An Application for Detecting Network Related Problems from Call Center Text Data

Ibrahim Onuralp Yigit, Engin Zeydan, and Ahmet Feyzi Ates

Türk Telekom Labs, Istanbul, Turkey 34770.
Email: {ibrahimonuralp.yigit, engin.zeydan, ahmetfeyzi.ates}@turktelekom.com.tr

***Abstract*—A Network Service Provider can receive different network related problems and complaints from various communication channels regarding their service activity in certain regions. One of the major and most important communication channels is customer call center. Detecting network related problems that customers are notifying during these calls are significant in order to provide solutions and increase customer satisfaction. However, due to sheer volume of the call records that are converted to text, it is quite difficult to analyze whole data using traditional approaches. In this paper, we study a topic modeling approach for detecting network related problems from call center text data. The analysis results demonstrate that for a major broadband service providers' *personal Internet at home tariff*, most of calls received in a customer call center is related to information, whereas the second majority of all calls are related to faults that are network related issues. These results signify the existence of network and service related issues in service providers' infrastructure.**

***Index Terms*—topic modeling, call center, text mining, network problems, network service provider**

I. Introduction

Today's service providers are receiving a growing number of data from various communication channels that they are interacting with their subscribers who are may evaluate their products and services. The data can arrive from different communication channels including social media, customer call center calls, emails, forums and blogs, customer reviews, surveys and chat [1]. This data, on the other hand, is a significant asset for gaining competitive advantages for service providers. It is mostly not easy to extract relevant information from this mostly unstructured data.

Recent advances in text mining have enabled powerful techniques to flourish. These techniques can be used to extract the desired information from a large corpus of text that most of the service providers are demanding. One of the most prominent such technique that can used along with this aim is *topic modelling* [2]. Using topic modelling, the organizations can identify the relevant topics from a huge cluster of text or corpus which are received daily from various channels in order to assist decision making process.

Topic modelling is used to find and observe topics in large cluster of text for information retrieval in an unsupervised manner. For example, in the telecommunication domain, a good topic model for a topic "fault" should include –"non-assignment", "mismatch", "disconnection"– and –"service", "campaign", "billing"– for a topic "information".

In this paper, we study a topic modeling approach for detecting network related problems from call center text data. The analysis results demonstrate that for a major broadband service providers' *personal Internet at home tariff*, most of calls received in a customer call center is related to information, whereas the second majority of all calls are related to faults that are network related issues. These results signify the existence of network and service related issues in service providers' infrastructure.

The rest of this paper is organized as follows. Section II presents related works. Section III discusses detecting network related problems of service providers. Section IV is about application details of call center text analysis. Finally, Section V gives concluding remarks.

II. Related Work

There exists many methods to obtain topics from a given text. Studies on textual data of call center voice recordings in the literature are mainly based on categorizing the complaints or creating various classification models based on the generated data at the call center. A call center data analysis of a company named Megaputer Intelligence that sells home products is performed using text mining techniques in [3]. In another study, it was aimed to extract useful information for decision making process from customer complaints by performing clustering analysis on call center voice-over text data [4]. In [5], the authors extracted the subjects on the Turkish tweets and topics under the titles, and managed to group them according to the emotional pole.

For topic modeling, there exists different works such as Latent Dirichlet Allocation (LDA) [6], Discrete infinite logistic normal [7] and Hierarchical Dirichlet processes [8]. Topic modeling can be applied to domains in computer vision, finance, bio-informatics, cognitive science, music and the social science. For example, LDA model is applied to online advertisement and search engine industrial application in [9] for big data systems.

However, none of the above approaches consider applying text mining and topic modelling approaches in the telecommunication domain using real data set of customer call records. In this paper, we investigate topic modeling techniques for extracting network related problems using the call center text data of a contact center company in Turkey. We study a generic methodology to extract text for creating topic and topic terms by utilizing MALLET topic modeling package that implements Gibbs sampling [10]. Through our methodology, we also demonstrate the distribution of call center data based on topics. Moreover, the methodology of the paper can be applied to other text-based communication channels such as emails, blogs, social media, etc. of the service provider.

III. Detecting Network Related Problems of Service Providers

Fig. 1 demonstrates a visual diagram illustrating various communication channels that can be used to contact with service providers. These communication channels may contain information arriving from customer call center calls, social media, emails, forums and blogs, surveys and chat. Customer call center data is first recorded in audio format. Then it is converted into digital text using speech-to-text analytics tools. Unstructured social media profile data can be from various sources such as Facebook, twitter, Foursquare, Linkedin, etc. Chats and Emails arrive from service provider's web sites portal. These communication channels help customers to pass their comments, suggestions and complaints information about a particular service and product into the service provider's agents.

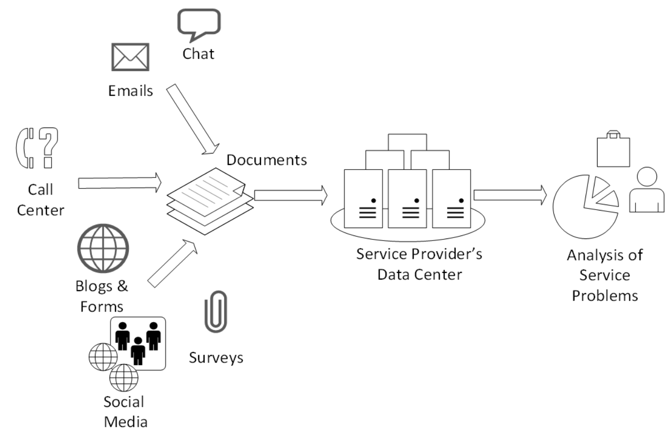

Fig. 1: Various communication channels to receive network related problems from service provider's subscribers

At the service provider's premises, all the information gathered from each communication channel is first converted into text based documents and then stored inside service provider's data center. After all data are gathered at data center, analysis of service related problems is performed as shown in Fig. 1. During analysis, customer's network related problems can be detected using the MALLET topic modeling approach in Section. IV. The methodology aims to increase the customer satisfaction via accurate analysis of all the customer related data received daily from various communication channels. This will also help in reducing the man-month efforts spent for customer analysis inside service providers.

IV. Call Center Text Analysis for Detecting Network Related Customer Complaints

In this section, we demonstrate a use case scenario for the analysis of call center text data that contains the customer call records converted voice to text. In our analysis, we ignored accuracy of voice to text converter and focused on only text analysis. The analysis is performed over a real dataset that is obtained from telecommunication provider's call data center in Turkey between September to October 2016 related to broadband service providers' "personal Internet at home tariff". The analyzed records contain $\alpha = 4,000$ calls where each call duration is approximately on average 5 minutes. The data size that is processed after converting to text is around 14 MB.

A summary of topic model building process is illustrated in Fig. 2. The first step in the process is voice-to-text conversion. In this step, all the voice records of agent and customers are converted into text document. Main component that we are interested in these documents are textual part of the call center audio records created by speech to text engine set. Additionally, structured portion of the call center logs also include call-related information such as location, call-ID, time of call, duration, originating number, terminating number, customer demographics, call agent and customers anger level, etc.

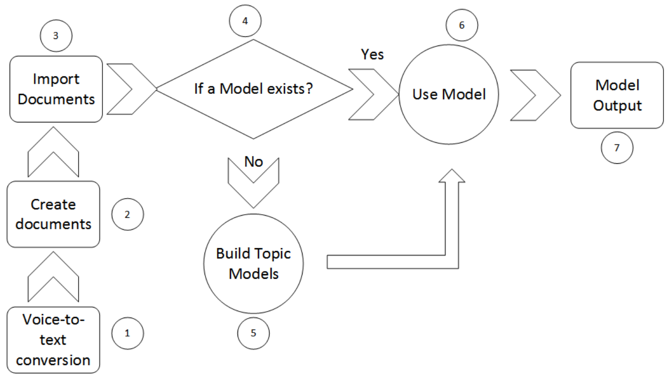

Fig. 2: The process of the execution steps of the utilized topic modeling technique

In second step, the documents are created based on the agents conversation text. Each separate call record of each agent is created as a separate text file inside a folder. In third step, all documents created in previous step are imported into MALLET's internal format. After third step, if a model exists, then we use the model in step 6. If it doesn't exist, we build the model in step-5 based on size of the collection and previous experience of customer call center agents. Some of the parameters that are used in step-5 are as follows:

- *The number of topics,* η to be extracted from the corpus is determined in this step.
- *The number of topic terms,* β in a given topic is also determined in this step.
- *The number of iterations,* γ is selected to be a constant for the convergence of the algorithm.

Additionally, in step-5 in order to build topic models, some common words are excluded from analysis. These words are selected based on the experience of customer call center experts. These words contain stop words such as names, numbers, daily routine expressions, etc.

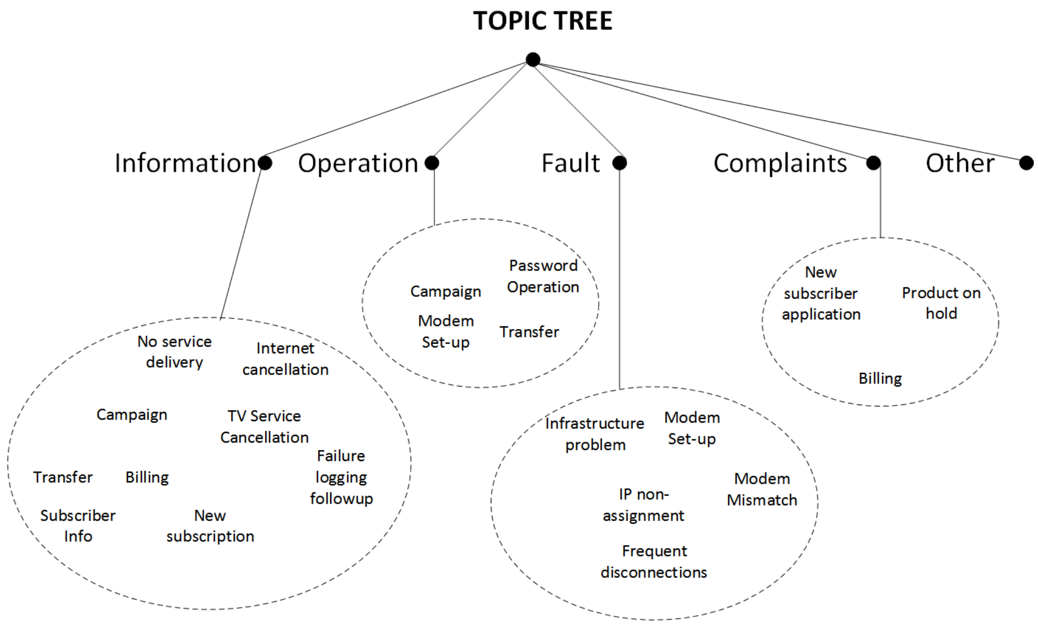

Fig. 3: Illustration of the structure of the topic tree with the topics and their corresponding sub-topics

In step-7, at the model output file, each line is a topic with individual topic terms and their corresponding weights. Some of the topic categories related to step-7 of the process are given in Fig. 3 as a *topic tree*. An example model outcome is as follows: *[0.34*(IP not-assignment) + 0.12*(modem set-up), 0.41*(Internet cancellation) + 0.02*(billing)]*. Here, the first topic can be termed as "Fault" and the second topic can be termed as "Information". It should be noted the topic terms are generated in step-7, but MALLET framework does not yield related topics. In order to generate the related topics, we have labelled the η outcome of MALLET's framework with the appropriate topic of Fig. 3 using the expertise of the call center expert.

Table. I gives an example of the studied topic numbers, topic terms and labelled topics of the outcome of the proposed methodology.

TABLE I: Topic Number, Topic Terms and Labelled Topic

Number	Topic Terms	Label
1	issue, technical, signal, modem ...	Fault
2	undertaking, address, infrastructure ...	Information
.	.	.
.	.	.
.	.	.
50	request, billing, activation, on hold, ...	Complaint

A. Analysis Results

In this section, we provide analysis results of the topic modeling methodology of Fig. 2 using the utilized data set explained in Section IV. In our analysis, we have used $\eta = 50$ as the total number of topics. Based on the specific requirement, we have selected a large number of topics (e.g. $\beta = 20$ in our analysis) due to aiming for extraction of themes and concepts. The number of iterations is selected to be $\gamma = 10$ by default.

Table II shows the distribution of the labelled topics after generating the topic terms and labelling each row of $\eta = 50$ elements with call center expert.

TABLE II: Distribution of the labelled topics

Topic	Count	Percentage
Complaints	6	12%
Fault	12	24%
Information	16	32%
Operations	8	16%
Other	8	16%

This distribution shows that among our topics 32% are about information and 24% of all topics is related to fault. This result shows that second majority of all calls arriving to service provider's call center is about faults which is related to *network issues*.

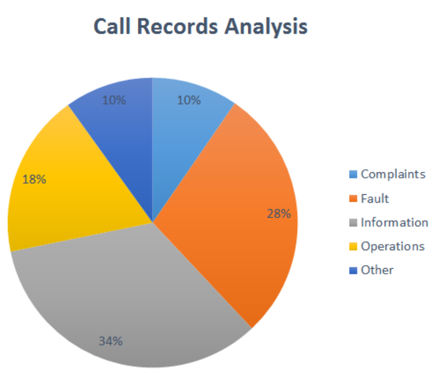

Fig. 4: The distribution of topics all call records after analysis of the proposed methodology

Fig. 4 demonstrates the distribution of topics in the data set after running the proposed methodology. In order to obtain this figure, we have mapped each call records of $\alpha = 4,000$ into one of the topics of Table II. It can be observed from this figure that the information related topic has again top percentage of 32%, whereas the network related issues with topic name fault has second most highest percentage of 28%.

V. Conclusions and Future Work

In this paper, we have analyzed the data set of customer call center records of a major broadband service providers' *personal Internet at home tariff* in Turkey. For analysis, we have utilized generic topic modeling methodology. Our analysis results have demonstrated that after the information related calls, second majority of the calls directed towards the customer call center of a service provider is related to network related issues. As a future direction, due to heavy number of records received daily in a customer call center, the call center records can also be analyzed using big data frameworks such as Apache Spark and recent unsupervised learning techniques such as deep learning framework.

References

[1] McKinsey & Company, Inc., *Higher satisfaction at lower costs: Digitizing customer care*, May 2013. http://www.mckinseyonmarketingandsales.com/higher-satisfaction-at-lower-cost-digitizing-customer-care/.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] "Call Center Data Analysis." http://www.megaputer.ru/ua/down/call_center_data_analysis_case.pdf, 2002. [Online; accessed 31-January-2017].

[4] G. Caputo, V. Bastos, and N. Ebecken, "Using text mining to understand the call center customers claims," *WIT Transactions on Information and Communication Technologies*, vol. 37, 2006.

[5] E. Akbaş, *Aspect based opinion mining on Turkish tweets*. PhD thesis, Bilkent university, 2012.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] J. Paisley, C. Wang, D. M. Blei, *et al.*, "The discrete infinite logistic normal distribution," *Bayesian Analysis*, vol. 7, no. 4, pp. 997–1034, 2012.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes.," in *NIPS*, pp. 1385–1392, 2004.

[9] Y. Wang, X. Zhao, Z. Sun, H. Yan, L. Wang, Z. Jin, L. Wang, Y. Gao, C. Law, and J. Zeng, "Peacock: Learning long-tail topic features for industrial applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 4, p. 47, 2015.

[10] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.