Improvement of a Ticketing System

WENLIANG PENG, YUANHAO ZHONG, CHANG ZHAO, and YUJUN LIU

In the context of the advanced practical course “Machine Learning for Information Systems Students” our task was to improve the Ticketing System of UCC company. During the project, we introduced machine learning algorithms to improve the system to meet the needs of both the users and the company as much as possible. Our main task is to use machine learning algorithms to cluster the problem descriptions initially entered by users. First of all, we separated the text information from the dataset, and then preprocessed the data for different languages (mainly English and German) after language classification. We then built the models. After building the model we carried out post-analysis and model evaluation, and finally identified possible future directions.

1 BACKGROUND AND PROBLEM

1.1 Background and Purpose

Ticketing system is a very important channel for companies to get feedback and provide support to users. Nowadays most companies on the market have their own support systems. Our project is for UCC’s support system. The UCC is an education-as-a-service company at the TU Munich that provides SAP services (hosting, support, training) to educational institutions (e.g., other universities). It internally uses an SAP ticket system based on a standard configuration to handle incoming customer requests. Requests can be submitted by creating a ticket in the system. The overall goal of this project is to identify ways to improve this support system, with the help of Machine Learning, to improving classification accuracy, reduce the time and effort needed from the staff and improve customer experience.

1.2 Conclusion

We can introduce Machine Learning algorithms to improve the system and meet the needs of users and companies. The optimization scheme we propose is to isolate the problem description input by the users for the first time, extract the effective information from it, and use the clustering algorithm to provide analysis and basis for subsequent problem classification, and mine ticket text information. For example, the number of clusters is 2, one cluster is the problems with specific description which can be directly assigned to different departments, another is the rest problems with ambiguous description which should be assigned by supporters.

2 MODELLING TASK

2.1 Data Extraction

The original data we need to use is the descriptions of the problems sent by users for the first time, that is, the text content of the type "Beschreibung" and "E-mail von Kunde". Due to Alert tickets, blank tickets and tickets with missing problem descriptions, we first need to isolate all valid tickets. After loading all the tickets, we built a new data frame to read all valid tickets. Then after verification, a total of 6,335 tickets we need to use are obtained. By extracting texts of type "Beschreibung" and "E-mail von Kunde", we got the problem descriptions sent by valid users.

This report is a part of the lecture, Master-Praktikum - Machine Learning für Wirtschaftsinformatiker (IN2106, IN2128, IN212812), Technical University of Munich.

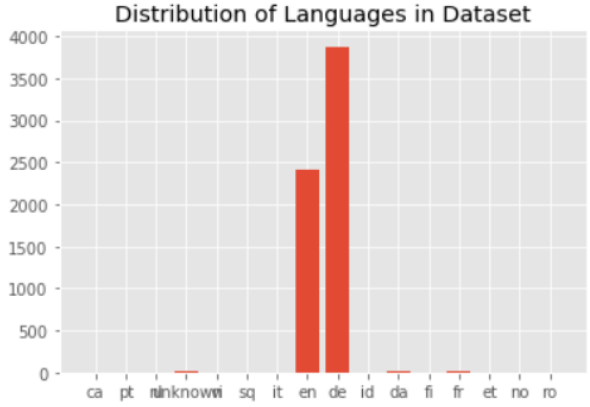


Fig. 1. Language distribution of the tickets.

2.2 Multi-language Processing

In this part, we are going to determine the language of each text in the data frame and also process tickets in different languages separately. First, the languages need to be identified so that we can know how to handle these instances. We used the library langdetect to detect the language in the dataset.

As can be seen from Fig.1, the main languages in the dataset are English and German. Some other individual tickets were detected as some other languages, but after validation, we found that these were misclassified and they are still basically English or German. Therefore, we can consider mainly English and German, and only use tickets that are recognized as German and English text. If a subsequent user enters a ticket in another language, the system can simply forward the ticket to a member of staff for manual classification. The total number of tickets used and the corresponding languages are shown below:

Total number	6283
English tickets	2406
Germany tickets	3877

2.3 Preprocessing

In this section, we performed preprocessing operations on the text data of all valid tickets (6283 in total). A series of common natural language processes were used to make the data form more uniform, remove useless distractions and facilitate subsequent vectorization. The preprocessing we used is as follows.

1) Removal of sensitive words. By removing the names of people and locations in the text, on the one hand, the interference of these unimportant words can be reduced, and on the other hand, the dimensionality of the text vector can be reduced.

2) Removal of spaces and line feeds. This step can remove useless information such as placeholder spaces.

3) Lower casing. Converting all words in the dataset to lower case is an effective way of removing redundancy. This is because words may appear several times in the text, some of them in lower case and others in upper case.

4) Removal of punctuations. This step can remove punctuation marks that have no real meaning. It is worth noting that the meaning of the words containing some punctuations may change if the punctuations are removed directly. For example, for the word "u-user", if we directly remove the hyphen, it will turn into another wrong word "uuser". So, we used spaces instead of punctuations to solve this problem.

5) Removal of stop-words. Usually, text data contains a large number of stop-words. These words haven't any significant impact on the data, so we can remove them.

6) Removal of frequent words. There are a lot of high-frequency words appearing in a lot of tickets, such as some greetings, which may cause some interference to our model. Therefore, we filtered out the 130 highest frequency words, but manually kept the words that were actually meaningful to the problem description.

7) Removal of Rare words. In this step, we removed words that only appeared once/twice. This greatly reduces the dimensionality of the text data and prepares for the subsequent vectorization.

8) Removal of numbers and words starting with a number. The text data also contains a lot of numbers that are not relevant to the content of problem descriptions, such as time information. So, we also should remove them.

9) Lemmatization. This is a very important step in preprocessing. In simple terms, lemmatization is the process of removing the affixes from a word and extracting the main part of the word. Lemmatization is based on the dictionary and transforms the complex form of a word into its most basic form. Instead of simply removing the prefixes and suffixes, lemmatization transforms the word according to the dictionary. This method can help us to find valid text information quickly to a large extent.

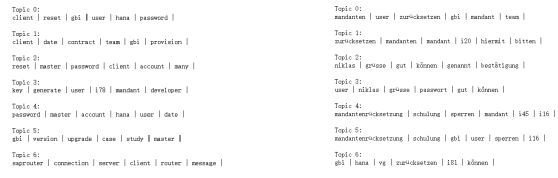
As a comparison, we also used the method of stemming. Stemming is the process of converting any word in the data to its root form. It is a process based on patterns in linguistic morphology that removes inflected or derived forms of affixes and unifies the different forms of a word into one representative standard form (stem). By comparing the results, we found that lemmatization was better at reproducing German words, so we chose to use lemmatization.

10) Vectorization. This step is to transform the preprocessed text data into vectors that can be used as input to the clustering model. We used the TF-IDF algorithm to implement the text vectorization. TF-IDF (term frequency-inverse document frequency) is a common weighting technique used in information retrieval and data mining. TF is the Term Frequency and IDF is the Inverse Document Frequency. The algorithm uses the frequency of word occurrences as the feature of text data. To implement this algorithm, we use the "Tf-idf Vectorizer" in the library scikit-learn to preprocess and vectorize for the algorithms. As shown in the figure below, we can call the ". get_feature_names ()" function to view the generated features.

2.4 ML Model

2.4.1 Truncated SVD. After preprocessing, we got a vector X, which stores the text information of all tickets. We now want to apply a clustering algorithm to this vector X to study the topic focusing on the text information of all ticket samples. We first tried the method of latent semantic analysis (LSA). Through topic analysis by LSA, we can get the key topic in the corpus, which means the importance of each word in the topic, and the tendency of each ticket sample on each topic. And based on them, keywords and representative texts corresponding to the theme can be obtained. LSA is just a truncated singular value decomposition of a (very high-rank and sparse) document-term matrix, only retaining the largest singular value of n_components. The process of LSA is as follows: Using Truncated SVD, transform the original feature vector X with a scale of N*D (number of ticket's samples, number of words) into a new feature vector X_topics with a scale of N*T (number of ticket's samples, number of topics).

Svd_model.components_is a vector with a scale of T*D (number of topics, number of words), and the elements in (i, j) represent the weight of word j on topic i. The higher the weight of word j, it can be considered more representative on the topic I, and we use this to select the most representative keywords of the topics. The result of English and German keywords after truncated SVD are shown in Fig.2.



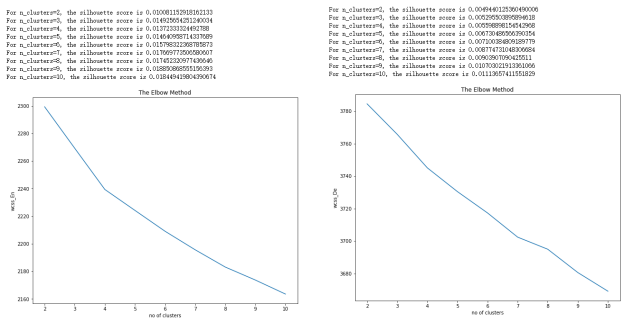
(a) English Data (b) Germany Data

Fig. 2. Topics after Truncated SVD.

Although we got the topics using LSA, we did not find a suitable method to match the topics with the ticket's samples, so we tried to use K-means for clustering.

2.4.2 K-means Clustering. A very important part of K-means is to find the best number of clusters. The basic step of any unsupervised algorithm is to determine the optimal number of clusters to which the data can be clustered. The Elbow method and Silhouette Analysis are the popular methods to determine the optimal value of k. The elbow curve is similar to a human elbow, and the value of k corresponding to the "elbow joint" part is the most appropriate k, but sometimes the "elbow joint" of the elbow curve is not obvious, so we can use the silhouette score to analyze, the silhouette score is a criterion for judging whether the clustering is good or bad, and it is calculated by combining two indexes of the degree of aggregation within a cluster and the degree of separation between clusters. Fig.3 shows the output result of using the elbow method and silhouette analysis to determine the optimal number of clusters.

We choose the number of clusters to be 4, 6, and 7 to visualize the data. As we mentioned before, the feature vector X we got after

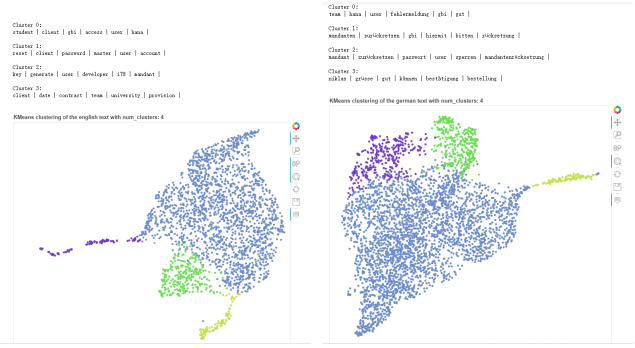


(a) English Data (b) Germany Data

Fig. 3. Results of elbow method and silhouette analysis.

Tf-idf Vectorizer processing is high-dimensional, and we need to visualize the cluster distance space obtained after K-means clustering on the 2D plane. Uniform Manifold Approximation and Projection (UMAP) is a brand-new dimensionality reduction technology, especially suitable for the visualization of high-dimensional data sets. Its goal is to find a representation of high-dimensional points in a low-dimensional space (usually a 2D plane). In terms of visual quality, compared with t-SNE, UMAP retains a more global structure and has superior operating performance.

Similar to the svd_model.components_ we referred to before, kmeans.cluster_centers_ is also a vector with a scale of C*D (number of clusters, number of words), and the elements in (i, j) represent the weight of word j on cluster i. The higher the weight of word j, the more representative it is on cluster i, and we use this to select the most representative keywords of clusters. Fig.4 is an example of when the number of clusters is selected to be 4, output 6 keywords for each cluster, and visualize the data points of all tickets' samples.



(a) English Data (b) Germany Data

Fig. 4. Result after clustering.

When we click on one of the tickets' samples with the mouse, we can get some information about this ticket's sample, such as information type, time, ID, etc. The most important is the text information, as shown in the Fig.5 and Fig.6, we have selected a sample from each cluster as a display. By looking at the keywords in the

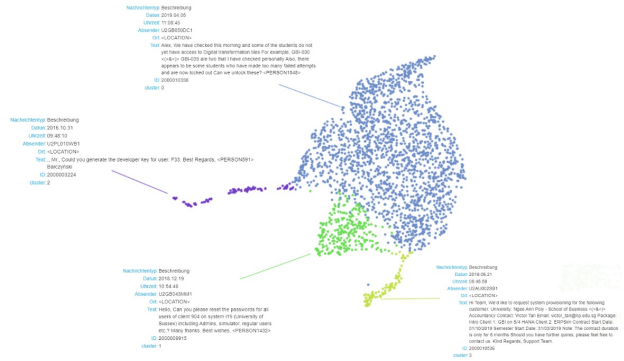


Fig. 5. Clustering visualiztion for English data.

Fig. 6. Clustering visualization for German data.

cluster corresponding to the sample we can see that the textual information of the sample corresponds roughly to the keywords of the cluster to which the sample belongs. In addition, we found that English Cluster 3 contains a lot of sensitive information. Almost all of the names of people, schools, and contact details are not treated for sensitive words. Therefore, the sensitive word processing in the ticket support system can be paid more attention to the data in this cluster.

2.5 Post-analysis

Through the establishment of the models, we obtained the topics that are most relevant to the descriptions of the problems entered by all users and the clusters according to the current datasets. In this step, we will analyze topics and clusters separately.

2.5.1 Meaning of Topics after Truncated SVD. Through the truncated SVD process, we downscaled the high-dimensional text data to the most important 7 dimensions/feature space and printed out the first 6 keywords of the 7 topics, the results are shown in the Fig.7.

From the Fig.7, we can see that our users ask more questions about password reset, key development, gbi upgrade in both English and German. By comparing the keywords in the English and German topics, we found that the German topics had more meaningless

NO. of topic	Keywords (English)	Keywords (German)
1	client | reset | gbi | user | hana | password	mandanten | user | zurücksetzen | gbi | mandant | team
2	client | date | contract | team | gbi | provision	zurücksetzen | mandanten | mandant | i20 | hiermit | bitten
3	reset | master | password | client | account | many	niklas | grüsse | gut | können | genannt | bestätigung
4	key | generate | user | i78 | mandant | developer	user | niklas | grüsse | passwort | gut | können
5	password | master | account | hana | user | date	mandantenrücksetzung | schulung | sperren | mandant | i45 | i16
6	gbi | version | upgrade | case | study | master	mandantenrücksetzung | schulung | gbi | user | sperren | i16
7	saprouter | connection | server | client | router | message	gbi | hana | vg | zurücksetzen | i81 | können

Fig. 7. Keywords of the topics.(The German and English topics in the table do not correspond to each other.)

NO. of clusters	Keywords (English)	Keywords (German)
1	key | generate | user | i78 | developer | mandant |	mandantenrücksetzung | schulung | sperren | vertragsende | i45 | i16 |
2	gbi | version | upgrade | client | access | update |	gui | domain | adresse | zugreifen | remote | folgend |
3	password | master | reset | user | a ccount | lock |	hana | team | vg | können | fehlermeldung | gut |
4	student | case | create | study | user | account |	mandanten | gbi | zurück | rücksetzung | setzen | mandant |
5	client | date | contract | team | university | provision |	niklas | grüsse | gut | können | bestätigung | bestellung |
6	reset | client | refresh | many | erpsim | i45 |	zurücksetzen | mandanten | mandant | bitten | i20 | passwort |
7	access | hana | client | domain | university | team |	user | master | passwort | anlegen | team | entwicklerschlüssel |

Fig. 8. Keywords of the clusters.(The German and English topics in the table do not correspond to each other.)

words such as bitten, grüsse, etc. The English topics are more meaningful. These topics give us a general idea of the common questions asked by the users. We can also improve the FAQs by including more frequently asked questions in the FAQs based on these topics. For users, this will improve the efficiency of solving common problems. For the company, these topics are a summary of the user's problems and can be used as a part of the overview of the data.

2.5.2 Meaning of Clusters after K-means Clustering. Through the K-means clustering method, based on the principle of the minimum sum of Euclidean distance, we clustered the preprocessed text data into 7 clusters. As shown in Fig.8, it is the first 6 keywords of the 7 clusters of English data and German data.

Through the comparison of keywords in English and German clusters, similarly, we find that German clusters have more meaningless words, and the clustering of the English model is more meaningful. Therefore, for English data, we have obtained the significance of 7 clusters in practical applications through analysis:

As can be seen from Fig.9, the clusters obtained by the K-means algorithm are meaningful for users' problems classification. For example, for the cluster of Password/Account reset, the keywords given by the K-means algorithm include master as well as lock, which can give some information about the reason why a ticket belongs to the cluster of Password/Account reset.

3 EVALUATION

In this section, we evaluated our model as well as the project from two perspectives. By analyzing, we gained a deeper understanding of the model as well as the project.

NO. of clusters	Keywords (English)	Meaning
1	key | generate | user | i78 | developer | mandant |	Key generation or developer
2	gbi | version | upgrade | client | access | update |	GBI version upgrade
3	password | master | reset | user | a ccount | lock |	Password/Account reset
4	student | case | create | study | user | account |	Questions/Case study from students
5	client | date | contract | team | university | provision |	Questions about contracts
6	reset | client | refresh | many | erpsim | i45 |	ERPsim relevant
7	access | hana | client | domain | university | team |	Questions about HANA

Fig. 9. Keywords and meaning of the English clusters.

3.1 Evaluation of the Model

In our previous modeling work, we modeled the English and German datasets separately to obtain separate clustering models for German- and English-speaking users. By comparing the German and English models we found:

1) In the loss function graph, when finding the best number of clusters by the elbow rule, the elbow of the English model is more pronounced, which allows us to find the number of clusters with better performance.

2) The English clustering model has a better Silhouette score in terms of the performance of the clustering results.

3) In the English model, the keywords in each topic and cluster are more representative and meaningful.

The results of the above comparison indicate that the English model has a better performance compared to the German model. Possible reasons for this are:

1) The libraries used for preprocessing are more suitable for English data.

2) German itself is more difficult to preprocess. For example, for German conjugation, separable verbs can be not handled well, resulting in many stems being read incompletely or incorrectly.

3.2 Evaluation of the Project

We carried out a comprehensive and systematic evaluation of the entire project by using SWOT, which is a strategic planning technique used to help us identify strengths, weaknesses, opportunities, and threats related to project planning.

The strengths of our project are:

1) The project provides the most frequent topics by users, which gives a better understanding of the data entered by users, and can also provide a reference for the setting of FAQs.

2) The user's problem descriptions are broadly clustered and can be used as part of the preprocessing for the subsequent classification task.

3) The models don't rely on data from the ticket support system forms that may contain incorrect information but use the initial text information directly, which can ensure the correctness of the model.

4) The project provides a very clear process and approach to modeling using text data, and the analysis is very intuitive and can provide a basis and reference for subsequent modeling.

The weaknesses of our project are:

1) The project is not directly user-oriented and can only be used as a small early part of the ticket support system. Therefore, it doesn't have sufficient product value.

2) In terms of the model results, the keywords of some topics and clusters obtained are not clear, and some keywords have no practical meaning, such as bitten, grüsse, etc.

3) The preprocessing process doesn't work as expected: some irrelevant information cannot be completely removed, and some words cannot be restored, especially for the German data.

The opportunities and development prospects for our project are diverse. There are many directions for optimization and subsequent application of the project. The details will be carefully presented and discussed in the next section.

The threats of our project are:

1) Inaccurate keywords of topics as well as clusters can have a direct negative impact on subsequent applications.

2) In contrast to the classification model, the samples in each cluster generated by our unsupervised learning model are not clearly labeled. Therefore, the clustering model is not as straightforward and efficient as the classification model when user problems are assigned to different departments for processing.

4 POSSIBLE FUTURE WORK

4.1 Improvements of Language Classification

When we performed language detection and classification on the initial data, we found that a portion of the data was misclassified as other languages. As we model different-language data separately, language classification errors can directly affect the accuracy of our models and results. Based on this problem, improving the accuracy of language detection and classification is an important research and development direction for our ticket support system.

4.2 Improvement of Preprocessing

From the previous analysis, we found that the German model doesn't perform as well as the English model. The main reason may be found in the preprocessing. Therefore, for German data we should adopt some more suitable preprocessing methods: 1) Use German-specific preprocessing libraries, such as "german_stopwords_full". 2) Use German-specific preprocessing methods, such as Part of speech tagging (POS), the Compound split of words, etc.

4.3 Improvement of the Model - LDA

Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. We have tried LDA to analzye our dataset and showed part of results in Fig.10, but the result might be not good enough due to ticket number limitations.

```
Topic of cluster 0:
0.014*"client" + 0.014*"service" + 0.013*"message" + 0.012*"server" + 0.010*"remote" + 0.009*"download" + 0.009*"reset" + 0.009*"team" +
0.009*"group" + 0.009*"gbi"
Topic of cluster 1:
0.027*"client" + 0.021*"hana" + 0.016*"university" + 0.013*"team" + 0.011*"gbi" + 0.010*"request" + 0.010*"contract" + 0.010*"domain" +
0.009*"query" + 0.008*"erpsim"
Topic of cluster 2:
0.056*"client" + 0.038*"team" + 0.022*"date" + 0.021*"hana" + 0.019*"contract" + 0.018*"gbi" + 0.016*"customer" + 0.015*"university" +
0.014*"erpsim" + 0.013*"provision"
```

Fig. 10. Top 3 clusters after LDA for English data.

5 PERSONAL EVALUATION

After working on this project for the last month, we are satisfied and happy with the structure of our project and the final model. Starting with the analysis of requirements, we had a lot of brainstorming and discussions. Based on the suggestions of our supervisors, considering objective factors such as time, we worked together to develop the main structure and timeline of our project, and we basically completed our project as planned. Our main time is invested in the third stage, which is the modeling task part and also the main body of our project. A sensible division of labor kept our project moving smoothly. We have also systematically analyzed our model and the project as a whole from various perspectives and have finally identified possible future directions. Our truncated SVD model, as well as the clustering model, can be used as part of the preprocessing for subsequent work, such as a reference for the FAQs setting, providing a reference for the labels of the classification model, etc. Finally we want to express our appreciate to our supervisors, Simon Fuchs and Omar Shouman, for their help!