From Rule Based Expert System to High-Performance Data Analysis for Reduction of Non-Technical Losses on Power Grids

Juan Ignacio Guerrero, Antonio Parejo, Enrique Personal, Íñigo Monedero, Félix Biscarri, Jesús Biscarri, and Carlos León

Department of Electronic Technology
EPS
University of Seville
C/ Virgen de África, 7
41011, Seville, Spain
e-mail: juaguealo@us.es, aparejo@us.es, epersonal@us.es, imonedero@us.es, fbiscarri@us.es, jbiscarri@us.es, cleon@us.es

***Abstract*—The Non-Technical Losses represent the non-billed energy due to faults or illegal manipulations in customer facilities. The objective of the Midas project is the detection of Non-Technical Losses through the application of computational intelligence over the information stored in utility company databases. This project has several research lines, e.g., pattern recognition, expert systems, big data and High Performance Computing. This paper proposes a module which uses statistical techniques to make patterns of correct consumption. The main contribution of this module is the detection of cases, which are usually classified as consumers with Non-Technical Loss increasing the false positives and decreasing the total success rate. This module is integrated with a rule based expert system made up of other modules, such a text mining module and a data warehousing module. The correct consumption patterns (consumers without Non-Technical Losses) are generated using rules, which will be used by a rule based expert system. Two implementations are proposed. Both of them provided an Intelligent Information System to reach unapproachable goals for inspectors. Additionally, some highlighted cases of detected patterns are described.**

Keywords - non-technical losses; pattern recognition; expert system; big data analytics; high performance computing; high-performance data analysis.

I. INTRODUCTION

Information systems have provided a new advantage: the capability to store, manage and analyze great quantities of information without human supervision. This paper proposes one solution to a very difficult problem: the Non-Technical Losses (NTLs) reduction in power utility. This paper is an extended version of a conference paper [1].

NTLs represent the non-billed energy due to the abnormalities or illegal manipulations in client power facilities. The objective of this work (called Midas project) is the detection of NTLs using computational intelligence and Knowledge Based Systems (KBS) over the information stored in Endesa databases. The Endesa company is the most biggest distribution utility in Spain with more than 12 million clients. Initially, this project was tested with information about low voltage customers. The system used consumer information with monthly or bimonthly billing. Although the system can analyze large volume of data, and which poses a very high cost in time when there are more than four million consumers. Notwithstanding, this volume of information would be unfathomable to analyze for an inspector. In order to reduce this cost, a hybrid architecture based on big data and high performance computing is currently applied to create a High-Performance Data Analysis (HPDA). This architecture has been successfully applied in biomedical topics [2], text data classification [3], and other scientific datasets [4].

Moreover, Smart Grids have provided a new scope of technologies, for example, Advanced Metering Infrastructure (AMI) with smart metering, Advanced Distribution Automation (ADA), etc. There are several references about the advantages of Smart Grids, and there are a lot of initiatives related to Smart Grids in the world (e.g., [5][6][7], etc.). Additionally, several studies about the utilization of AMI in Smart Grid (e.g., [8][9][10], etc.) to improve the power quality (e.g., [11][12], etc.) and demand management (e.g., [13][14], etc.) can be found in the current state of art. These new infrastructures increase the information about consumers, taking hourly or even quarterly measurements.

This paper proposes a model which uses statistical techniques to detect correct consumption patterns. These patterns are used to generate rules, which are applied in a Rule Based Expert System (RBES). The RBES is described in [15][16] and the module of text mining is described in [17]. In this paper, an improvement on the data mining module functionality is proposed.

The proposed solution is described as follows. Section II shows a review over the state of the art in NTLs detection. Following this review, the architecture and technical characteristics of a new proposed solution is described in Section III. Section IV provides a brief description of RBES, and whose text mining module and statistical pattern generator modules are described in Section V and Section VI respectively. In Section VII, the consumption characterization of consumers without NTLs is proposed. In Section VIII, the improvements of the proposed HPDA architecture are included. Section IX shows the new parameters for consumption characterization of consumers without NTLs in the proposed HPDA architecture. In Section X, a brief description of evaluation and experimental results are presented. After that, Section XI shows a description of several highlighted cases, which traditionally are wrongly classified as an NTL. Finally, Section XII poses the conclusions and future research lines.

II. Bibliographical Review

In terms of consumption in utilities, a great spectrum of techniques can be applied; data mining, time series analysis, etc. Basically, the use of any type of statistical technique is essential to detect anomalous patterns. This idea is not new. Several researchers usually apply statistical or similar techniques to detect or analyze anomalous consumption (e.g., [18][19][20][21][22]). Some of these techniques are based on studies of the historical customer consumptions; for example, Azadeh et al. [23] made a comparison between the use of time series, neural network and ANOVA, always with reference of the consumption of the same customer. However, these techniques have several problems, the main one being that it is necessary to have large historical data about customer consumptions. Other researchers use different studies to make good patterns of consumption, which compare the consumption of a customer with others who have similar characteristics. For example, Richardson [24] compared both neural networks and statistical techniques; in the performed tests, statistical techniques are 4% more efficient than neural networks. Hand and Blunt [25] proposed the identification of some characteristics, which make the identification of consumption patterns applying statistical techniques that use them as anomalous patterns possible. Other methods propose the use of advanced techniques to make other profiles or patterns of consumption. In this sense, Nagi et al. [26] used support vector machines and [27] applied rough sets, both of them in NTLs detection.

There are other examples, e.g., Aguero [28] proposed a method to improve the efficiency of the distribution systems for reducing technical and non-technical losses. They use the utility information system, which includes computational models of feeders and advanced modeling software systems and is based on the implementation smart grid approaches. In the same way, Paruchuri and Dubey [29] proposed the use of smart metering and advanced communication protocols to detect NTLs. Iglesias [30] proposed an analysis of energy losses for activity sectors (domestic, etc.) using a load balance, by means of consumer information, the distribution transformers and several measurement points. Alves et al. [31] suggested an upgrade of the measurement equipment by means of electronic devices such as alarm systems, connection systems, remote reconnection, and protections of drivers.

Nizar et al. [32] described a series of detection rules based on feature selection and clustering techniques, using the costumer consumption history. Depuru et al. [33] proposed the use of consumption patterns, which are generated starting from the consumer history, using a Support Vector Machine (SVM) and the gathered data from smart meters; the consumers are classified according to the pattern that they propose. Ramos et al. [34] proposed the use of an Optimum Path Forest (OPF) clustering technique based on supervised learning; so, a dataset with obtained frauds of a power distribution company is used. These references are based on the use of learning techniques using available information about the consumer consumption to model a pattern for anomalous consumption.

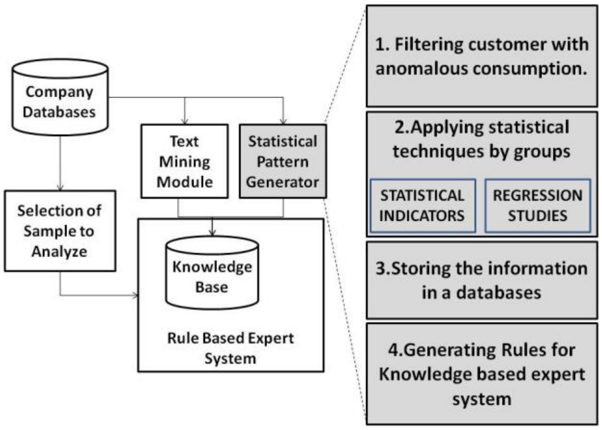
Company Databases
Text Mining Module
Statistical Pattern Generator
Selection of Sample to Analyze
Knowledge Base
Rule Based Expert System
1. Filtering customer with anomalous consumption.
2.Applying statistical techniques by groups
STATISTICAL INDICATORS
REGRESSION STUDIES
3.Storing the information in a databases
4.Generating Rules for Knowledge based expert system

Figure 1. Local Expert System Architecture and Details of Statistical Pattern Generator Module.

Other applications of advanced techniques, mainly Artificial Neural Network (ANN), which are not typically used for detection of NTLs, but could be used, are the applications for demand forecasting. In this sense, the forecasting can be done in short [35], medium [36], or long [37] term.

III. Architecture and Technical Characteristics

Initially, the architecture of original RBES is shown in Figure 1. This architecture is detailed for the Statistical Pattern Generator, showing the different stages of this process. The system was run in a single machine, and it has been successfully tested on four million clients. This volume of analysis forced the system to do partitions in order to analyze more than four million customers.

Currently, the new architecture applies big data and High Performance Computing (HPC). The big data architecture is based on Apache Spark with a database stored in HBase implemented in Apache Hadoop. The analytics are implemented in MLlib, GraphX, and library to send jobs to Graphics Processor Units (GPUs, based on Compute Unified Device Architecture or CUDA® cores). The architecture is shown in Figure 2.

The proposed system has still not been deployed over a real cluster of machines. However, a prototype was successfully implemented over a simple cluster of two nodes. The first node had an Intel® Core^{TM} i7 (3GHz), 16GB RAM and an Nvidia® GeForce® GTX750 graphic adapter (2GB and

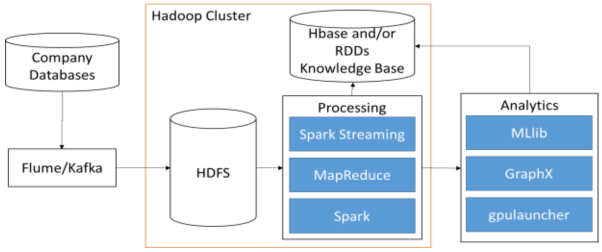

Figure 2. Architecture of Expert System in a High-Performance Data Analysis.

640 CUDA® cores). The second node had an Intel® Xeon® E5 (2GHz), 64GB RAM and an Nvidia® Quadro® K1200 graphic adapter (4GB and 512 CUDA® cores).

IV. RULE BASED EXPERT SYSTEM (RBES)

This RBES is described in [38]. This system uses the information extracted from the Endesa staff. The RBES has several additional modules, which provide dynamic knowledge using rules. The expert system has additional modules, which use different techniques: data warehousing (it is used as a preprocessing step), text mining, and statistical techniques.

The RBES can be used as additional methods to analyze the rest of information about the customer. The company databases store a lot of information, including: contract, customers' facilities, inspectors' commentaries, customers, etc. All of them are analyzed by RBES using the rules extracted from the Endesa staff and others obtained from the statistical techniques and text mining modules.

The system can be used alone or with other modules to provide additional methods to analyze the information. These modules are described in the following Sections.

V. TEXT MINING MODULE

The text mining module, which is described in [17], uses Natural Language Processing (NLP) and neural networks. This method is used to provide a tool to analyze the inspectors' commentaries. When an inspection is made in a customer's location, the inspector should register their observations and commentaries. This data is stored in company databases.

This information is not commonly analyzed because the traditional models are only based on consumption studies. The text mining module complete these studies using additional information because the inspectors' commentaries provide real information about the client facilities, which may be different from the stored in database.

This technique uses NLP and fuzzy algorithms to extract concepts from inspectors' commentaries. This process is implemented and performed in the SPSS Modeler. The NLP process consists of four engines (as described in Figure 3):

- String matching engine. This engine is based on fuzzy logic with synonym dictionaries. A fuzzy ratio is added to each word to identify similar words and mistakes. The mistake correction can be applied according to the length of words.
- Syntactic engine. This engine assigns a function to each word, according to its position and the previous and following words.
- Concept engine. This engine generates several concepts, the words with the same syntactic function and meaning in the same sentence are grouped in the same concept.
- Semantic engine. This engine assigns a semantic function to each concept. The different semantic categories are defined according to the inspectors' knowledge.

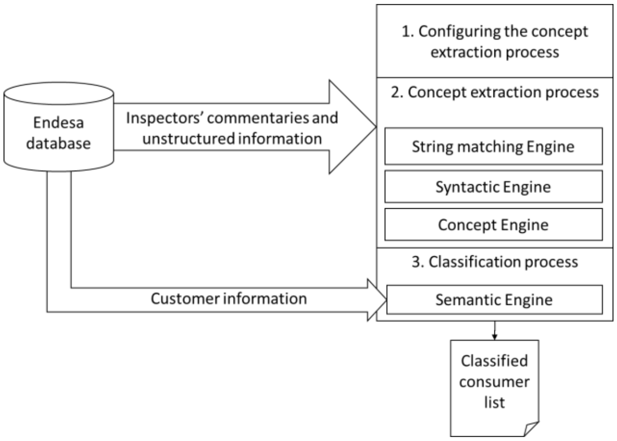
1. Configuring the concept extraction process
2. Concept extraction process
String matching Engine
Syntactic Engine
Concept Engine
3. Classification process
Semantic Engine
Endesa database
Inspectors' commentaries and unstructured information
Customer information
Classified consumer list

Figure 3. Diagram flow of Text Mining Module.

The different dialects were extracted by means of synonym dictionaries, these dictionaries allowed this module to include the different dialects in the extraction process. The extraction process involves the application of fuzzy techniques for string matching and correction of mistakes.

This set of categories was used in the second step as a training set:

- CLOSED. This category groups the concepts that represent consumers who are: closed, uninhabited, on holiday, demolished, and so on. These scenarios are usually confused with NTL by the detection of an algorithm because of the consumption pattern.
- CORRECT. This category identifies consumer installations that are correct or without NTLs. This category could prevent false positives.
- INCORRECT. This category identifies consumers whose consumer installation might have an NTL. This category represents concepts that usually identify a measurement problem in the consumer facilities.
- LOW CONSUMPTION. This category identifies consumers who usually have a low or very low consumption, due to their activity. The consumers classified in this category are filtered because the correct consumption is irregular or very low. For example, some consumers with agricultural activities

have water pumps, which have irregular and low consumption.

- UNUSEFUL. This category has 101 subcategories, which are not used in the filtering process. These subcategories include the UNKNOWN category, which contains the concepts that could not be classified. Additionally, there are several subcategories that contain information about names, numbers (currency, telephone numbers, address, etc.) and dates. These three subcategories represent 23% of the total number of concepts. This category is excluded from the set of the most frequent concepts.

These concepts are classified initially according to their frequency of appearance. The most frequent concepts are classified manually according to their meaning. Additionally, consumption indicators, date of commentary, number of measurements (estimated and real), number of proceedings, source of commentary, frequency of appearance, time discrimination band and some others are associated to each concept. Some of these indicators are generated by a Statistical Pattern Generator and applied in a Semantic Engine (Figures 4 and 5) This data is used in an ANN, which is trained with data of the most frequent concepts and is tested with the less frequent concepts. This ANN can be used to classify the new concepts, which could appear.

Additionally, the Statistical Pattern Generator is applied to whole samples when the system is in modelling time (Figure 4). Initially, this modelling process was applied to all the consumers in the Endesa database (around 12 million clients). In the analysis time, only limited size samples are usually analyzed. Thus, the Statistical Pattern Generator (Figure 5) is applied to generate some indicators, but is only applied in order to classify the consumer in a previously defined category. This classification can affect the final classification of consumer.

VI. STATISTICAL PATTERN GENERATOR

The statistical pattern generator is based on basic statistical indicators such as: maximum, minimum, average and standard deviation. These indicators are used as patterns to detect correct consumption. Additionally, the slope of regression line is used to detect the regular consumption trend. Each of these calculations is done for different sets of characteristics. These characteristics are: time, contracted power, measurement frequency, geographic location, postal code, economic activity and time discrimination band. Using these characteristics it is possible to determine the patterns of correct consumption of a customer with a certain contracted power, geographic location and economic activity.

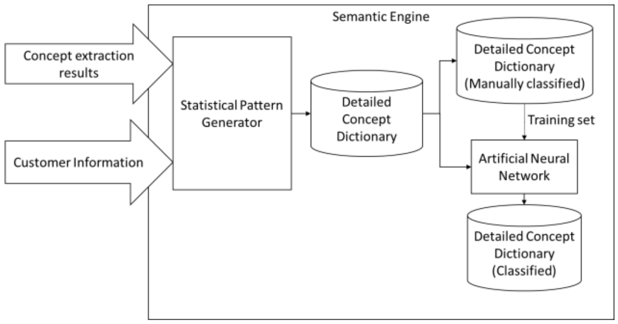
Semantic Engine
Concept extraction results
Customer Information
Statistical Pattern Generator
Detailed Concept Dictionary
Detailed Concept Dictionary (Manually classified)
Training set
Artificial Neural Network
Detailed Concept Dictionary (Classified)

Figure 4. Role of Statistical Pattern Generator in Semantic Engine in modelling time

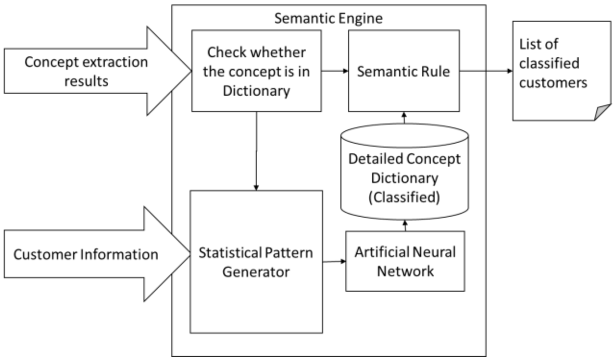
Semantic Engine
Concept extraction results
Check whether the concept is in Dictionary
Semantic Rule
List of classified customers
Detailed Concept Dictionary (Classified)
Customer Information
Statistical Pattern Generator
Artificial Neural Network

Figure 5. Role of Statistical Pattern Generator in Semantic Engine in analysis time.

The creation of these patterns needs to study a lot of customers. In this study, all customers are not used because the anomalous consumption of the customers with an NTL is filtered. This idea allows the system to eliminate the anomalous consumption getting better results.

Several tables of data are generated as a result of this study. This data is used to create rules, which implement the detected patterns. If a customer carries out the pattern, this means that the customer is correct. But if a customer does not carry out the pattern, this does not mean that the customer is not correct. These patterns are described in the next Section.

VII. CONSUMPTION CHARACTERIZATION

To find out what characteristics have more influence on consumption is a very difficult task because there is a lot of consumption information available. An in-depth analysis shows that some characteristics have more influence over the consumption: time, geographic location, postal code, contracted power, measurement frequency, economic activity and time discrimination band. The importance of these characteristics has also been analyzed in other utilities such as gas utility. Moreover, the results of these analysis have been compared with the knowledge provided by Endesa inspectors.

Each characteristic by itself is not efficient because the consumption depends on several characteristics at the same time. Thus, grouping characteristics can help find patterns of correct consumption, because these characteristics can determine the consumption with a low level of error rate providing, at least, one consumption pattern. These groups have a series of characteristics in common: geographic location, time, contracted power, and measurement frequency. These are named Basis Group because these are the main characteristics. The values for each of these

TABLE I. GROUPS OF CONSUMPTION CHARACTERISTICS

Consumption Characteristics	Description
Basis Group	This group provides consumption patterns by general geographic location: north, south, islands, etc.
Basis Group and Postal Code	This group provides patterns useful for cities with coastal and interior zones.
Basis Group and Economic activity.	The granularity of geographic location is decreased. In this way, the economic activity takes more importance. Nevertheless, the geographic location cannot be despised because, as for example, a bar has not the same consumption whether it is in interior location or coastline location.
Basis Group and time discrimination band.	There are several time discrimination bands. Each band registers the consumption at a different time range. This group provides consumption patterns in different time discrimination bands. These are useful because there exists customers who make their consumption in day or night time.

characteristics are wide; therefore, each of them shows great variations of consumption. A description of Basis Group and the other groups are shown in Table I.

Some characteristics have different granularity because they have continuous values or have a lot of possible values. The granularity is used because there are some problems related to the measurements. For example, the proposed framework performs a discretization of contracted power in 40 ranges. In the graph of Figure 6, the 14th range of contracted power is shown. This range groups the contracted power between 46,852 kW and 55,924 kW in the North of Spain. This Figure shows an abnormal level of consumption at 2002; this fact represents errors in measurements, which cannot be filtered.

In the graph of Figure 7, the average consumption in monthly periods for the 14th range is shown. In this case, the granularity of time is increased; therefore, it is possible to get another pattern, which is better than the one obtained from the graph of Figure 6. In this case, the consumption can be analyzed monthly.

Thus, several time ranges are used: absolute, monthly, yearly and seasonally. For example, the average consumption calculation provides different results: total average consumption (absolute), twelve/six average, monthly/bimonthly consumption, one average yearly consumption (when the measurements are available), and four average seasonal consumption. In the same way, the contracted power should be discretized in equal consumption ranges. In lower contracted power, the ranges are very narrow because there are a lot of consumers. When the contracted power is higher, the quantity of consumers is smaller, although the consumptions are very different. The reason for adding or aggregating the consumption (of supplies without NTLs) in different groups is because there are scenarios in which it is necessary to have other patterns.

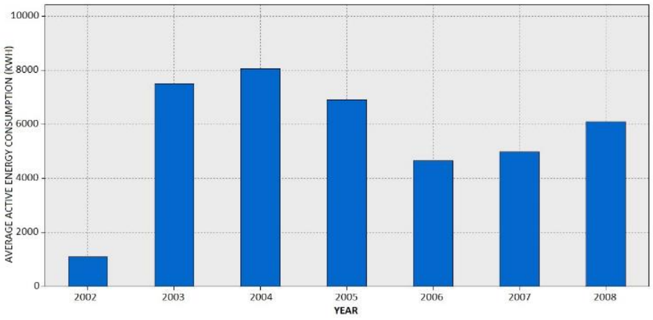

Figure 6. Average yearly consumption graph in power range 14th.

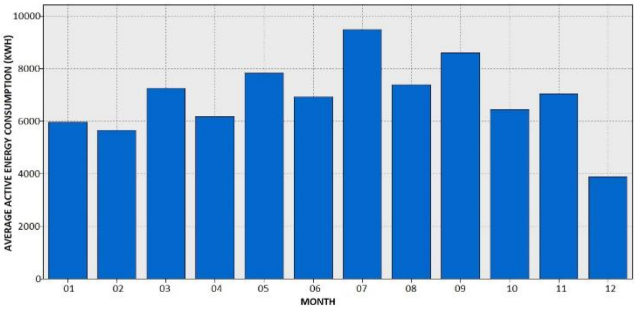

Figure 7. Average monthly consumption graph in power ranges 14th.

These groups provided dynamic patterns, which can be updated according to the time granularity. Once the characteristics are identified, it is necessary to design a process, which finds patterns automatically. Initially, these studies were made bimonthly and were applied as a part of an integrated expert system to model correct consumption patterns (Figure 1). Currently, the process can be performed hourly through the architecture proposed in Figure 2. The system applies statistical techniques to get consumption patterns using the process detailed in Figure 1.

When the rules are created, they are used to analyze the customers in order to determine if there exist any NTLs. There are defined series of rules in RBES, which use the information generated by the proposed module. The antecedents of the rules are generated dynamically using the patterns generated in the described process and according to the characteristics of the customer who will be analyzed. In this way, the use of memory resources is minimized because only the necessary antecedents of the rules are generated.

When the consumption of a customer is analyzed, several rules can fit with the characteristics of that customer. Initially, the rules are applied in the most restrictive way; this means, the customer consumption will be correct if it fits in any correct consumption pattern. Moreover, the system notifies us if the pattern fails for each customer. For example, the correct consumption ranges of active energy for specific geographic zones, different contracted power ranges, and different measurement frequency (monthly or bimonthly) are shown in Figure 8. The proposed study includes a geographic study of consumption could replace the studies related to

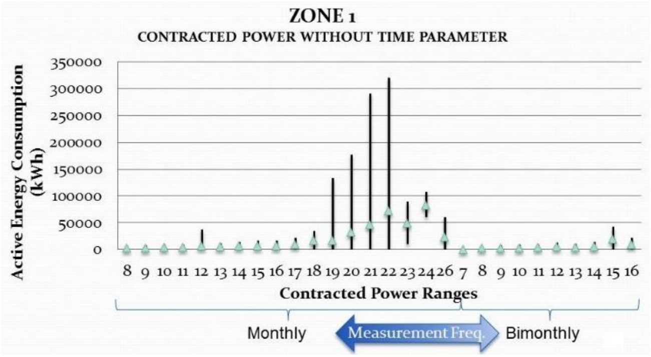

Figure 8. Graph of Active Energy Consumption Ranges vs. contracted power ranges without time parameter for a specific geographic zone.

weather conditions, because the consumers in the same zone have probably the same weather conditions. But, in this case, the contribution of adding information about weather probably does not explain the quantity of efforts and time consumption in analyzing it in front of the increment of success rate.

VIII. High Performance Data Analysis

The solution proposed in Figure 2 is a novel architecture. This solution is based on big data and HPC. The emergent technologies related to smart grids provide a great quantity of data with better temporal resolution. This solution is proposed to analyze a great quantity of information in an NRT period.

The main problem in the application of the proposed solution is the unavailability of data sources. Traditionally, this type of data sources has a very high security level and are used by other systems, which requires all resources. In these cases, it is very difficult to apply the solution, because it is necessary to load the data onto the proposed solution. Flume and Kafka are used to load the data from relational and non-relational databases. This load processes in the company databases are usually performed during the night. However, if the load of data were performed when a change in database is made, the system could operate in NRT. But, when the data is loaded in the proposed solution the operation is in real-time.

The architecture is based on Apache Hadoop and Spark, enhanced with a new daemon to take advantage of HPC architectures. This daemon, named gpulauncher, is invoked by processes in order to send jobs to GPUs. The processes can be part of a MapReduce or Analytics. Although the system implemented statistical and regression algorithms, the new algorithms will also apply multivariable inferences.

The gpulauncher daemon implemented several algorithms for analytics, functions for streaming the data to GPUs and functions for synchronization of nodes. These synchronization functions are in development. The main objective is the synchronization between GPUs of different nodes and working in near-real-time (NRT).

The loaded data is stored into HBase. The loading process includes several preprocessing steps in order to guarantee the data integrity, coherence, and anonymization. There are several processes to convert this data into a Resilient Distributed Dataset (RDD). In these processes, the data is preprocessed and transformed to the format, which is directly used by analytics algorithms.

The methods, techniques, and algorithms applied on the proposed solution (Figure 1) to fraud detection in monthly and bimonthly billing periods were adapted to the new smart grid scenario (Figure 2).

IX. Consumption characterization in HPDA solution

The consumption characterization in the proposed HPDA solution is increased with more temporal resolution in the data, adding additional time dimensions: hourly, daily, and weekly. A calendar is included and updated with working days and holidays in each geographic zone.

In the previous described solution, the generation of each pattern is updated monthly or bimonthly (according to the billing period). The solution based on HPDA makes it possible to update the pattern hourly. This scenario provides more accurate patterns. It is possible to establish the daily consumption pattern according to the type of day (working days or holiday) and the hours where the consumption is centered.

The solution based on HPDA increased the number of rules, which can be applied on a consumer. However, the criteria for applying the pattern is the same (like previously described solution). The consumer is initially selected as correct if the consumption carries out with any pattern. If the consumer does not fit any pattern if the consumer is selected as a possible NTL.

X. Evaluation and Experimental Results

The proposed module provides patterns of correct customer consumption. The analysis made by the mentioned expert system uses this module to create rules. The customer consumption analysis applies these rules according to the contract attributes: contracted power, economic activity, geographic location, postal code, and time discrimination band. Traditionally, the systems used to detect frauds or abnormalities in utilities make patterns for NTLs detection. However, in the proposed system, models of correct consumption ranges and trends are made. The use of these patterns increase the efficiency of the RBES. The Statistical Pattern Generator module is essential to analyze the customers. The RBES has been applied in real cases getting better results in zones with a lot of clients. The success of the RBES (Figure 1) is between 16,67% and 40,66% according to the quantity of clients and the quality of data, this success rate is related only to NTL detection. This fact is shown in Figure 9. But, the proposed solution classifies customers in NTL or CORRECT (without NTLs). Therefore, the success rate will be the sum of both cases.

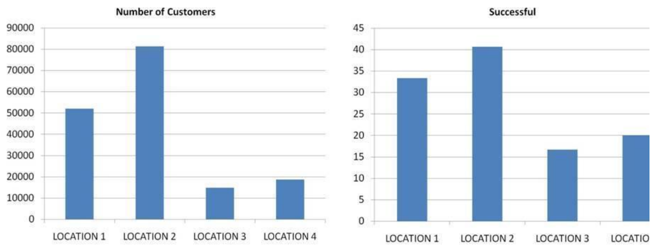

Figure 9. Number of customers vs. successful.

The evaluation of these processes is traditionally done using methods based on sensitivity and specificity measurements of the performance [39] or confusion matrix: true positive, true negative, false positive, and false negative. True and False are related to predicted and non-predicted values. Positive and Negative are related to correct and incorrect classifications. In this case, the correct classification is when the system successfully classifies the consumers (with and without NTLs). In this case, the consumers classified without NTLs are not usually inspected, and it is very difficult to get an exact value because it is not possible to know the results for all classified customers. Thus, this success rate is estimated because the consumers classified as CORRECT (without NTLs) are not inspected. Moreover, the authors check the information of results of other types of inspections made after the study (until three months) in order to verify the conclusion provided by RBES. In this case, the estimation provided a success rate between 82.3% and 91.2%, according to the quantity of clients of the corresponding location and quality of data.

The RBES was designed to deal with consumers with monthly or bimonthly billing periods. However, the new architecture based on HPDA makes the application of the expert system in NRT possible with hourly measurements. Thus, this system will be useful in the new Smart Grid infrastructures, based on AMI.

The solution proposed in Figure 1, has been successfully applied, as can be seen in previous published results [17] and [38]. The solution proposed in Figure 2, has not been tested with real inspections, it has been tested with real data provided by a retailer company but without inspections. The application of the proposed framework over the data provided several patterns for a determinate geographic location.

XI. Highlight Cases

The proposed framework in Figure 1 has been more efficient in analysis. There are some cases, which traditionally were very difficult to detect. Specifically, two cases are treated in this Section.

The first case is a client with an irrigation activity. The consumption of this type of client is strongly influenced by climate. The consumption of this client is very irregular, and difficult to analyze. These clients decrease their consumption when rainfalls increase. In this system, data about climate are not available, and only use the information about the client. Sometimes, variations of climate conditions make the data mining or regression analysis techniques select the client with irrigation activity. This client is analyzed by expert system, and normally it is dismissed according to the elapsed time since the last inspection.

The second case is a client with seasonal consumption. This type of client is very difficult to detect with traditional methods. The consumption of these clients shows one or two great peaks, which can be classified as a fraud. This type of clients can be hotels on the coast weather, which only has consumption on months with good climate or on holiday periods. The use of descriptive data mining and expert system allows the system to detect these cases.

The framework proposed in Figure 2 is useful in this case. Moreover, it is faster. However, this framework provided the identification or classification of another very difficult case: domestic clients. This case showed very high variability in the consumption, because it depended on a lot of factors: number of residents, age of each of them, housing area, etc. This information is usually unavailable for retailer and distribution companies. However, this framework has been applied over a sample of 20677 customers with a smart meter data. Thus, several patterns are identified in the groups previously described:

- High domestic consumption: These clients are characterized by periodic consumption. Inside this category there are several subcategories:
  - Full working day (only for working days). This type of clients only has high consumption between two high consumption peaks and the other periods are low consumption (Figure 10).
  - Part-time day (only for working days). The clients present low consumption several hours, between 4 to 10 hours. In these cases, there are three types centered in the morning (Figure 11), in the evening, or at night.
  - Holiday with consumption. This pattern usually models a day or a week consumption. This pattern is characterized by an irregular consumption, centering at lunch and dinner time (Figure 12).

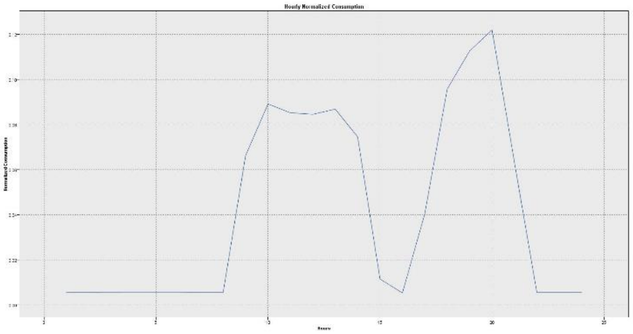

Figure 10. Graph of consumer who shows Full work day (only for working days) pattern.

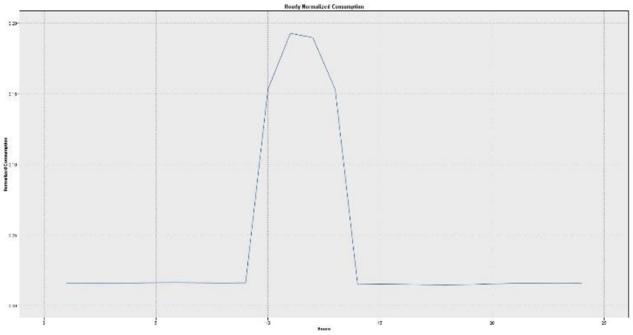

Figure 11. Graph of consumer who shows Part-time day (only for working days) centered in morning pattern.

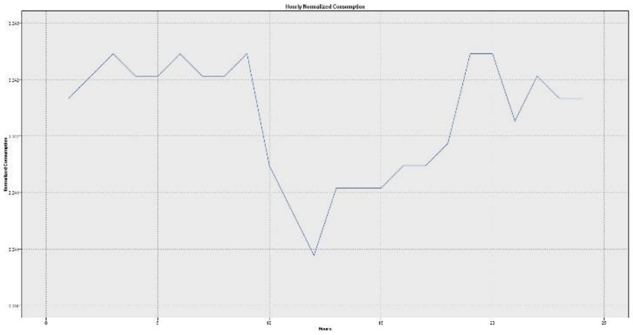

Figure 12. Graph of consumer who shows Holiday with consumption daily pattern.

- Holiday without consumption. This pattern usually models a day or a week consumption. This pattern is characterized by a constant consumption.
- Weekend with consumption. This pattern is equivalent to holiday with consumption, but in weekend.
- Weekend without consumption. This pattern is equivalent to holiday without consumption, but in weekend.

- Low domestic consumption. These clients are characterized by irregular consumption. Although, it is possible to identify the same subcategories like in high domestic consumption. In this case, two new subcategories could be described:
  - Irregular. This type of clients shows a very irregular consumption. Commonly, this pattern is overlap the other patterns. In this case, the analysis of consumption should be made for all time resolutions (weekly, monthly, and bimonthly).
  - Periodic. This type of clients shows periodic and low consumption.

Additionally, these patterns are mixed with some other patterns, which could make the interpretation or application of patterns very difficult. One of these patterns is the night shift consumption. This type of pattern modifies the pattern, increasing the consumption at night time (Figure 13). Usually, this type of clients has contracted a time discrimination band.

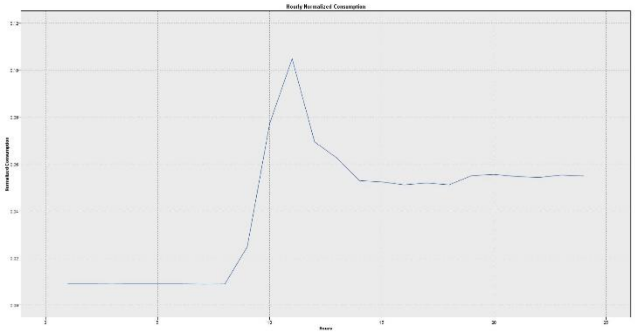

Figure 13. Graph of consumer who shows Part-time day (only for working days) with night shift.

The domestic consumers usually show a weekly pattern consumption. Usually, the consumption in domestic clients has three possible patterns: working days (Monday to Friday), weekend (Saturday and Sunday), and holidays. However, these different patterns fit in the previously proposed patterns. In this way, one consumer could carry out several patterns. Additionally, there are several problems that make it very useful the proposed solution:

- The domestic consumers are usually the largest sector in company databases.
- The domestic consumers are usually measured remotely (in Spain). The new smart meters, which support standards related to Meters&More, IEC 60870-5, IEC 61850, etc. are being installed in the power grid. This new technology has increased the quantity of information available about consumption.
- The consumption of domestic consumers has a very high variability, because it depends on several factors, which are not registered by distribution or retailer companies.
- The consumption of domestic consumers is very low if it is compared to other applications like service or industrial consumers.

XII. Conclusion and Future Work

The main contribution of the present paper is the definition of two frameworks for NTL reduction. Both frameworks are able to analyze great quantities of consumers. The first framework is applied in traditional power distribution grid with monthly and bimonthly measurements. The second framework is applied in Smart Grid scenario, with smart meter deployment. Both frameworks make detecting anomalous consumption and identifying consumers without NTLs possible.

The proposed framework in Figure 1 was implemented, tested and deployed in a real Power Distribution Company. This framework is part of a RBES. This model establishes a series of similarities with other utilities. For example, the utilization of frequency billing, geographic location, and time can be made in all utilities. However, the contracted power can be replaced by the contracted volume of flow in gas or water utilities.

The different modules of proposed solution, for example Statistical Pattern Generator or Text Mining module, can be added to other systems of NTLs detection to increase their efficiency by using rules or a translator of the knowledge generated by the module.

The solution based on HPDA was tested with real information from a Retailer Company with hourly consumption data. The test of this solution was focused in Statistical Pattern Generator module, which provides several consumption patterns, discovering some shifting of these patterns in clients. This module has not been tested for NTL detection based on inspections, but it was tested with results of other studies.

Usually, an inspector takes between 5 to 30 minutes to analyze the information about a customer in order to confirm whether an NTL exists. This period depends on the quantity of information to be analyzed; the average time of the analysis process takes 16.3 minutes. This means that the time to analyze four million customers (the maximum number of customers in case proposed in Figure 1) would be 1086666.6 hours of work. In the first case, the proposed system in Figure 1 takes 22 milliseconds per customer in the analysis process (24.4 hours in total). The HPDA provides the possibility of analyzing the information in NRT, without a limit in the number of customers. Notwithstanding, the analysis of the inspector will always be better than the machine analysis because inspectors usually work in the same zone and they have additional knowledge of facilities, which is not stored in the system. However, the analysis of the previously mentioned quantity of consumers is an unapproachable goal for inspectors. Additionally, the new AMI technologies provide information, which improves the efficiency of proposed methods.

The proposed RBES provides a double success rate because it is able to classify customers with and without NTLs. Therefore, the success rate is more difficult to evaluate and compare with because the traditional references deal with NTL detection. They do not deal with detection of customers without NTLs. In addition, there are some other problems:

- The total success rate is estimated according to the inspection results and the manual review of all customers classified without NTLs. Thus, the number of customers with informed inspections from other studies is very limited.
- This manual review takes a lot of time. When the proposed system analyzes great quantities of customers, the number of customers without NTLs is greater than the number of customers with NTL. The manual analysis of all cases classified without NTLs has a very big time period.
- Each zone has different cultural environment. This environment reduces the NTL in each of these zones and traditionally they have a low NTL level. Therefore, the number of customers and the success rate is low in these zones, due to the Statistical Pattern Generator establishes a very general pattern, which could provoke the increase of true negatives easily. Thus, the total success rate (consumers classified correctly with and without NTL) was estimated between 82.3% and 91.2% according to the location of sample and the quality of data. Notwithstanding, the success rate (only consumers with NTL) is not estimated, and it is calculated according to the results of study between 16.67% and 40.66%. Therefore, in case of the success rate of 82.3% (with and without NTLs), the success rate for customers with NTL is 16.67%. Therefore, in these cases, the proposed solution has more success rate identifying customers without NTLs.

Finally, several research lines for improving the efficiency of the proposed framework will be addressed:

- Application of techniques related to Information Retrieval, to increase the information about consumers.
- Test the new approach in a big scenario, based on AMI and with hourly measurements.
- Application of the proposed framework in other utilities.
- Enhance the analysis with application of multivariable inference.

Acknowledgment

The authors would also like to thank the backing of SIIAM project (Reference Number: TEC2013-40767-R), which is funded by the Ministry of Economy and Competitiveness of Spain.

References

[1] J. I. Guerrero, A. Parejo, E. Personal, F. Biscarri, J. Biscarri, and C. Leon, “Intelligent Information System as a Tool to Reach Unaproachable Goals for Inspectors - High-Performance Data Analysis for Reduction of Non-Technical Losses on Smart Grids,” *INTELLI 2016, The Fifth International Conference on Intelligent Systems and Applications*, 2016, pp. 83–87.

[2] E. Elsebakhi et al., “Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms,” *J. Comput. Sci.*, vol. 11, pp. 69–81, Nov. 2015.

[3] A. Rauber, P. Tomsich, and D. Merkl, “parSOM: a parallel implementation of the self-organizing map exploiting cache effects: making the SOM fit for interactive high-performance data analysis,” *IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 2000, 2000, vol. 6, pp. 177–182 vol.6.

[4] J. Liu and Y. Chen, “Improving Data Analysis Performance for High-Performance Computing with Integrating Statistical Metadata in Scientific Datasets,” *High Performance Computing*, Networking, Storage and Analysis (SCC), 2012 SC Companion:, 2012, pp. 1292–1295.

[5] V. Giordano et al., “Smart Grid projects in Europe: Lessons learned and current developments,” *European Commission*, Luxembourg, EUR 25815 EN, 2013.

[6] “The Global Smart Grid Federation Report,” 2012.

[7] P. Lewis, “Smart Grid 2013 Global Impact Report,” Ventix, VaasaETT Global Energy Think Tank, 2013.

[8] M. P. McHenry, “Technical and governance considerations for advanced metering infrastructure/smart meters: Technology, security, uncertainty, costs, benefits, and risks,” *Energy Policy*, vol. 59, pp. 834–842, Aug. 2013.

[9] C. Selvam, K. Srinivas, G. S. Ayyappan, and M. Venkatachala Sarma, “Advanced metering infrastructure for smart grid applications,” 2012 *International Conference on Recent Trends In Information Technology (ICRTIT)*, 2012, pp. 145–150.

[10] L. Dan and H. Bo, “Advanced metering standard infrastructure for smart grid,” *2012 China International Conference on Electricity Distribution (CICED)*, 2012, pp. 1–4.

[11] M. Naglic and A. Souvent, “Concept of SmartHome and SmartGrids integration,” *2013 4th International Youth Conference on Energy (IYCE)*, 2013, pp. 1–5.

[12] Z. Luhua, Y. Zhonglin, W. Sitong, Y. Ruiming, Z. Hui, and Y. Qingduo, “Effects of Advanced Metering Infrastructure (AMI) on relations of Power Supply and Application in smart grid,” *2010 China International Conference on Electricity Distribution (CICED)*, 2010, pp. 1–5.

[13] P. Siano, “Demand response and smart grids—A survey,” *Renew. Sustain. Energy Rev.*, vol. 30, pp. 461–478, Feb. 2014.

[14] E. Valigi and E. Di Marino, “Networks optimization with advanced meter infrastructure and smart meters,” *20th International Conference and Exhibition on Electricity Distribution - Part 1*, 2009. CIRED 2009, 2009, pp. 1–4.

[15] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millán, “Integrated expert system applied to the analysis of non-technical losses in power utilities,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10274–10285, Agosto 2011.

[16] J. I. G. Alonso, C. L. de Mora, F. B. Triviño, I. M. Goicoechea, J. B. Triviño, and R. Millán, “EIS for Consumers Classification and Support Decision Making in a Power Utility Database,” *Enterp. Inf. Syst. Implement. IT Infrastruct. Chall. Issues Chall. Issues*, p. 103, 2010.

[17] J. I. Guerrero, C. León, F. Biscarri, I. Monedero, J. Biscarri, and R. Millán, “Increasing the efficiency in Non-Technical Losses detection in utility companies,” *Melecon 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*, 2010, pp. 136–141.

[18] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, “Using regression analysis to identify patterns of non-technical losses on power utilities,” *Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2010, pp. 410–419.

[19] C. C. . Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcão, “A New Approach for Nontechnical Losses Detection Based on Optimum-Path Forest,” *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.

[20] M. E. de Oliveira, D. F. . Boson, and A. Padilha-Feltrin, “A statistical analysis of loss factor to determine the energy losses,” *Transmission and Distribution Conference and Exposition: Latin America*, 2008 IEEE/PES, 2008, pp. 1–6.

[21] M. Gemignani, C. Tahan, C. Oliveira, and F. Zamora, “Commercial losses estimations through consumers’ behavior analysis,” *20th International Conference and Exhibition on Electricity Distribution - Part 1*, 2009. CIRED 2009, 2009, pp. 1–4.

[22] A. H. Nizar and Z. Y. Dong, “Identification and detection of electricity customer behaviour irregularities,” *Power Systems Conference and Exposition*, 2009. PSCE ’09. IEEE/PES, 2009, pp. 1–10.

[23] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, “Forecasting electrical consumption by integration of Neural Network, time series and ANOVA,” *Appl. Math. Comput.*, vol. 186, no. 2, pp. 1753–1761, Mar. 2007.

[24] R. Richardson, “Neural networks compared to statistical techniques,” *Computational Intelligence for Financial Engineering (CIFEr)*, 1997. Proceedings of the IEEE/IAFE 1997, 1997, pp. 89–95.

[25] D. J. Hand and G. Blunt, “Prospecting for gems in credit card data,” *IMA J. Manag. Math.*, vol. 12, no. 2, pp. 173–200, Oct. 2001.

[26] J. Nagi, A. M. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, “Non-Technical Loss analysis for detection of electricity theft using support vector machines,” *Power and Energy Conference*, 2008. PECon 2008. IEEE 2nd International, 2008, pp. 907–912.

[27] J. E. Cabral and E. M. Gontijo, “Fraud detection in electrical energy consumers using rough sets,” *2004 IEEE International Conference on Systems, Man and Cybernetics*, 2004, vol. 4, pp. 3625–3629 vol.4.

[28] J. R. Aguero, “Improving the efficiency of power distribution systems through technical and non-technical losses reduction,” *Transmission and*

Distribution Conference and Exposition (T D), 2012 IEEE PES, 2012, pp. 1–8.

[29] V. Paruchuri and S. Dubey, “An approach to determine non-technical energy losses in India,” *2012 14th International Conference on Advanced Communication Technology (ICACT)*, 2012, pp. 111–115.

[30] J. M. R. Iglesias, “Follow-up and Preventive Control of Non-Technical Losses of Energy in C.A. Electricidad de Valencia,” *Transmission Distribution Conference and Exposition: Latin America*, 2006. TDC ’06. IEEE/PES, 2006, pp. 1–5.

[31] R. Alves, P. Casanova, E. Quirogas, O. Ravelo, and W. Gimenez, “Reduction of Non-Technical Losses by Modernization and Updating of Measurement Systems,” *Transmission Distribution Conference and Exposition: Latin America, 2006*. TDC ’06. IEEE/PES, 2006, pp. 1–5.

[32] A. H. Nizar, Z.-Y. Dong, and P. Zhang, “Detection rules for Non Technical Losses analysis in power utilities,” *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, 2008, pp. 1–8.

[33] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, “Support vector machine based data classification for detection of electricity theft,” *Power Systems Conference and Exposition (PSCE)*, 2011 IEEE/PES, 2011, pp. 1–8.

[34] C. C. O. Ramos, A. N. Souza, R. Y. M. Nakamura, and J. P. Papa, “Electrical consumers data clustering through Optimum-Path Forest,” *2011 16th International Conference on Intelligent System Application to Power Systems (ISAP)*, 2011, pp. 1–4.

[35] B. F. Hobbs, U. Helman, S. Jitprapaikulsarn, S. Konda, and D. Maratukulam, “Artificial neural networks for short-term energy forecasting: Accuracy and economic value,” *Neurocomputing*, vol. 23, no. 1–3, pp. 71–84, Dec. 1998.

[36] M. Gavrilas, I. Ciutea, and C. Tanasa, “Medium-term load forecasting with artificial neural network models,” *Electricity Distribution, 2001. Part 1: Contributions. CIRED. 16th International Conference and Exhibition on (IEE Conf. Publ No. 482)*, 2001, vol. 6, p. 5 pp. vol.6-.

[37] K. Padmakumari, K. P. Mohandas, and S. Thiruvengadam, “Long term distribution demand forecasting using neuro fuzzy computations,” *Int. J. Electr. Power Energy Syst.*, vol. 21, no. 5, pp. 315–322, Jun. 1999.

[38] J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, “Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection,” *Knowl.-Based Syst.*, vol. 71, pp. 376–388, Nov. 2014.

[39] D. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.