Arcada Working Papers 4/2018
ISSN 2342-3064
ISBN 978-952-5260-93-9

Towards Better Customer Experience: Models for Automatic Detection of Potential Misuse or Mistakes in Customer Care Process

Shuhua Liu, Patrick Jansson

www.arcada.fi

Arcada Working Papers 4/2018
ISSN 2342-3064
ISBN 978-952-5260-93-9

Towards Better Customer Experience: Models for Automatic Detection of Potential Misuse or Mistakes in Customer Care Process*

Shuhua Liu[i] and Patrick Jansson[ii]

Abstract

In this study we explore AI methods and models for intelligent customer support solutions that could help automatically supervise the support case handling process and detect potential mistakes or misuse in case handling that may lead to an unwanted customer experience. We developed different types of proof of concept classification models for automatic prediction of case status based on its conversation/activity history. We also conducted LDA topic modelling analysis to help obtain insights into the characteristics of different types of customer care cases.

Keywords: AI in customer service management; support case handling; machine learning; clustering and classification models

1 INTRODUCTION

Customer service management is a field that can potentially benefit greatly from AI (Artificial Intelligence) and ML (Machine Learning) technology powered solutions. It is possible that AI enabled solutions will change customer experience for the better, with the right technology to address the right problems.

AI's impact on customer experience can come from different forms. To many, AI in customer service means automated interaction with customer through a chatbot. A chatbot in its best form could get to know the customer or some common service questions a lot better over time, thus provide better or more personalized information service. That's certainly one way to help customer engagement and experience, to make customers feel

* Funding from our industry partner, Arcada TUF Foundation (http://tuf.arcada.fi) and Katumetro research program are gratefully acknowledged.
[i] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [shuhua.liu@arcada.fi]
[ii] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [patrick.jansson@arcada.fi]

you know them well and is able to provide the accurate and relevant information they need. On the other hand, there are so much more content in customer care than just talking to customer, especially in B2B settings, where specialized knowledge and solutions are critical to customer experience. AI can potentially make a big impact here as well in helping to automate certain processes and to complement human experts in their tasks.

In this study we explore AI methods and models for intelligent customer support solutions that could help automatically supervise the customer case handling process and detect potential mistakes or misuse in case handling that may lead to an unwanted customer experience. Our major focus is on the automatic classification of case status based on the conversation history concerning the customer support case. Our starting point is a dataset provided by our industry partner that contains a collection of email conversations concerning extracted customer cases. Our deliverable includes a machine learning powered status prediction solution that consists of components of preprocessing, feature extraction, model training, model ensemble and feedback mechanism.

2 TASK DEFINITION, DATA AND PRE-PROCESSING

Our task is to assign status label to a case given a snapshot of its conversation history at a specific time. The dataset provided by our industry partner allows us to train classifiers that can predict case status in a number of ways: 2-classes (Correct or Wrong status), 3-classes (Correct, Doubtful, Wrong) or 6-classes (their actual status in the case handling system). The labelled dataset contains 665 sample cases.

Classes	Number of unique cases
Correct	297
Doubtful	109
Wrong	259
Total	665

The major data source for case status assessment are the conversation history (i.e. email exchanges between customer and support staff) concerning the cases[1]. Before we can start to use the email texts to develop models, some careful pre-processing is needed to clean up the text and remove irrelevant or noisy content that could potentially be harmful for modelling tasks. Our preprocessing included text encoding normalization, removal of repeating whitespace characters, removing redundancy (text blocks already appear earlier), split text blocks into header and body, retaining time stamps. Output of the pre-processing is the input to the analytical process. It contains the core content of the email conversations, in sequential order[2].

[1] In addition we have access to some extra information – "Activtydetails" that are highlighted as important content in the case conversation history – we also make use of this information in our modelling efforts.

[2] We consider email "Sender" information would have been very helpful in segmenting the long case description into meaningful units. However there are actually much complexity and difficulty in doing this so we have to live with a very primitive solution for this issue during the project.

3 FEATURE EXTRACTION

Feature extraction covers both text features and non-text features. Text features are extracted from both case "Description" (full conversation history, raw data) and "Activitydetails" (highlights of important activities). We considered the options of either one feature vector for the full set of descriptions per case or N feature vectors for different parts of the case (eg. beginning, middle, end).

Word level feature is mainly the TF-IDF weighting of words and ngrams. Sentence level features are three types: (1) sentence vector of word tf-idf weights; (2) Openai language-model sentence features[3]; and (3) Facebook Infersent[4] pre-trained model sentence features. In addition, we also considered topic similarity features, to measure the similarity between a case and a case class (correct, wrong, doubtful). A case class is represented by Activitydetails of the collection of cases with the same label. The similarity is calculated as cosine similarity of tf-idf vectors.

Non-text features covers three types of information: case duration from first to last activity (time in hours as a continuous feature; average/median/max time delta between first and previous activity in case), number of interactions and current case status.

4 METHODS AND MODELS

4.1 Models for case status prediction

We included both traditional machine learning methods and neural network models in our study. While traditional machine learning algorithms require manually engineered features to reach optimal results, deep learning algorithms are capable of formulating their own features from raw data, based on what is learned to be relevant during training. Deep learning models have achieved state of the art results in building language models and analyzing sequential text data. However, proper training of deep learning models would require much larger amount of labeled data than currently available. So we focused on building smaller models using traditional learning methods and simpler neural networks.

The traditional machine learning methods included five types:

- Logistic regression
- TruncatedSVD followed by Logistic regression/other models
- Gradient boosted trees and other tree-based models
- Naive Bayes
- SVM with linear kernel and non-linear kernels

[3] Improving Language Understanding with Unsupervised Learning, available at: https://blog.openai.com/language-unsupervised

[4] Infersent at Facebook research (https://research.fb.com/downloads/infersent/).

Cross-validation is applied. When deemed necessary, we tested with both training as is and with balanced class weights.

The neural network models included the following:

- Feed-forward neural network on per-case features
- Linear layer applied per activity in case, combined to one vector for classification
- RNN(s) where each activity in case is used as a single "timestep"
- "Self-attention" model using a learned query-vector
- Autoencoder(s) for pre-training making use of unlabeled data

Our complete solution also includes model ensemble and a feedback mechanism to facilitate active learning. Ensembles of different types of single method models are expected to help bring model performance to more stable status. Feedback mechanism is incorporated to facilitate further testing on unlabled cases, to integrate user feedback on unlabeled cases and expand the labeled dataset.

In the Ensemble, models are a mix of: Logistic Regression, SVM with linear kernel, Multinomial Naive Bayes, Extra Trees classifier, Multi-layer Perceptron and Gradient Boosting classifier. Each model has one out of the text features together with other features. Ensemble is based on simple soft voting, probabilites of each model are averages together, prediction is the class with the highest average probability. The models are trained separately per classification task.

The results from our experiments on 2-class and 3-class classification models are summarized in the following tables[5]. Considering the amount of labeled samples we have, and the complexity of the problem, the performances should be considered decent. The binary classification results on the labeled dataset is notably better than the 3-class prediction results, partly due to binary classification is a smaller prediction task, and partly due to the exclusion of the doubtful class, which is the most difficult class to get right among the three. For 3-class classification, the performance on the Doubtful class is very poor, missing out on many positive examples. The logloss values are pretty high with both 3-class, indicating the probability/confidence for assigning a case to a class is relatively low. Detecting the "Doubtful" cases is inherently difficult as it is even difficult for human experts to make correct judgement about them. We have also less samples for such cases.

2 Classes: Correct, Wrong

Class	Precision	Recall	F1-Score	Support	Accuracy	Log loss
Correct	0.72	0.67	0.69	61		
Wrong	0.67	0.72	0.69	57		
Avg / Total	0.70	0.69	0.69	118	69.49%	0.5863

3 Classes: Correct, Doubtful, Wrong

[5] For privacy reason we cannot show the 6-class models.

Class	Precision	Recall	F1-Score	Support	Accuracy	Log loss
Correct	0.60	0.72	0.66	61		
Doubtful	0.40	0.08	0.13	26		
Wrong	0.62	0.72	0.67	57		
Avg/Total	0.57	0.60	0.57	144	0.6042	0.9163

4.2 LDA topic modelling analysis

Topic modeling offers a sophisticated treatment of the topic extraction problem with an unsupervised approach. Topic modelling has been widely used for tasks such as corpus exploration, document classification and information retrieval. It offers a powerful means for finding hidden thematic structure in large text collections. LDA topic modeling and its variations represent the most popular topic modelling methods (Blei et al, 2003; Blei, 2012). In topic models, topics are defined as a distribution over a fixed vocabulary of terms and documents are defined as a distribution over topics. Topic modelling results can be visualized using LDAvis visualization tool (Sievert and Shirley, 2014).

We developed an interactive tool for exploring collection of case descriptions and activitydetails using LDA analysis and visualization. We first conducted LDA analysis on preprocessed full case description (on an unlabeled dataset). It seems difficult to find meaningful clusters this way. The topics can be random and wide ranging. So we turned to analyzing the collection of Activitydetails in the labeled dataset. This seem to reveal much more meaningful clusters and patterns, help us to explore case collection from multiple perspectives, and to better understand characteristics of cases in different groups. The LDA analysis can be simple word based (with or without certain POS items) or ngram based[6].

5 DISCUSSION AND CONCLUSION

In this report we introduced our proof of concept solutions for automatic assessment of case status which is a critical element in case handling support and supervision for better customer experience. Our solution included components for pre-processing, features extraction, machine learning models, ensembles and feedback integration mechanism. We also delivered LDA topic analysis tool for exploring case collections.

To our best knowledge, automatic assessment of case status in customer care system is a problem that there is no ready solution. Research efforts are still needed to have a fully functional solution for the problem. Through this project we have developed better understanding about the problem and have identified alternative paths to better solutions.

The performances of our status prediction models for the 2-class and the 3-class tasks are decent, considering the limited amount of labelled samples we have, and the com-

[6] For data privacy reason we cannot show the topic model examples here.

plexity of the problem. However, the models as such are not ready to be put into work. In general the model performance level seems to have reached a limit set by the amount of data. The limited amount of labeled data basically rules out the use of more complicated deep learning models, which is an approach that could be revisited when much bigger amount of labeled cases become available, as deep learning models have been shown to achieve state of the art results in building language models and analyzing sequential text data. With larger labeled dataset, and more fine tuned model, or incorporation of other features, it is very hopeful that the binary classifier can arrive much higher accuracy. The models and tools we developed in this project can be used by experts for various tests and experiments, to get familiar with model behavior, to further explore the data and outputs, and to get insights about the new ways of approaching the problem and improving the solution.

LDA topic analysis and visualization tool is based on simple but powerful methods for discovering hidden patterns in large data collections. Although in the scope of this project, we could not directly benefit the case prediction models from LDA analysis results, the exercise actualy helped us to get a hint on what content are in the cases. The more meaningful and useful interpretations of the text clusters can be done by domain experts, for example, to see if multiple ways to look into the text collection can give insights into what are important in the different classes of cases, and what are the good indicators for distinction, what rules can be enforced in the final judging of case status.

Our complete solution also included model ensembles and a feedback mechanism. Ensembles of different types of single method models helps to make model performance become more stable. The feedback mechanism is incorporated to facilitate further tests on unlablled cases, to integrate user feedback on the predictions and expand the labelled dataset.

REFERENCES

Blei, D, Ng, A., and Jordan, M. I. 2003. Latent dirichlet allocation. Advances in neural information processing systems, pp.601-608.

Blei, D. 2012. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012

Hoffman M, Blei DM, and F Bach. 2010. Online learning for Latent Dirichlet Allocation, Advances in Neural Information Processing Systems 23, pp.856-864.

Sievert C. and K. Shirley. 2014. LDAVis: A Method for Visualizing and Interpreting Topics, ACL Workshop on Interactive Language Learning, Visualization and Interfaces, June 27, 2014, Baltimore, USA.