Data Science Project Competition Automated Customer Email Classification

Miha Likar and Neža Marija Slosar

Abstract

Often emails must first be assigned to a category and sometimes even subcategories before we can determine who should respond and how. Our task is to make such a classification into (pre)determined categories. First we did some basic preprocessing and tokenization. Second we adjusted those features for Slovene language. Third we visualize data and analyzed the trend. Fourth we tried some models and found out that logistic regression works best. We finished with clustering and optimising categories to best fit in deep learning models. At the final step we have narrowed the number of categories to 9 and reached highest accuracy: 0.84 and F1 (macro avg.) 0.69. In the future use of BERT or RoBERTa (for Slovene language) would be good way to continue.

Keywords

Email classification, Natural language processing, Slovene, Supervised classification, Deep learning models

Advisors: doc. dr. Slavko Žitnik, Andrej Miščič

Introduction

Email classification is one of the first step in good customer support, since it reduces the responding time and offers employees to concentrate on some perhaps more relevant task instead of manual email classification. [1][2]

Email classification is based on artificial intelligence, combining linguistics and machine learning. Based on huge datasets, which are cumulating on the internet, new sufficient approaches in natural language processing (NLP) are developed. Still, it is important to mention the aspect of language itself. We are working on Slovene emails so we should take all the differences from English into account. On that point we can help with CLARIN Slovenia - Common Language Resources and Technology Infrastructure, Slovenia, which is pioneer on Computer and Corpus Linguistics, and Digital Humanities. Researchers, working on that project had also developed CLASSLA, pipline from Python NLP library Stanza.

Furthermore email classification is not a new thing, it began in clustering *ham* and *spam* emails. [3] Nowadays lot of different learning models are widely used: BERT (Bidirectional Encoder Representations from Transformers); DNN (deep neural network) with BiLSTM (bidirectional Long Short Term Memory) layer; classifier k-NN (k-nearest neihbours) and NB (Naive Bayes). [3]

Based on all these knowledge, models and NLP Course from our supervisor doc. dr. Slavko Žitnik, our aim is to develop autonomous email classifier and make faster and more efficient customer support.

Methods

We got data (emails) separated by months (from January 2021 till June 2021). Firstly we merge them together into one data frame. Then we cleaned not labeled categories and we got all together 35.698 emails, separated in 15 categories. The designated categories were: Asistenca, Asistenca brez primera, Banke, CORIS, Drajv, GDPR, Help desk - interna javnost, Marketinške akcije, Odjava B2B, PDPZ, Premoženje, Zavarovanja - zunanji partnerji, Zdravje, drugo, Življenje.

We started with a general preprocessing procedure, which included lowercasing, punctuation removal, digit removal, email adresses removal and web links removal. These processes were made by applying regex functions. Secondly we have also applied stopwords removal based on Slovene library. After the first analysis we have found out that some uninformative words remained in the data, so we made list of additional uninformative words based on this primary analysis. In the second round of preprocessing we have also used CLASSLA, mainly to achieve lemmatisation. We used KMeans for the whole process of clustering which involved many different

trials from sklearn library and worked on different visualisations with seaborn library. The original number of designated categories in was 15, so we have decided to test a range of clustering the data provided from 20 to 8 clusters.

We decided to visualize data in tables, charts and confusion matrix in all phases:

1. after the basic pre-processing,
2. after CLASSLA,
3. at the end (with joint categories).

Since the vocabulary in the emails is not that rich, a lot of informative words were recognized as different words and this had a negative impact on the models. Lemmas expectedly turned out to be much more efficient in both the model training and testing, as in the case of clustering as well. Thirdly, we count out how many emails are in each category. We found out, that "Marketinške akcije" and "GDPR" have under 55 emails, "Banke" and "CORIS" under 20 and "Zavarovanja - zunanji partnerji", "Help desk - interna javnost" under 10 emails. Based on different trials of clustering, we have found that all of these categories contain highly non-unique and thus uninformative vocabulary which was, based on additional text analysis, practically impossible to categorize with our knowledge and model. Lastly we visualized the data and put it into the final model.

After that we did clustering and consequential categorization we ran few models:

- dummy classifier,
- naive Bayes,
- linear support vector classifier and
- logistic regression.

Results

From preprocessing we find out, which words have potential to be key words of specific category, we visualize those data to get insight, if our preprocessing makes sense and provides useful information to further work with them on models. After using CLASSLA for additional processing we got far better results of word frequencies in lemmatisation. Afterwards we made clustering process using the lemmas and got some interesting results regarding data.

About data

We compared data from frequency tables, charts and clustering. This comparison gave us additional information, what else put in stop words and what is the general representation of informative words across the categories. We found out that words: "imeti", "podatek" and "vožnja" are very common and present in all categories. On the other side we found that feature words for each category are: "vozilo" (Asistenca brez primera), "stranka" (Asistenca), "drajv" (Drajv), "oseben" (Drugo), "seminar" (Odjava B2B), "pokojninski" (PDPZ), "zavarovalen" (Premoženje), "zdravstven" (Zdravje), "stečajen" (Življenje).

Initial stages of clustering turned out to be quite uninformative but as we have performed additional preprocessing and systematically changed the parameters, we have started to form some clustering which resembled some of our best-classified categories and excluded some of the smaller, and some of the categories with little of non-representative samples which guided us which categories to combine. One such example is shown in Figure 1. For example "gebrüder" or "impresum", which alert us and we paid attention again to stop words and informative words for each category to provide best preprocessed data. At the end, we have figured that we went through some cycles of cleaning and clustering since beside frequency tables it provided with some useful additional insight into the data.

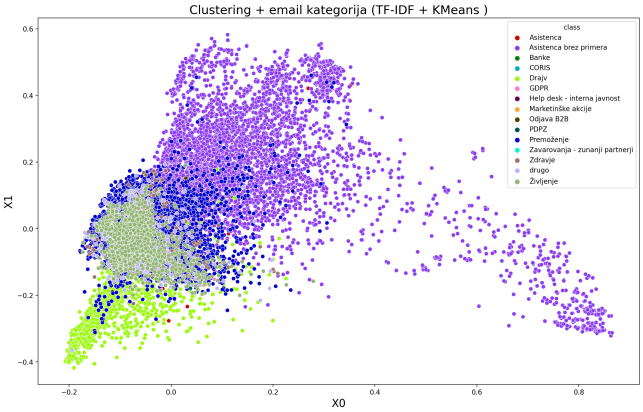

Figure 1. Clustering. Visualization of clustering merged with categories.

Based on the clustering and models we have concluded that the categories: "Marketinške akcije", "GDPR", "Banke", "CORIS", "Zavarovanja - zunanji partnerji" and "Help desk - interna javnost" does not have enough data to provide efficient classification accuracy. Because of that we merged together all those categories in category "Zdruzeno" first, which provide better accuracies, and in the last stage we decided to merge this category with "Drugo", because we agreed upon both categories either have not enough samples or are not classified. In the last stage we had 9 categories and best macro average F1-score.

Test scores from dummy classifier are:

1. most frequent: 0.43,
2. stratified: 0.28,
3. uniform: 0.067,
4. constant: 0.43.

We are showing also confusion matrix for logistic regression – Figure 2, which was the best model.

Tables

Clustering process provide those clusters with useful inside in top keywords for some categories. We got 15 clusters.

We made example of frequency table, which is first inside into category understanding and visualizing of key words.

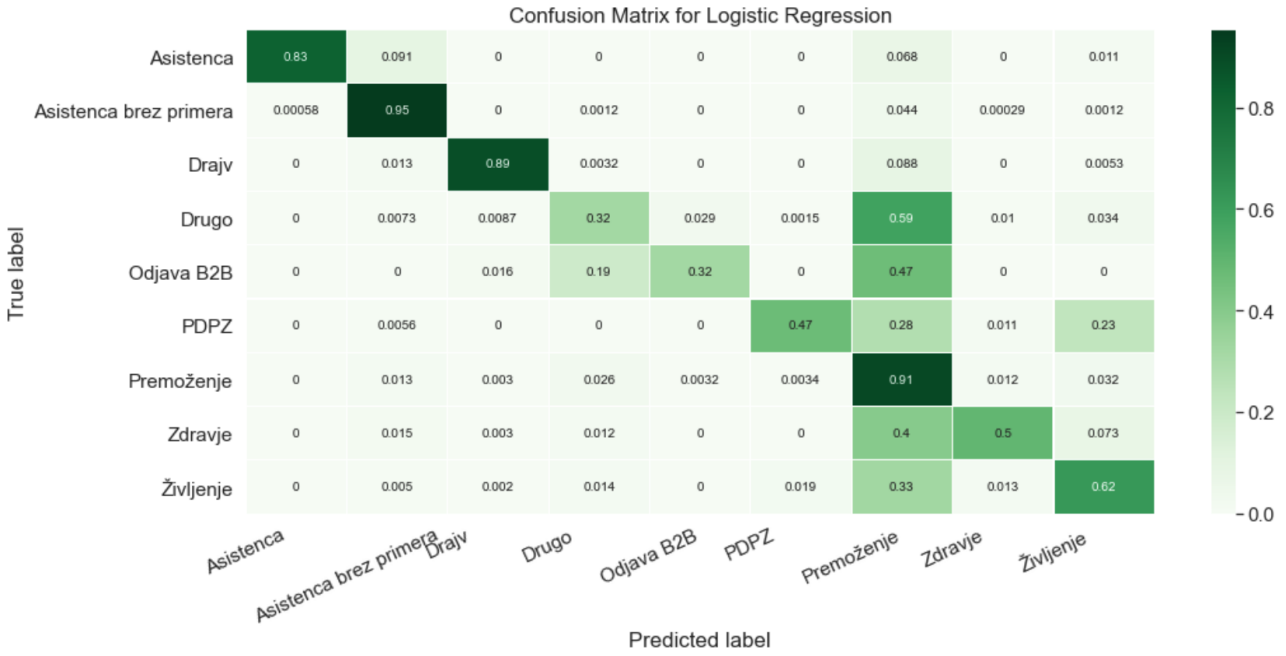

Figure 2. Confusion matrix. In the final stage logistic regression produce this result.

Second we are reporting accuracies from models. Naive Bayes provide:

Linear support vector classifier report slightly better results compare to naive Bayes.

Logistic regression model perform best from all models, with accuracy of 0.84. So, we conclude to work on with this model.

Discussion

We have decided that the beginning of our first dive into NLP should be data understanding and visualization. At the very beginning we recognized the first obstacle, which is not enough data for some categories and heavier challenge in NLP of Slovene language.

We have pretty much finished with preprocessing, regarding tokenizing, lemmatizing, cleaning and clustering data. Second we did few meaningful visualizations. Third we tried few basic learning models, to see if preprocess works well and what should we take into consideration for other models.

Based on learning outcomes we worked on two directions for next few months. First, we prepared reasonable categories and did on data as much as we can with CLASSLA, regarding clustering and other features. Second, we tried to use apart logistic regression some other, according to literature more powerful models – we started to study BERT or RoBERTa 2.0 (which is BERT for Slovene language), but we stopped because we did not have enough processing power on our computers to work on with it.

On the other hand our main goal was to develop a code, which does good pre-processing from the point email step into the system, and second connect this to classification model, which will classify email (based on a bit different categories from now) with at least 0.85 classification accuracy and 0.7 macro avg. F1 score.

We managed to improve our classification models from the very beginning (0.42) for 60% (0.7) till the end of project. Even tough our last accuracy at logistic regression was 0.69 we still think based on clustering and cleaning data, that this result is good and not over fitted to data we used. We did not manage to try among other models also BERT, but according to results our model provide, we believe that could moderately perform as a company classifier.

If we consider results from the Figure 2 in the context of potential contribution to the whole classification method. Based on our raw data, there are approximately 5500 mails classified manually on average each month. If we apply results of our model to different categories, not all of them would be classified equally successfully and as a result the number of correctly classified emails differs among the categories. Categories of Asistenca brez Primera, Drajv and Premoženje are all approximately 90 % accurately classified - since these are by far the biggest categories, they serve as the most informative ones, and consequentially the highest amount of emails would be successfully classified in a automated way. These categories combined contribute to almost 4000 out of 5500 successfully classified emails (Asistenca brez primera: cca. 2200, Premoženje: cca. 1300, Drajv: 400). Classification accuracy for other, much smaller categories is between 30 % and 60 % meaning that out of remaining 1500 emails, half of them on average would still be classified in a successful way, resulting in total of 4750 emails and leaving 750 emails classified incorrectly. Speaking in practical manner, the department with

Table 1. Frequency table for Drajv category.

Word	Frequency
drajv	4676
vožnja	3785
aplikacija	3523
cdrajv	2948
eglasnik	1747
popust	1510
koda	1272
uporaba	1034
mesec	878
telefon	814
varen	672
pripis	630
pogoj	599
uporabnik	582
obvestiti	563
pravilno	549
napaka	543
elektronski	540
napisati	526
pripisati	518

Table 2. Naive Bayes results.

	precision	recall	F1-score
accuracy			**0.73**
macro avg	0.55	0.33	0.34
weighted avg	0.73	0.73	0.65

the highest amount of wrongly classified emails in their inbox would be the department covering category "Premoženje", which is an anticipated consequence based on our text analysis – result of highly frequent non-unique words. They would receive approximately 450 emails which would need to be manually resent to another, correct department. The result of our model would therefore contribute to a high number of correctly classified emails, but would leave members of some departments with potentially a larger number of incorrectly classified emails in their inboxes, which would consume their time to manually re-classify or resend the emails at this point.

Table 4. Logistic regression results.

	precision	recall	F1-score
accuracy			**0.84**
macro avg	0.76	0.65	0.69
weighted avg	0.83	0.84	0.83

Acknowledgments

We would like to thank our supervisors – Slavko Žitnik and Andrej Miščič, who are giving us support and help us with all NLP understanding. We would also like to thank Jure Demšar and his team for marvelous work on organisation and last but not least, we are thankful for support from Zavarovalnica Triglav, where we can always get useful information, regarding data and concepts.

Table 3. Linear support vector classifier results.

	precision	recall	F1-score
accuracy			**0.77**
macro avg	0.62	0.43	0.45
weighted avg	0.74	0.77	0.71

References

[1] Michael Wenceslaus Putong and Suharjito Suharjito. Classification model of contact center customers emails using machine learning. *Advances in Science, Technology and Engineering Systems Journal*, 5(1):174–182, 2020.

[2] Anton Borg, Martin Boldt, Oliver Rosander, and Jim Ahlstrand. *E-mail classification with machine learning and word embeddings for improved customer support*. Blekinge Institute of Technology, Master's thesis, 2017.

[3] Isra AbdulNabi and Qussai Yaseen. Spam email detection using deep learning techniques. *Procedia Computer Science*, 184:853–858, 1 2021.