Proposal of Customer Feedback Topic Extraction

Author: Yuhou Zhou

Company Supervisors: Sadok Ben Yahya, Fatih-Mehmet Inel

University Supervisor: Prof. Dr. Adalbert F.X. Wilhelm

Introduction

Customer feedback is information provided by clients about whether they are satisfied or dissatisfied with a product. In CEWE, one method to collect customer feedback is asking customers to complete an online questionnaire.

By collecting customer feedbacks, the extracted useful information can act as a guide to improve products. Feedback gives the company the clue of customers' general experience and the satisfaction rate. Sending surveys to ask customers' suggestion also shows the attitude of listening and valuing opinions from them. Except the description provided by sellers, the publicly posted feedback is useful information to other consumers, imparting them an overview how the product really is. This information also helps managers to make correct business decisions.

The traditional way of processing user feedback requires much manual effort, and this can be very expensive and inefficient. At one time, people can only process a small batch of text from recent period. Thus, the useful information, which concerning long lasting but less frequently appeared issues, may lost.

Current situation

How the feedback is collected?

Few days after customers purchase products, they will receive emails asking them to complete questionnaires. One of the questions is asking the satisfaction level of the purchased product (from extremely unsatisfied 0 to extremely satisfied 10) and them to explain why they give their rates. After customers finish and send their questionnaires. The feedback is collected and stored in the database.

How the feedback is processed?

Currently, by using the dictionary method, the feedback are categorized into nine categories. Eight out of nine categories are concerning certain issues or products; the feedback which is hard to be categorized will be put into the category 'Other'. The categorization process is automatically executed every morning. After a certain time period, working staff will manually check the categories and summarize topics under each category. The topics will be used in different data products, for example, the explanatory text on bar charts.

Goals of the project

Though we already established an automatic customer feedback classification system, it can be improved in following aspects:

- The categories are human predefined. The category "Other" may contains topics of interest.
- Extracting topics from categories are completed by human. This is expensive and inefficient and may neglect some issues lasting for a long time but appearing with a low frequency.
- As releasing of new products or new versions of software, the present categories will be out-of-date. Much information will be miscategorized or categorized into 'Other'.

The project will tackle these challenges by

- Using machine learning methods to extract topics
- Setting up a workflow to deploy the model and, when necessary, update it

Difficulties

Two obstacles are about to be solved.

The feedback is written in German. Nowadays, many natural language processing libraries have poor support to German language. Either they cannot train models based on German, or they only provide limited features to German.

The variation of text is dramatic and unpredictable. This difficulty is reflected on two aspects. First, the quality of the feedback is varying. Some feedback contains totally useless information, such as "Nein", the copy of the survey email. The length of the feedback varies from one word to a long text containing hundreds of word. Second, the content of new feedback is unforeseeable. It may include new products, new services, etc.

Solutions

For the task of topic extraction. Data cleaning will be performed to filter useless information. Since there is no labelled data, an unsupervised topic extraction method will be chose to train a model on previous customer feedback. Some metrics will be picked to measure the function of the model.

The tools expected to be used and compared are Google Cloud Platform (GCP) Natural Language, Python spaCy, and Gensim. GCP Natural Language provides German support and offers easy to use APIs. Spacy also supports German and is highly optimized. It has very good interfaces with other machine learning libraries, such as TensorFlow, scikit-learn. Gensim is a python library which is highly efficient and scalable towards topic and semantic modelling.

For model deployment, the model will be retrained based on a time period. Outdated observations will be removed, and new observations will be included into the training process. This process can be scheduled using an ochestrator tool, such as Apache Airflow.

Summary

In this project, two goals are about to be achieved. I will compare different tools, such as GCP AI products, Python spaCy, etc. to find the optimal solution and retrieve literature to find the ideal learning methods in the current case. Sample experiments will be performed to test the functionality of the workflow. Clustering metrics will be picked to evaluate the learning result. After training, the deployment process will be considered.