><th>Security</th><td>12446</td><td>263</td><td>173</td><td>303</td><td>51</td></tr>
<tr><th>Health</th><td>645</td><td>1226</td><td>55</td><td>57</td><td>19</td></tr>
<tr><th>Economy</th><td>995</td><td>192</td><td>1543</td><td>394</td><td>54</td></tr>
<tr><th>Employment</th><td>591</td><td>78</td><td>218</td><td>3938</td><td>41</td></tr>
<tr><th>Education</th><td>344</td><td>43</td><td>51</td><td>83</td><td>1211</td></tr>
<tr><th colspan="6">Naïve Bayes with 10 folds cross-validation</th></tr>
<tr><th>\</th><th>Security</th><th>Health</th><th>Economy</th><th>Employment</th><th>Education</th></tr>
<tr><th>Security</th><td>12450</td><td>267</td><td>158</td><td>302</td><td>59</td></tr>
<tr><th>Health</th><td>662</td><td>1210</td><td>46</td><td>57</td><td>27</td></tr>
<tr><th>Economy</th><td>1005</td><td>194</td><td>1525</td><td>394</td><td>60</td></tr>
<tr><th>Employment</th><td>592</td><td>78</td><td>219</td><td>3934</td><td>43</td></tr>
<tr><th>Education</th><td>342</td><td>43</td><td>49</td><td>82</td><td>1216</td></tr>
</table>

In addition, considering that this specific experiment was done in a period of three weeks, by 65 capturers, using the here proposed capture application, this could be done in the same three weeks by 20 people. This means that the time reduction is translated into an economic saving in logistics and human resources.

It is important also to emphasize that the accuracy of this proposal is determined by at least two elements: the OCR precision, and the classifier training, The OCR outcome depends on the quality of the algorithm, but also on the input (handwritten text is harder to be recognized).

Besides, during the classifier training task, the manual classification may be subjective taking into account the principles that were established in Section three "all responses must fit in a category, and every answer can only be labeled with one class, even if two or more are applicable". Subjectivity in this phase may lead to incorrect training, or inaccurate results during the validation. For instance, consider the following document taken from the case of study: "*falta de trabajo causante de la inseguridad*" which translated to English, is read as "*lack of employment causes insecurity*", and may be manually classified either as an *employment*, or as a *security* matter.

In regards to the Bayesian net classifier which was used, it was possible to contrast it with a Naïve Bayes classifier, and to evaluate both alternatives using a 10 folds cross-validation. The complete results are presented in Table 3. From those results, it is possible to calculate total precision for each alternative, and moreover for each class as it was done in results section for the Bayes net classifier. The comparison of those results is provided in Table 4.

Table 4. Comparison of results for the different classifiers.

\	Security	Health	Economy	Employment	Education
Bayes net with 65% of data in training set	0.95	0.72	0.58	0.81	0.78
Bayes net with 10 folds cross-validation	0.95	0.72	0.58	0.81	0.77
Naïve Bayes with 65% of data in training set	0.94	0.61	0.48	0.80	0.69
Naïve Bayes with 10 folds cross-validation	0.94	0.60	0.47	0.80	0.70

Using that comparison, it is possible to notice a slight advantage of the Bayes net towards the other alternative, which is translated as a 0.3% benefit on the usage of this method. After reporting results and obtaining insights about them, it is possible to have conclusions from this work, and to establish new objectives. Next section is dedicated to describe those findings inspiring future work on this subject.

6 Conclusions and Future Work

In this paper, the development of a web based application for gathering, classifying, and presenting results from social surveys, using OCR, ML, and IV techniques, was addressed. When considering the specific objectives that were settled in the introduction, it is possible to notice that they were aimed and reached.

The process of gathering data from paper based, and digital questionnaires was covered by developing a tool for either surveying users, or uploading answered surveys, and to extract information from them supported with Google´s OCR. The gathered information was then classified using a Bayes net classifier, which performance was measured and contrasted in terms of precision and complexity order. Finally, the presentation of the classified information was supported with the implementation of a real-time updated module that receives information specific to the domain, and using the InfoViz library, provides graphical outcome representing the results of the elaborated surveys.

Following the idea of having control of the whole process, and giving a more straightforward execution to the user, it was identified that using an OCR API is not the ideal option. Having a tailored implementation of an OCR off-line algorithm supported by digital image processing techniques for segmentation, would allow the evaluation of it in terms of precision and complexity, leading to the correct validation of the entire proposal. Regarding the classifier, in order to increase the precision of this proposal, two lines may be followed as future work: exploring automatic unassisted ML techniques, and considering other assisted techniques, and compare their results with those reported in this paper. From the domain perspective, not all data may be susceptible of being displayed in a tabular way, but could be more illustrative when shown in a geographical manner through information layers on maps. The InfoViz library allows such sort of tasks, but a modification on the questionnaire creation, data gathering and information classification processes will be required.

References

1. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. Mining text data, Springer Science & Business Media, Springer, Heidelberg, 163–213 (2012)
2. Althaus, S.L.: Collective preferences in democratic politics: Opinion surveys and the will of the people. Cambridge University Press, New York (2003)
3. Barlett, J.E., Kotrlik, J. W., Higgins, C.C.: Organizational research: Determining appropriate sample size in survey research. Information technology, learning, and performance journal, 19, 43–50 (2001)
4. de Mello, C.A., Lins, R.D.: A comparative study on OCR tools. In: Vision Interface'99 Conference, 224–231 (1999)
5. Downie, N.: Open source charts for your website. http://www.chartjs.org
6. Fekete, J.D., Van Wijk, J.J., Stasko, J.T., North, C.: The value of information visualization. Information visualization, LNCS, 4950, Springer, Berlin, Heidelberg, 1–18 (2008). doi: 10.1007/978-3-540-70956-5_1

7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine learning, 29, Springer, 131–163 (1997). doi: 10.1023/A:1007465528199
8. Gilbert, D.: The jfreechart class library. Developer Guide, Object Refinery, 7 (2002)
9. Handel, P. W.: U.S. Patent No. 1,915,993. Washington, DC: U.S. Patent and Trademark Office (1933)
10. Ilieva, J., Baron, S., Healey, N.M.: Online surveys in marketing research: Pros and cons. International Journal of Market Research, 44, 361 (2002)
11. Joachims, T., Sebastiani, F.: Guest editors' introduction to the special issue on automated text categorization. Journal of Intelligent Information Systems, 18, 103–105 (2002)
12. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. Frontiers in artificial intelligence and applications, 160, IOS Press, Netherlands (2007)
13. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. Association for the advancement of artificial intelligence journal, 90, 223–228 (1992)
14. Martınez, A.M., Webb, G.I., Chen, S., Zaidi, N.A.: Scalable learning of Bayesian network classifiers. Journal of Machine Learning Research, 17, 1–35 (2016).
15. Mathers, N., Fox, N.J., Hunn, A.: Surveys and questionnaires. Trent RDSU, United Kingdom (2007)
16. Mori, S., Suen, C. Y., Yamamoto, K.: Historical review of OCR research and development. In: Proceedings of the IEEE, IEEE Press, 80, 1029–1058 (1992). doi: 10.1109/5.156468
17. Ravallion, M., Chen, S.: What can new survey data tell us about recent changes in distribution and poverty? The World Bank Economic Review, Oxford University Press, 11, 357–382 (1997). doi: 10.1093/wber/11.2.357
18. Risch, J.S., Rex, D.B., Dowson, S.T., Walters, T.B., May, R. A., Moon, B.D.: The STARLIGHT information visualization system. In: IEEE conference on Information Visualization, 1997, IEEE Press, 42–49 (1997)
19. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), ACM, New York, 34, 1–47 (2002)
20. Tauschek, G.: U.S. Patent No. 2,026,330. Washington, DC: U.S. Patent and Trademark Office (1935)
21. Yalniz, I.Z., Manmatha, R.: A fast alignment scheme for automatic OCR evaluation of books. In: International Conference on Document Analysis and Recognition (ICDAR), 2011, IEEE Press, 754–758 (2011)
22. Zhang, H., Liu, C.L.: A lattice-based method for keyword spotting in online Chinese handwriting. International Conference on Document Analysis and Recognition (ICDAR), 2011, IEEE Press, 1064–1068 (2011)