Proposal of Customer Feedback Topic Modeling

Yuhou Zhou, yu.zhou@jacobs-university.de

July 24, 2019

Company Supervisors: Sadok Ben Yahya, Fatih-Mehmet Inel

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

1. Introduction

Customer feedback is information provided by clients about their general experience with a product or a service. In CEWE, one method to collect such feedback is asking customers to complete online survey.

By collecting customer feedback, the extracted information can act as a guide to improve products. Feedback gives the company the clue of customers' general experience and the satisfaction rate. 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 current way of processing user feedback requires that working staff reads the feedback and later extracts topics. This process can be costly and inefficient, because of manual intervention. At one time, people can only process a small batch of text from a recent period. Thus, the information, which concerning long lasting but less frequently appeared issues, may lost. This proposal will present a workflow, which can extract topics from customer feedback without manual effort.

2. 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, the feedback is categorized into nine categories, by using the “dictionary method”, where feedback falls into one category if the feedback contains keywords of the category. Eight out of nine categories are concerning certain issues or products; since the dictionary cannot include all the keywords of one category, and the eight categories do not include all information, the feedback which is hard to be categorized will be put into the category 'Other'. The categorization process is automatically executed every morning. The insight collected from the categorization of the customer feedback will be addressed to different employees in their respective department.

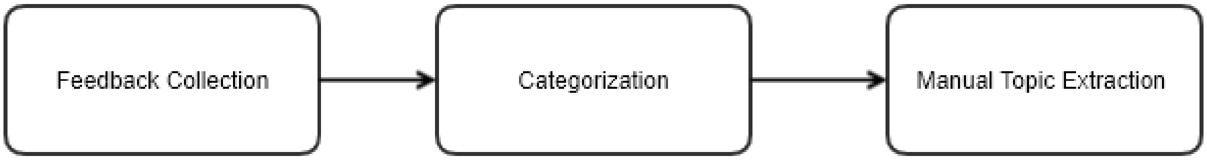

Figure 1 - Current Topic Extraction System

3. Goals

- Using machine learning methods to extract topics.

  Since the categories are human predefined. The category "Other" may contains topics of interests. Machine learning methods can catch the information which predefined categories do not cover.

  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. On the contrary, machine learning methods are executed automatically and able to make use of a large amount of data.

- Setting up a workflow to deploy the model and, when necessary, update it

  As new products or new versions of software are released, the present categories will be out-of-date. Much information will be miscategorized or categorized into 'Other'. The trained machine learning model will be updated, when the system finds new data should be included to train a new model.

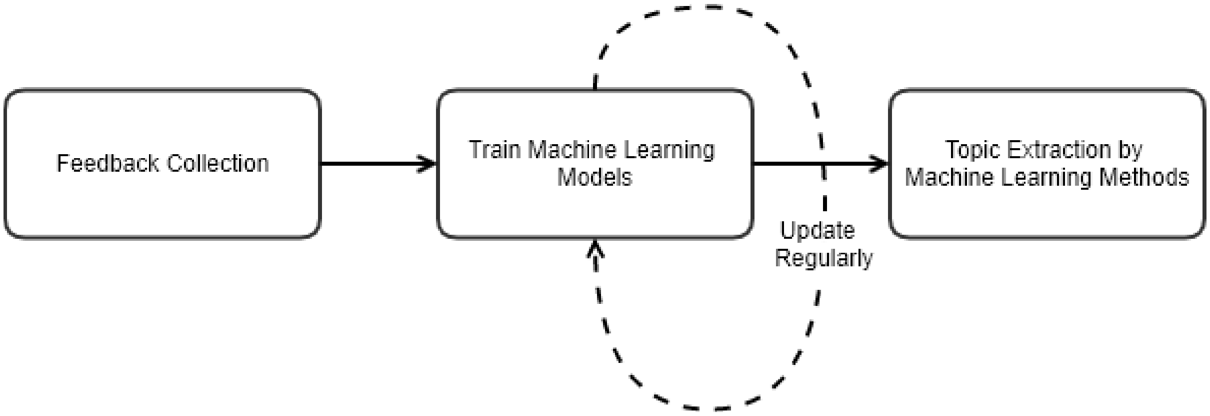
Feedback Collection
Train Machine Learning Models
Topic Extraction by Machine Learning Methods
Update Regularly

Figure 2 - Proposed Topic Extraction System

4. Related Work

Discovering “topics” from a collection of documents has been a long-standing problem in natural language understanding. Latent Semantic Analysis (LSA) acts as the base of recent topic models [2]. Hofmann proposed Probabilistic Latent Semantic Indexing (pLSA) [3], bringing topic models from deterministic to probabilistic. In 2003, Blei described Latent Dirichlet Allocation (LDA) which is a Bayesian alternative to pLSA [1]. LDA makes use of word co-occurrence information in documents, to generate document-topic distribution. However, the lack of word co-occurrence in short texts prevents LDA yielding satisfying results [9]. Such short texts are news headlines, tweets, user feedback, questions and answers, and so on. One strategy to overcome this data sparsity problem is to concatenate a set of short texts according to available metadata [6, 7, 8]. For example, aggregating texts by timestamps, locations or users. The evaluation plans for topic models includes perplexity, topic coherence, etc [4 ,5].

5. Proposed Approach

Data preparation 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.

Two obstacles are about to be solved. The feedback is written in German is one of the difficulties. 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 other difficulty is the variation of text is dramatic and unpredictable. This 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 words. Second, the content of new feedback is unforeseeable. It may include new products, new services, etc.

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.

6. 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.

7. References

[1] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003).

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990).

[3] Hofmann, T. Probabilistic Latent Semantic Analysis. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), (1999).

[4] Wallach, H. M., Murray, I., Salakhutdinov, R. & Mimno, D. Evaluation methods for topic models. in Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09 1–8 (ACM Press, 2009). doi:10.1145/1553374.1553515

[5] Röder, M., Both, A. & Hinneburg, A. Exploring the Space of Topic Coherence Measures. in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15 399–408 (ACM Press, 2015). doi:10.1145/2684822.2685324

[6] Weng, J., Lim, E.-P., Jiang, J. & He, Q. TwitterRank: finding topic-sensitive influential twitterers. in Proceedings of the third ACM international conference on Web search and data mining - WSDM '10 261 (ACM Press, 2010). doi:10.1145/1718487.1718520

[7] Hong, L. & Davison, B. D. Empirical study of topic modeling in Twitter. in Proceedings of the First Workshop on Social Media Analytics - SOMA '10 80–88 (ACM Press, 2010). doi:10.1145/1964858.1964870

[8] Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13 889 (ACM Press, 2013). doi:10.1145/2484028.2484166

[9] Li, C., Wang, H., Zhang, Z., Sun, A. & Ma, Z. Topic Modeling for Short Texts with Auxiliary Word Embeddings. in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16 165–174 (ACM Press, 2016). doi:10.1145/2911451.2911499