Using Text Mining for Automated Customer Inquiry Classification

Raoul Praful Jetley and Jinendra K. Gugaliya
ABB Corporate Research
Software Research Group
Bangalore, India
Email: {raoul.jetley, jinendra.gugaliya}@in.abb.com

Sana Javed
International Institute of Information Technology
Department of Computer Science
Bangalore, India
Email: sanaiiitb@gmail.com

***Abstract*—Understanding the customer's needs and issues is central to business survival and growth. Typically, customer inquiries are captured by a dedicated customer service center using unstructured narrative text. The number of such customer inquiries can vary from several thousands to millions depending on the business size and customer penetration. Analyzing such a huge number of customer inquiries manually is cumbersome, prone to errors and costly. More importantly, it is not possible to identify broad signals and trends among this data. In this paper, we describe an automated approach to analyze customer inquiries, and show how this approach can help classify customer inquiries. We illustrate the application of this approach through an example related to a leading multinational automation company.**

***Keywords*—text mining; information retrival; intention analysis; query classififcation; service center.**

I. INTRODUCTION

Customers are the biggest asset for any business organization. It becomes imperative, therefore, for an organization to ensure that its customer opinions are heard and their feedback and concerns addressed appropriately. In most organizations, this function is performed through a dedicated customer service center. For a typical organization, the customer support center is responsible for:

- receiving and addressing inquiries related to different products and services offered by the organization,
- providing a sounding board for customer questions and complaints, and directing these to the appropriate business unit, and
- interfacing with the customers to gather feedback for specific products.

A typical large organization could receive hundreds of thousands or even millions of customer inquiries every year, originating from different parts of the world. It is the responsibility of the customer service representative (CSR) to identify and channel these to the concerned person within the organization who can respond to the inquiry.

Traditionally, this has been done manually, with the CSR reviewing individual inquiries and addressing them accordingly. This approach, however, becomes infeasible as the number of inquiries increase and data becomes too large to analyze manually. What is needed is a method to automate the process of classifying customer inquiries. The automated classification can aid the CSR by reducing the effort required to manually scan each inquiry. Moreover, it ensures that a uniform classification scheme is used for classifying all inquiries and the CSR doesn't need to rely on personal expertise and judgment to make these decisions.

In this paper, we describe one such automated approach to identify and classify customer inquiries. The approach is based on text retrieval and data analytics techniques, which are used for identifying the intent of the inquiry and classifying these based on the narrative text contained within the customer inquiries.

We use this approach to classify customer inquiries for a large multinational automation engineering company. This classification helps the customer service center direct the inquiries to appropriate (teams and) business units. In addition, the classification also helps in identifying trends in the inquiries received.

As an example, consider a scenario where many complaints are received regarding licensing of a certain product. Inquiry classification helps isolate inquiries related to the product. Intention analysis isolates sales/procurement/licensing related queries within these. Finally, sentiment based filtering can be used to identify complaints among these inquiries - indicating that users are unhappy with the licensing scheme. This can then be raised as a signal by the analysis tool which is used by the analyst to take appropriate action (in this case notify the related product's sales team to review licensing policy).

The rest of the paper is organized as follows. Section II lists existing research related to this study. Section III describes the approach used for analysis. Section IV lists the results of the analysis. Finally, Section V summarizes the paper and discusses future directions.

II. RELATED WORK

Traditional analysis of customer service data has been limited to pure data and statistical analysis. Shen and Huang describe an approach for analyzing this data using singular value decomposition (SVD) [8]. Brown et. al. present several statistical techniques for analyzing data collected at a telephone call center [1]. These include forecasting based on a Poisson distribution and estimation of the mean function in a nonparametric regression.

In the context of using text analytics for custom service centers, IBM has implemented a text analysis and knowledge mining (TAKMI) tool at its own customer service centers to ease the task of the CSR. They have used this to demonstrate inquiry classification on the Statistical Package for the Social Sciences (SPSS) platform [4]. The focus of the tool is on analyzing survey forms to identify the frequency at which types or patterns exist. Moreover, the tool relies on experts

 ISBN: 978-1-61208-395-7

to feed in a comprehensive listing of product types, and does not glean this information from existing data, as proposed in our approach.

A study by Carlos et. al. [2] uses a nave Bayesian classifier to identify common patterns across customer service reports and categorize them into different buckets. The work analyzes the intentions in customer inquiries without considering which service or product is being referred. Weiss et. al. [9] on the other hand, provide a generic approach for text mining and handling unstructured information.

In our study, we demonstrate how text analytics based on information retrieval techniques can be used for performing a combination of intention analysis and data classification analysis using customer service reports.

III. Approach

Figure 1 shows an overview of the workflow used for text analysis of inquiries received by the custmomer service center. As shown in the figure, the text mining workflow employed comprises of four main processes. The first of these is indexing, which is used to extract text data from source documents and store them in a well-defined index. The Indexing process constitutes of tokenization, text pre-processing and attribute definition. Tokenization involves parsing of the text data and identifying individual terms that need to be stored in the index. The parsed tokens are then subject to pre-processing to clean the data and remove superfluous elements. Attributes or fields are identified for classifying the individual tokens before storing them as a record, or document in the index.

The next step, Information Retrieval, defines means and methods for extracting relevant information stored in the index. This is done by defining a search mechanism to query the index and retrieve relevant records that match the query. The results from the information retrieval process are then used to perform detailed analysis to identify signals and trends within the data. The analyses employed in this study include inquiry classification and intention analysis. The analysis algorithms can be applied individually or in tandem. Finally, results of the different analyses are displayed to the user in an interactive manner through generalized visualization schemes.

A. Indexing the Customer Inquiries

The process of indexing customer service inquiries entails parsing individual customer inquiries to identify the terms used therein. These terms are then stored in an index (essentially an inverse lookup table) [6]. The terms identified within the individual inquiries are added to the index generated from the data. Each token identified is added as a term in the index, and is annotated based on the corresponding field for the document.

Tokenization and Text Pre-processing: Once the tokens within each document are identified, they are added to the index, with the corresponding link to the document and associated term frequencies. However, simply adding all identified tokens to the index can introduce a number of inefficiencies, owing primarily to the commonly used terms (like articles, prepositions, conjunctions, etc.), which not only clutters up the index, but can also adversely affect the search/retrieval results. To avoid this, a common strategy is to clean the tokens to eliminate unwanted words before populating the index. Another pre-processing strategy that is commonly employed is to group together similar words (based on their meanings) to provide a semantic search capability and improve the hit rate when searching for related terms.

Commonly employed text pre-processing methods include stop word removal, stemming and lemmatization. Stop word removal entails filtering the tokens identified in the data to identify and remove commonly used words. Such words could be articles (‘a’, ‘an’, ‘the’), prepositions, conjunctions or numeric literals, pronouns among others. Additional stop words may be defined by the analyst based on the specific corpus of documents being indexed.

Stemming is another commonly employed method of text pre-processing. Stemming attempts to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form - it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

An associated pre-processing method, called lemmatization is usually used to remove inflectional endings and to return the base or dictionary form of a word, known as the lemma. If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was as a verb or a noun.

In our study, we evaluated all three text pre-processing techniques, and found stop word removal as the most effective. For the rest of the paper, we present results of our analysis when using only stop word removal during indexing.

Handling Multi-lingual Data: A special consideration needed to be made for the customer inquiry data, as it is collected from feedback received from across the globe. Consequently, the (text) data is in multiple languages, reflecting the region that the record originated from. In order to adequately process text in these multiple languages, special care is needed to translate this text to a common language before storing it in the index.

During Index Generation, the textual (narrative text) data is extracted from the multi-lingual repository a single record at a time. Each record corresponds to a user inquiry in a specific language. The record may thus consist of various fields, including a narrative text field containing the natural language user input. In a customer inquiry scenario, this could correspond to customer feedback, questions regarding specific products, request for quotations, general advertising related data, etc.

A statistical language detection algorithm is used to detect the language used within the record. The statistical detection method maintains a database for commonly used words in various languages (for each of the handled languages). When the narrative text from a particular record is parsed, the individual words in the record are matched against the commonly used words database. The matches are ranked in order of relevance. Some of these may match common words for more than one language. In such cases, the primary language and secondary languages are identified and scored accordingly. Customized heuristics are used to determine number of tokens to match to identify the language.

 ISBN: 978-1-61208-395-7

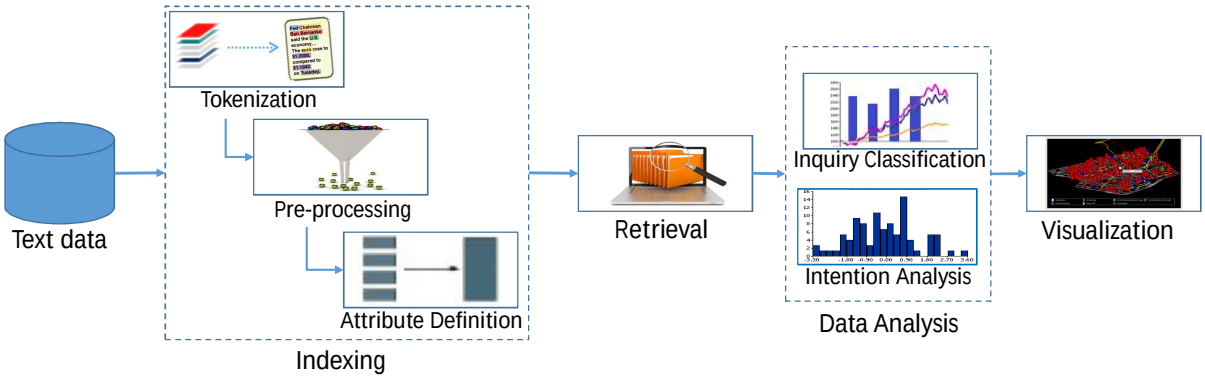

Fig. 1. Overview of the Text Mining Process for analyzing Customer Inquiries

If the narrative text data is in a language other than base language (English in our case), it is translated to generate a version of the textual data in the common language. We use an open source API, based on the Google translation service [5] to perform the translation. The translation here does not need to be perfect; simply identifying and translating key words in the text should suffice. The indexer is not so much concerned with grammar here, but simply aggregates words from transliterated text. The translated tokens are stored in a separate field. Text in English, is parsed normally (without any translation) and stored as it would be in the general case.

When searching against the index, the search string is provided in the common language. It is then searched against the translated text field for any matches. The results of the search/information retrieval thus include matches for documents in English as well as those in other languages.

B. Information Retrieval

Information Retrieval (IR) is the process of obtaining information resources relevant to an information need from a corpus of available data (the index of customer inquiries). The IR process begins with the user entering a query into the system. The query is then evaluated against the index based on the specific syntax and semantics of the query description language. Results of the query are returned as a set or list of matches, ordered according to their relevance score. The relevance score is computed using a product of term frequencies and inverse document frequencies (TF-IDF) for the given index [3]. The results obtained are then used to perform additional analysis on the data set, either manually or automatically.

It is important not only to take into account the actual words in the query, but also the meaning of the words searched for, the order of words in the query, related terms and synonyms and perform automated spell checking and diacritics restoration. The key goal of IR is to retrieve all the items that are relevant to a user query, while retrieving as few non-relevant items as possible. In order to maximize the relevance of the search results, sophisticated ranking algorithms and querying methods are used.

A number of different query methods are used to extract information from the customer inquiry index. The most trivial search method is a nave keyword based search, where individual keywords are matched against the terms contained within the index. The terms can be matched against specific fields in the index, by specifying which field they should be searched against. An extension of simple keyword search is querying using *wildcards*. Wildcards (such as '?' and '*') can be used to match for variations of a term or multiple words at once. For example, the query 'te?t*' will match the terms text, test, tester and tested.

A *Boolean query* is a simple method to combine results of two queries or add a modifier to existing queries. Most IR engines support the AND, OR and NOT operators for Boolean queries and searches. An example of a Boolean query would be "text: python OR text: boa AND class: reptile", indicating that the searcher should look for the terms python or boa in the field labeled text, while the (structured) field class is given as reptile.

Fuzzy queries are used to search for terms similar to the given word. The fuzziness factor is based on the Levenshtein Distance, or Edit Distance algorithm [7], and is used to match words that sound similar (or are spelled similarly) to the given query term. For example, when searching for a term similar in spelling to "roam" the fuzzy search algorithm will find terms like "foam" and "roams". Fuzzy searches are useful when searching for terms that may often be misspelled or have slight variations in spelling in different regions (for example, "analyze" vs. "analyse").

Proximity searches are used to search for words in a given document that are used within a specified distance of each other. For example, one may search for the words "voltage" and "meter" occurring within 10 words of each other in a document. Proximity searches are a useful tool when looking for correlations between terms.

Finally, *Range queries* allow one to match documents whose field(s) values are between the lower and upper bound specified by the Range Query. Range Queries can be inclusive or exclusive of the upper and lower bounds. Sorting is done lexicographically. A range query can be specified for numeric fields, text fields or date fields (for example, "mod_date:[20020101 to 20030101]", indicates the range of all dates from January 1^{st} 2002 to January 1^{st} 2003).

C. Intention Analysis for Customer Inquiry Data

Once the data is retrieved from the index, various analyses can be performed to discover knowledge that can aid in making

 ISBN: 978-1-61208-395-7

business decisions. The analysis techniques we have employed include *Intention Analysis* and *Inquiry Classification*.

Intention analysis was carried out to identify the intention of the customer from the text of the inquiry. The intention could be to purchase, to sell, to complain, to accuse, to inquire, etc. Manual analysis of sample data from the customer inquiry data indicated that there are four major classes in which the inquiry can be classified. These classes are summarized in TABLE I. Intention analysis was used to detect the basic four intentions and their various combinations, i.e., a total of 16 intentions. The documents for which intentions could not be found out were put under the "Unclassified" class.

TABLE I. INTENTION CLASSES FOR CUSTOMER INQUIRY DATA

Intention	Explanation	Need
Sales	To purchase or inquire (price and availability) of some product or service	To detect customer needs for products
Service	To ask for some help or assistance on some product or service	To detect customer common problems
Career Training	To seek training on some courses, job intention (job/internship)	To find suitable candidates for some job position
Complaint	To complain about some product or service	To detect customer pain points
Unclassified	Motives that cannot be classified	-

A manual analysis of data was carried out to find customers' common writing pattern. A list of intentions was created from this list. For every word of a document in the corpus, the word was matched with the intentions. If a match was found, the corresponding intention was assigned a point. The intention with the maximum weight was assigned as the class for the document. Fuzzy search was used to take care of misspelled words. The algorithm used for fuzzy searching is outlined in Figure 2. TABLE II gives sample results of intention analysis performed on the customer inquiry data.

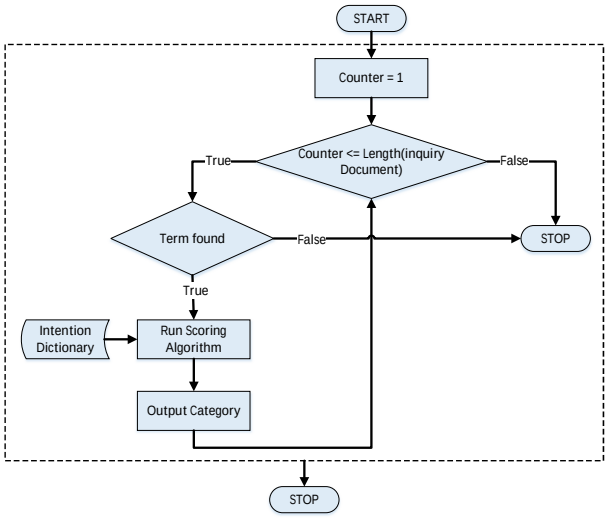

Fig. 2. Workflow for Intention Analysis

In order to check the accuracy of the algorithm, 150 randomly selected queries were taken which were also manually classified by the Contact Center (CC) executive. The result of the manual classification vs algorithm classification are summarized in TABLE III. This comparison indicates that free-flow text messages can be properly classified based on the intention of the message. Sometimes, the classification can be even improved over manual classification.

TABLE II. SAMPLE RESULTS OF INTENTION ANALYSIS

Inquiry Details	Intention Class
Please quote us your best price lead time as below controller vfd for fdf motor 440v 55kw 101a 60hz rpm 891rev min	Sales
I have a technical question in regards to a frequency drive model XV550. We have a problem with the drive where parameter maximum speed tends to change on its own. The drive is programmed to a max speed of 3510 rpm and on one occasion, the parameter changed to 1740 rpm. No other parameter changed. This could be a one-time occurrence. However, I'm just wondering if there is something that could be causing that to happen; perhaps another parameter as a result of an alarm. Your help would be appreciated	Service
Dear sir, I'd like to know about any job opportunity for business development profile. I have 6 years' experience in ac drives field and ERP arena. Due to loss in my business venture I'm looking for a suitable job in gulf countries for my drives profile	Career Training
We purchased a 150 hp 960 rpm foot mounting motor from your dealer. Now we are finding huge ampere variation. We lodged verbal complaint with the dealer who in turn forwarded the complaint to your service center. Subsequently, we were promised that a service engineer will be sent to check the motor. In spite of all our efforts and promises made, no one has visited us as yet. The amp variation is causing huge increase in electricity cost and we are unable to run our plant causing huge financial losses. We need your instant resolution of the problem	Complaint
No communication	Unclassified

TABLE III. COMPARISON OF MANUAL CLASSIFICATION VS. ALGORITHM CLASSIFICATION

	Analysis Conducted by	
	CC Executive	Algorithm
Correct Classification	109	124
Incorrect Classification	39	18
Multiple Classification	-	6
Undetermined	2	2
% Accuracy (Approx.)	73%	83%

D. Inquiry Classification

The objective of inquiry classification is proper organization of datasets, so that data can be handled with much efficiency and ease. In this study, customer inquiry data was categorized by mapping it to specific product categories. Proximity search was performed for individual products to find patterns (like commonly used specifications, name of the product, etc.), and a list of product offerings created based on the product's name and specifications. The list was sorted based on term frequency.

For example, proximity search was performed for the product "breakers" on the customer inquiry data. As result of the search, a list was generated that included the terms "circuit breaker", "generator breaker", "sf6", "air circuit breaker", "vacuum circuit breaker", "sace", "voltage breaker".

A search for the product and its associated terms was then carried out for all documents in the corpus and the matching documents were associated with the product. A similar approach was used on other products to build dictionary of words corresponding to those products.

 ISBN: 978-1-61208-395-7

An example of product classification on various product inquiries is summarized in TABLE IV. TABLE V shows how the inquiry has been categorized. The last column in the table indicates which product category the inquiry pertains to.

Though this study develops library of key words for a few popular products, a comprehensive library could be built for each and every product within the organization. A comprehensive listing of the products and services offered by the organization would need to be generated for this purpose. The list could be generated using the proximity search based approach as described above.

TABLE IV. PROXIMITY SEARCH FOR THE WORD "BREAKERS"

Term	Frequency
circuit	1575
generator	149
part	135
sf6	124
model	121
vacuum	114
air	108
sace	100
pole	90
manufacture	82
breaker	67
voltage	55
kv	49
case	49

TABLE V. PRODUCT CLASSIFICATION EXAMPLES

Inquiry Details	Product
Dear sir, we are looking for distribution transformer type 250kva. We need offer price	250kva distribution transformer
Please send your quote for 3 units current transformer 5 kva. Thank you.	5kva current transformer
We immediately need to order the following transformer pl advice with techno commercial details type distribution transformer 315kva	315kva distribution transformer
We want information on distributed control systems. Need contact name, phone number	DCS
We are in Hawaii. Can we purchase a box cooling fan for a vfd drive?	Variable frequency drive (vfd)
Looking for some help on the XS350 setup	XS350
Sir, I need a sf6 breaker up to 15 kv. Immediately so can you send me specification of this breaker?	sf6 breaker
Kindly quote us the item below 1 circuit breaker 100amps 400volts	circuit breaker

IV. RESULTS

As described above, intention analysis is used to classify the customer's intention and inquiry classification is performed to identify which product is being mentioned in the customer inquiry. Combining the results from these two analyses can provide further insight into the types and trends within the available data. For instance, combining results from the two analyses can provide information for the number of sales or service queries related to a certain product. Some of these results are presented in this section.

Figure 3 shows the distribution of customer inquiries across different types of drives. Based on the data, it can be inferred that the maximum number of sales requests are for AC system (ACS) drives. However, these drives also have the largest number of service and training related inquiries, hinting that these drives have a large number of unresolved issues, and that

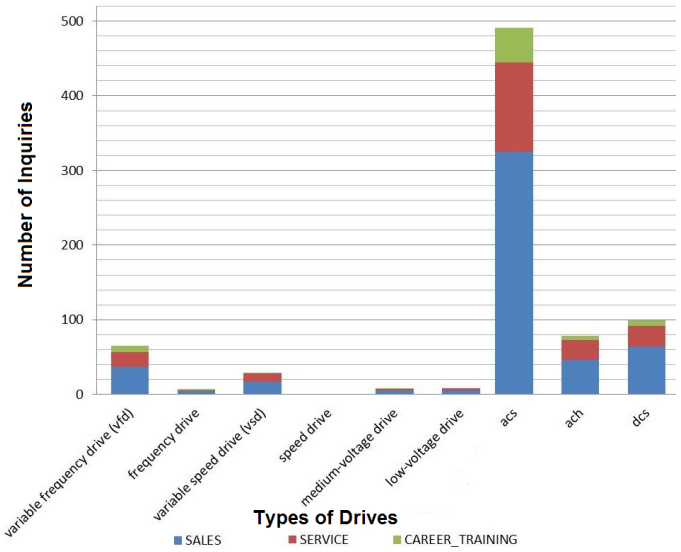

Fig. 3. Inquiries Related to Drives

customers feel the need for additional training in order to use them effectively.

Figure 4 drills down to specific products within ACS drives. The data indicates that ACS 800 and ACS 550 are the two most popular series of ACS drives, based on the number of sales inquiries. The data also shows a disproportionately large percentage of service inquiries for the drives ACS 355 and ACS 600. This could indicate that the drives have some inherent design defect that needs to be addressed. Further analysis of the specific service inquiries for these products will help identify the nature of the problem.

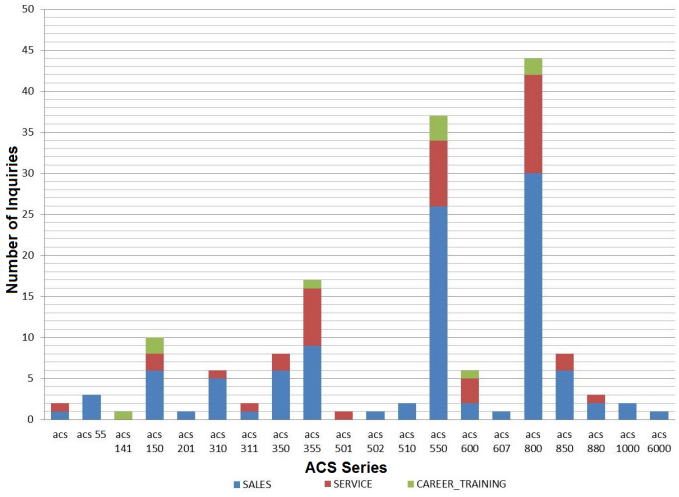

Fig. 4. Inquiries for the ACS Drives Product Series

Similarly, Figure 5 shows some data from analysis performed on transformers. The results indicate that the only service related inquiries for the transformers device class relates to power transformers. There are no service inquiries related to other transformers. This implies that customers are generally happy with the service provided for transformers.

Figure 6 indicates that the number of inquiries related to

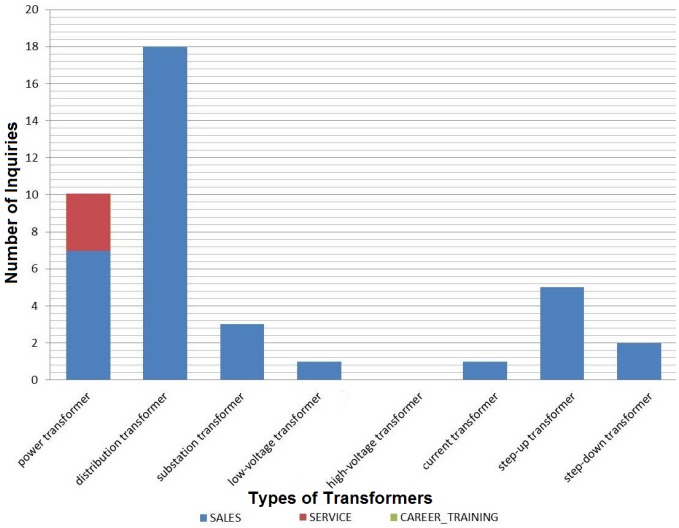

Fig. 5. Inquiries for different types of Transformers

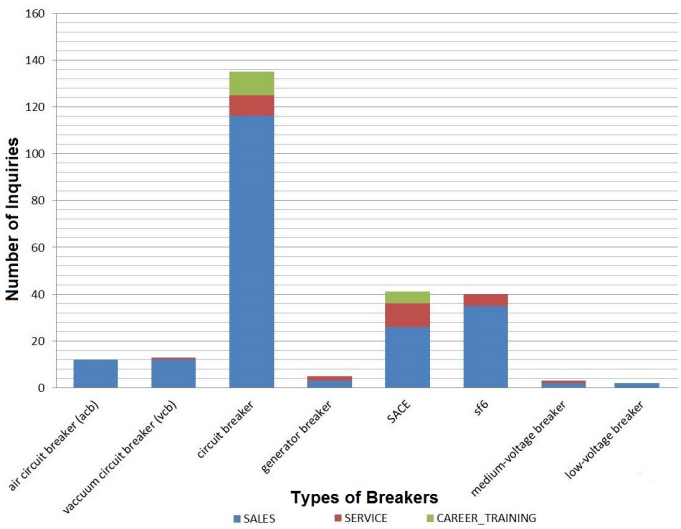

Fig. 7. Queries for different types of Breakers

training are significantly higher for Supervisory Control and Data Acquisition (SCADA) products as opposed to Distributed Control System (DCS) products. This implies that the SCADA products are comparatively more complex to configure and maintain. Thus engineers may need to focus on the usability aspects for these products.

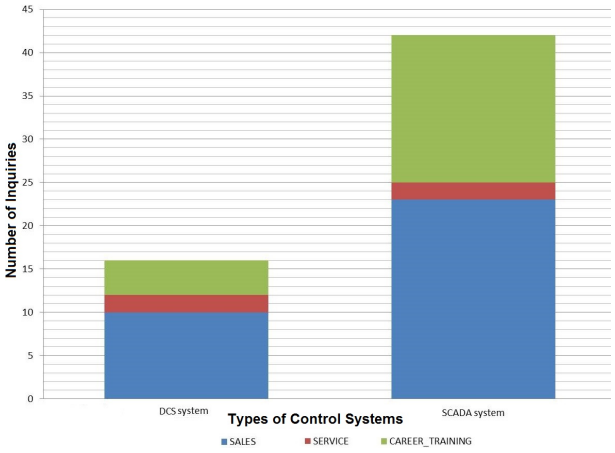

Fig. 6. Inquiries for types of Control Systems

Finally, a comparison between the circuit breakers sace and sf6 in Figure 7 shows that despite having a higher volume of sales inquiries, the sace breakers received fewer service requests and no training inquiries. This indicates that sace breakers perform better and are generally preferred by the customers when compared to sf6 breakers.

The examples above cite just a few examples of how text based analysis of customer inquiries could help organizations infer signals and trends within the data collected by its customer service center. Other such analysis can be performed at various levels of granularity to uncover additional signals and provide critical business intelligence. Further, once a signal is detected, it can be directed to the appropriate personnel for follow-up leading to reduction in manual efforts and errors and improved response time.

V. SUMMARY

This paper has illustrated the use of customer inquiry classification and intention analysis on customer inquiries, primarily in the form of unstructured multi-lingual text data. The paper has demonstrated an automated way of such analysis resulting in significant savings of manual efforts, without compromising on accuracy of analysis. In fact, in some cases the automated analysis has proved better than manual analysis.

The automated analysis provides insights into the customer inquiries not readily available to the customer service representative by simply browsing through the textual reports. The signals and trends identified through these analyses can be used to support more informed business decisions.

REFERENCES

[1] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, “Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective”. Published in Journal of the American Statistical Association, Volume 100, Issue 469, 2005, Pages 36-50.

[2] C. S. Carlos and M. Yalamanchi. “Intention Analysis for Sales, Marketing and Customer Service”. Published in the Proceedings of COLING 2012: Demonstration Papers, December 2012, pages 3340, Mumbai.

[3] T. E. Doszkocs. “From Research to Application: The CITE Natural Language Information Retrieval System”, in Research and Development in Information Retrieval, Berlin, Springer-Verlag, 1982, pp. 251-62.

[4] A. Field. “Discovering statistics using IBM SPSS statistics”. Sage, 2013.

[5] The Google Translate API. Available from https://code.google.com/p/gtranslate-api-php/

[6] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. “Sindice.com: A document-oriented lookup index for open linked data”. Published in the International Journal of Metadata, Semantics and Ontologies, Volume 3, Issue 1, 2008, Pages 37-52.

[7] V. Levenshtein. “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”. Published in Soviet Physics-Doklady,1966, 10(8):707710.

[8] H. Shen and J. Z. Huang. “Analysis of call centre arrival data using singular value decomposition”. Published in Applied Stochastic Models in Business and Industry, , May/June 2005, Volume 21, Issue 3, pages 251263.

[9] S. M. Weiss, I. N. Zhang, and F. Damerau. “Text Mining: Predictive methods for analyzing unstructured information”, Springer 2010.

 ISBN: 978-1-61208-395-7