ITSM Automation - Using Machine Learning to Predict Incident Resolution Category

Jorge COSTA
Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal, jtmca1@iscte-iul.pt

Rubén PEREIRA
Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal, ruben.filipe.pereira@iscte-iul.pt

Ricardo RIBEIRO
Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal,
INESC-ID, Lisbon, Portugal, ricardo.ribeiro@inesc-id.pt

Abstract

Problem resolution is a key issue in the IT service industry, and it is still difficult for large enterprises to guarantee the service quality of the Incident Management (IM) process because of the difficulty in handling frequent incidents timely, even though IT Service Management (ITSM) standard process have already been established (Zhao & Yang, 2013). In this work, we propose an approach to predict the incident solution category, by exploring and combining the application of natural language processing techniques and machine learning algorithms on a real dataset from a large organization. The tickets contain information across a vast range of subjects from inside the organization with a vocabulary specific to these subjects. By exploring the text-based attributes, our findings show that the full description of an incident is better than the short description and after stop words removal, the use of additional preprocessing techniques and the addition of tickets nominal attributes such as have no impact to the classification performance.

Keywords: ITSM, Incident management, Natural Language, Machine Learning

Introduction

Currently, organizations spend lots of resources to keep their IT resources incident free and running since there are very few areas that do not depend directly or indirectly on software systems. Therefore, they rely on IT Service Management (ITSM) processes to quickly detect, process and resolve incoming incidents to achieve that goal. Conventionally, the Incident Management (IM) process is largely manual, error prone and time consuming, especially the resolution step (Gupta, Prasad, & Mohania, 2008) and in many cases this process is not entirely systematic and may be incoherent and inefficient (Salah, Maciá-Fernández, Díaz-Verdejo, & Sánchez-Casado, 2016).

Normally, after an incident is solved the service agent records the information about the resolution steps applied such as resolution category, closed code and the resolution notes and description. The desire is to move to an automatic system that autonomously proposes a resolution to new incoming tickets saving time and resources by reallocating the man power to processes.

This research is the first step in the attempt to automate the incident resolution process, by applying machine learning to classify the resolution category of the incident, in order to further used it in the autonomous resolution by narrowing down the possible solution range. The incident tickets contain different attributes in structured and unstructured data, with the latter in the form of free text such as the incident short and full description. The dataset used to test and evaluate the algorithms is composed by samples from real-word incident tickets provided by a company whose name, by request, cannot be mention.

Background

ITSM is a subset of Service Science that focuses on IT operations such as service delivery and service support (Kang, Zaslavsky, Krishnaswamy, & Bartolini, 2010). It refers to the entirety of activities: directed by policies, organized and structured in processes and supporting procedures – that are performed by an organization to design, plan, deliver, operate and control IT services offered to customers and meets the needs of the business. It must be carried out effectively and efficiently and from the business perspective enables organizational high performance and value creation (Kosasi, Prabowo, & Budiastuti, 2017).

Referring to ITIL, service supports consist of IM, problem management, change management, exemption management, and configuration management. Service deliveries, on the other hand, include management service levels, finance management, IT continuance management, capacity management, and availability management (Wan & Chan, 2007).

The IM process is responsible for managing the lifecycle of all incidents, including any event which disrupts, or which could disrupt, a service. This includes events which are communicated directly by users, either through the Service Desk or through an interface from Event Management to IM tools (Marcu et al., 2009). An incident is defined by the ITIL as an unplanned interruption to an IT service or a reduction in the quality of an IT service and are reported by humans or automatically detected and generated by a monitoring system (IBM Tivoli Enterprise Console). As a key process, IM provides data record of each step-in incident resolution process, verifies resource configuration, management process and its operation quality to achieve service objectives, and provides data for developing service report, service plan, cost accounting as well as service workload assessment (OGC, 2007).

Related Work

In our research we did not find attempts to classify the tickets resolution category, but over the years there has been some approaches to the automatic problem category determination. In this section we describe some of the work developed, used algorithm and the obtained results of each implementations.

Silva, Ribeiro, and Pereira (2018) introduce automatic text classification to ticket problem, by applying NLP pre-processing techniques to the ticket's text data and term frequency–inverse document frequency (TF-IDF) based document that then feeds the Support Vector Machine (SVM) and k-nearest neighbors (KNN) algorithms. They achieved better accuracy with SVM with values up to 80% and concluded that the short description is better than the full description and the pre-processing had little impact on the categorization, mostly due the specific vocabulary used in the text. Agarwal, Sindhgatta, and Sengupta (2012) created a system focused on automating incident classification from the ticket description. They used weighted vectors of terms to build an SVM classifier achieving a performance between 69% and 89% accuracy in the IBM dataset that was used for training and testing. Altintas, Cuneyd Tantug, Tr, and Tr (2014) proposed an extension to ITS for auto-addressing the issue ticket to the relevant person or unit from the ticket category in Turkish ticket data. They used the bag of word approach with machine learning algorithms. They carried out a morphological analysis step to decrease the high feature vector size and also tested with TF-IDF. For classification they used SVM, KNN, Decision Tree and Naïve Bayes algorithms. They achieved an accuracy between 75% and 85% depending on the training data set. Zinner et al (2015), investigated the performance of machine learning algorithms in identifying relevant support tickets from a 12000 manually labeled tickets dataset from a German company. For preprocessing they used bag-of-words representation, tokenization and stop words removal and TF-IDF was also used. To categorize the tickets, the authors explored with SVM, KNN and Naïve Bayes and SVM outperformed the other two classifiers achieving up to 72% F1-score. In their experiment they were also more successful at detecting major issues at specific sites than categorizing each individual ticket accurately.

Proposed method

Our proposed method is based on the works introduced in the previous section. The most important attribute of an incoming ticket is the incident description, which is present in the short and full description fields in the form of unstructured natural language text. They are both provided by the person creating the incident ticket and by being in the form of text, an unstructured form of data that classifiers algorithms cannot directly process (Feldman; & Sanger, 2007) they must be converted into a more manageable form, during a preprocessing step.

In this study, we analyze the impact of some most common natural language pre-processing techniques (Aggarwal & Zhai, 2012), such as: tokenization, stop word and digits removal, word stemming and part-of-speech filtering by selecting only open grammatical classes.

We use TF-IDF, a vector space-based representation for the text data that are feed to the machine learning algorithms. It also serves as a feature selection method by selecting terms based on their weights (Patil & Atique, 2013).

For classification we explored key methods, which are commonly used for text classification (Aggarwal & Zhai, 2012), more specifically: Multinomial Naive Bayes SVM, KNN, Decision Tree and Logistic Regression. The goal is to compare the results from each approach and choose the one with the best performance.

Dataset

As previously stated, the dataset was used was provided by the IT department from a big multinational company. An example of an incident ticket can be observed in Table I.

The dataset contains 472,929 tickets in different languages such as English, Portuguese, German, Spanish, French, etc. For this study, we will only be using the data in English since it is the most used, representing more than 90% of the dataset. There are also other attributes in the tickets in the form of structured fields, such as the incident problem category, subcategory, severity, user department, location and the assignment group for the ticket. These attributes may have an impact and contribute to the classification, but the descriptions are the critical ones to obtain a good performance (Silva, Ribeiro, & Perreira, 2018). In

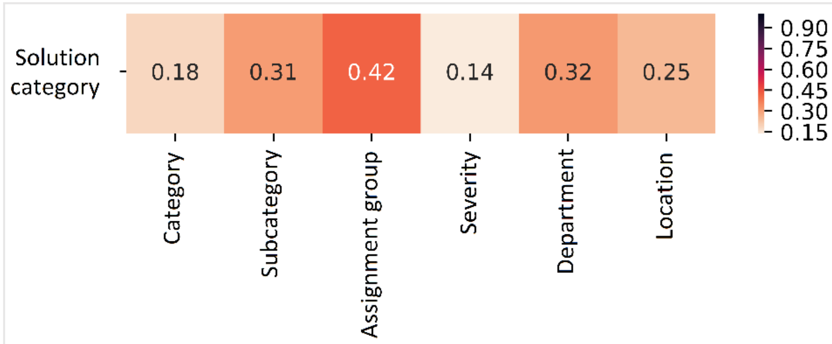

Figure ***I***, we show the correlation, using Cramér's V (Acock & Stavig, 1979), between the structured nominal attributes and the solution category. Based on the Chi square, Cramér's V is a measure of the strength of the relationship or association between two nominal variables. When calculated, compared to Chi square, it considers the dimensions of the attributes table, meaning that attributes with different dimensions can be meaningfully compared.

Table I- Incident ticket example

Short Description	Windows Password Reset
Full Description	User U needs to reset laptop password…
Incident Category	Security and Access
Incident Subcategory	Identity and Account Management
Affected Location	Location X
Severity	3 - Medium
User Department	Department Y
Assignment group	Group Z

The solution category is one of the following ten different categories: Information/Advice given, Request, Configuration, Security, Installation, Data, Other, Software, Hardware and Complaint. As it is possible to observe in

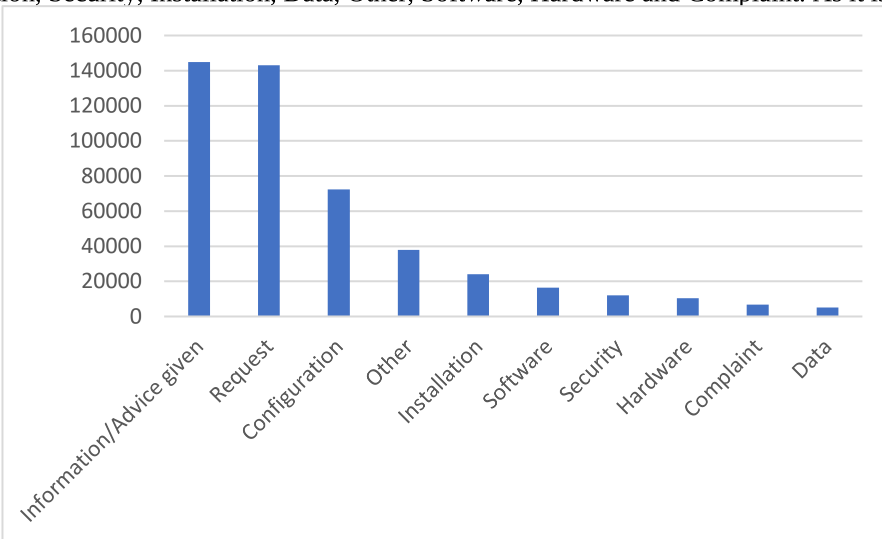

Figure ***II***, the dataset is highly unbalanced having two categories representing 60% of the data. This is usually a problem for classification tasks, but since we have of the large size of the dataset, we can extract a balanced sample to be used.

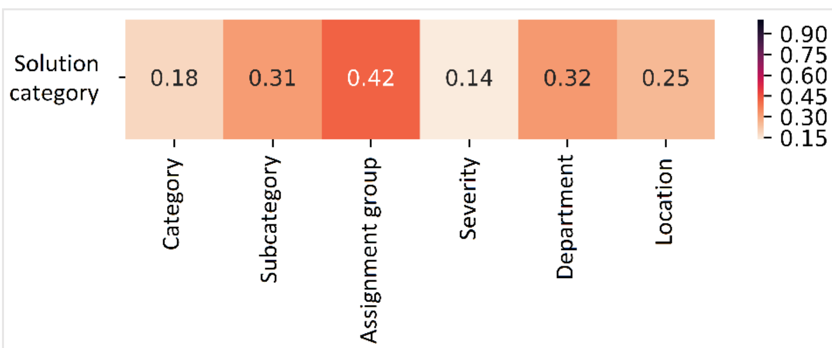

Figure I - Cramer V Correlation between Variables

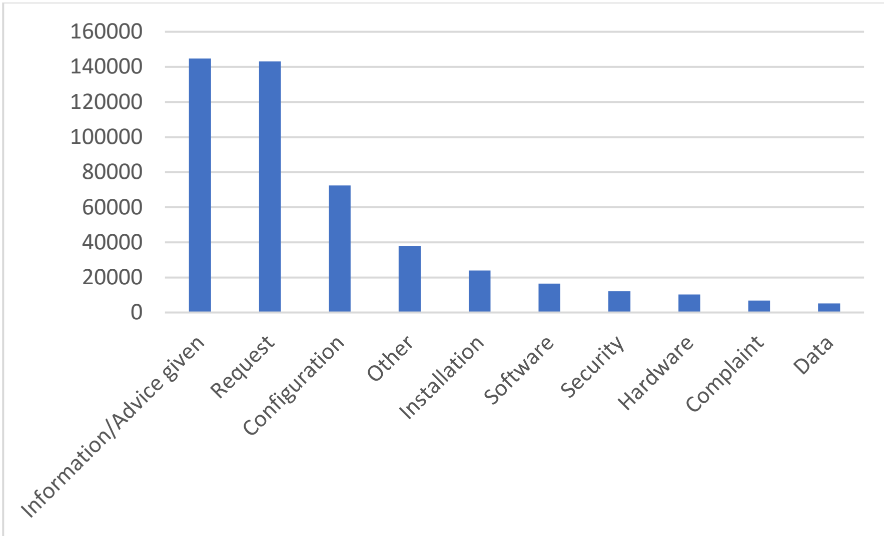

Figure II - Solution category distribution

The short description and full description are the most important feature in the tickets. Basically, the full description contains more detailed information about the incident whereas the short description is more of a one phrase short summary about the incident details. Table II shows the minimum, maximum and average word count for each solution category. All categories have presented a ticket with only one word in one of the descriptions that by our intuition is due to the unwillingness of the person that opened the ticket. Concerning large maximum word count for the full description, a couple of tickets had whole emails conversations or other forms of external reports that was pasted to the field. Regardless of those, the average for each description has a balanced distribution throughout the categories.

Table II - Word count per solution category for short and full description

	Short description			Full description		
	Minimum	Maximum	Average	Minimum	Maximum	Average
Information/Advice given	1	34	6.21	1	4801	46.47
Request	1	34	6.24	1	1676	40.62
Configuration	1	32	6.62	1	4173	43.74
Other	1	29	5.61	1	1639	46.11
Installation	1	32	5.94	1	759	46.45
Software	1	31	6.87	1	764	53.72
Security	1	32	5.61	1	2095	40.1
Hardware	1	32	6.38	1	857	53.31
Complaint	1	32	7.24	1	1737	54.15
Data	1	28	7.71	1	1832	38.23

Implementation

To train the algorithms we selected 5,000 tickets from each solution category and tested different pre-processing techniques previously specified and their combinations for better performance.

We analyze the impact of using the short and full description on the classification and also the impact of the nominal attributes for the performance, more specifically the incident subcategory and the assignment group since these are the ones that presented a stronger correlation with the solution category as it was showed in

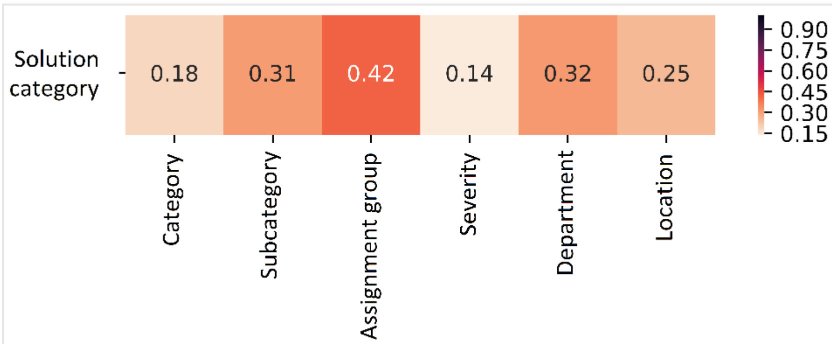

Figure ***1***.

For the execution of the different approaches, we used cross-validation to divide the dataset in subsets of 80% of the dataset for training and 20% for testing aimed at assessing the models results. The goal of cross-validation is to test a model's ability to predict new data that was not used in the training process, in order to flag problems like overfitting or selection bias (Cawley & Talbot, 2010).

The obtained results are presented in the results section.

Evaluation metrics

To assess the classification performance two evaluation metrics were used: accuracy and F1-Score which is the harmonic average of the precision and recall. The definition of the metrics is as follows:

- True Positive (TP) means that a solution category has been correctly identified;
- False Positive (FP) means that a ticket that does not belong to a solution category, was incorrectly identified to that category;
- True Negative (TN) means that a ticket that does not belong to a solution category, was incorrectly identified as not belonging that category;
- False Negative (FN) means that a ticket belonging to a solution category, was incorrectly identified as not belonging to that category;

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Results

Figure III presents the accuracy and f1-score results after testing the classification algorithms with only stop words and numbers removal for the short and full description. As it is possible to see to contrary to the findings of Silva, Ribeiro, and Perreira (2018) for incident classification, the full description presented better results, achieving an accuracy of 51% with SVM, whereas when using the short description the accuracy decreases to 43%. Overall in the training process, because of its bigger size the full description resulted in a larger vector dimension representation which even though achieved better results, also took a longer time to train.

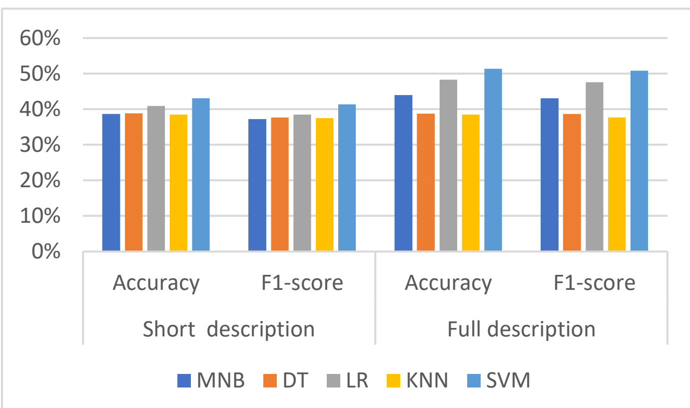

Figure III - Accuracy and F1-score results for Short vs full description

Regarding additional pre-processing techniques, we explored the use of stemming, that consists in reducing terms in a document to their basic forms or stems (Kotu & Deshpande, 2015) that results in a smaller dictionary size. More specifically, we used the Porter Stemmer (Porter, 1980) and compared the results to not using stemming at all. The use of stemming did not have any impact to the obtained results, presenting variations of less than 1%. Another aspect that we explored was the part-of-speech filtering. By tagging each word with its respective grammatical category and filtering and selecting only the words that belong to the open grammatical class. With this we sought to reduce the noise in the text data helping in the feature selection. Again, the variation in the performance was minimal, and considering increase in training time we conclude that part-of-speech brought no benefits.

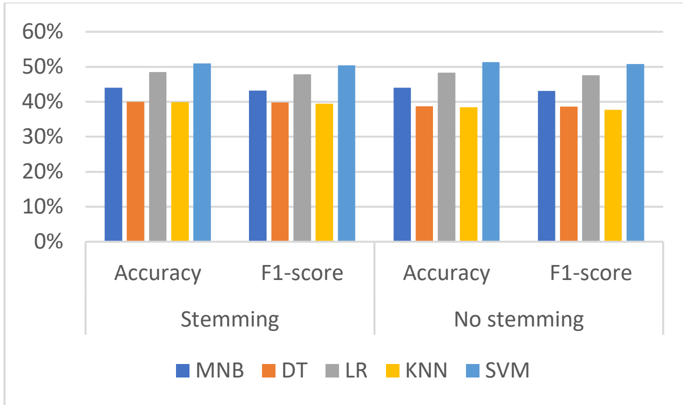

Figure IV - Accuracy and F1-score results for Stemming vs No Stemming

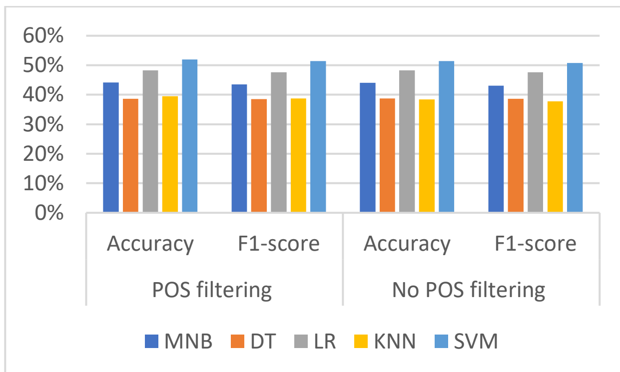

Figure V - Accuracy and F1-score results for Stemming vs No Stemming

As previously indicated, the incident tickets contain structured nominal attributes related to the incident. Based on the correlation observed in

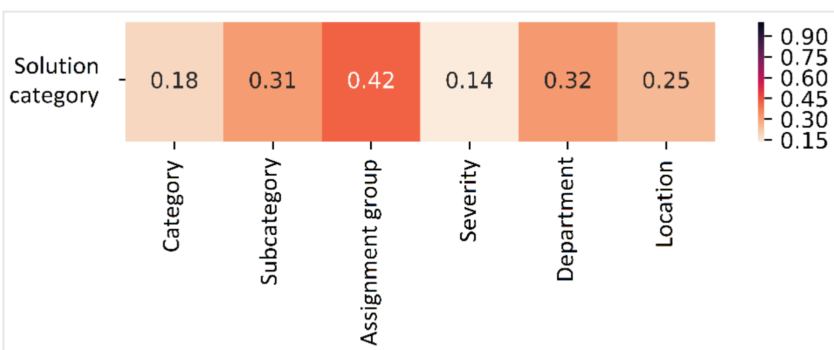

Figure ***I*** we explored the Department, Assignment group and the ticket subcategory for the training since they presented a stronger correlation with the solution category.

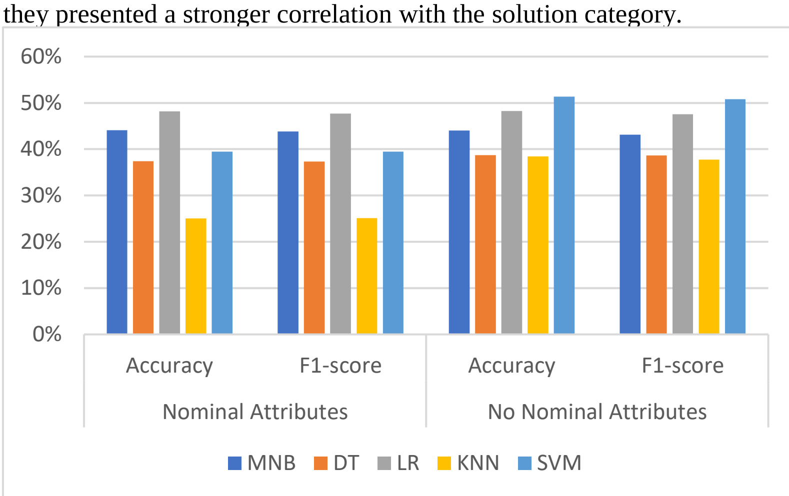

Figure ***VI*** presents the obtained results and for our surprise, this process did not improve the results at all and actually decreased the accuracy and f1-score for SVM and KNN, achieving the highest accuracy at 48% with Logistic Regression.

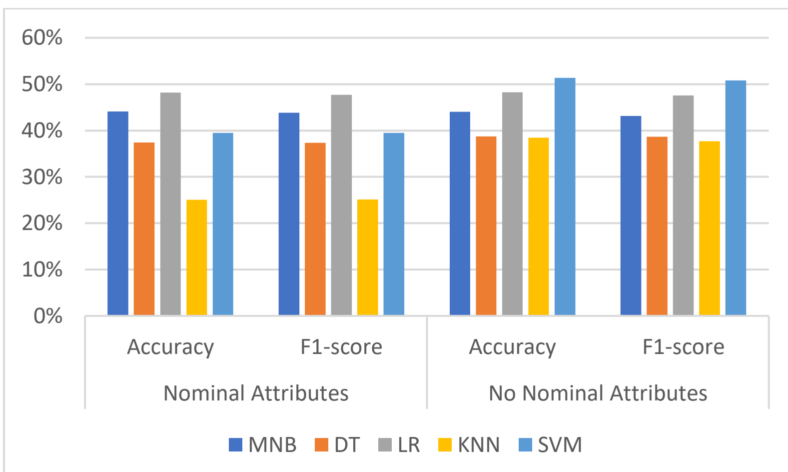

Figure VI - Accuracy and F1-score results with nominal attributes vs without nominal attributes

Conclusions and future Work

In this article, we present an approach to predict the solution category of an incident ticket using machine learning. We analyze different natural language processing techniques such as stop words removal, stemming and part of speech filtering, and then evaluate their impact on the classification of real incident tickets dataset.

The best results were achieved by SVM with TF-IDF vectors as input and stop words removal reaching 51% accuracy followed by Logistic Regression at 48% and KNN in the last place at 38%. Considering that 60% percent of the dataset tickets belong to two of the ten categories, the accuracy results are not much better than chance or optimal to be applied in to the real-world scenario.

We concluded that the use of the full description produces better results when compared to the short description. A possible justification might be the fact that the full description contains a more detailed explanation about the incident. It was also observed that the used of additional processing techniques and the structured nominal attributes from the ticket present no improvements to the results.

In a large organization thousands of man hours are consumed every year analyzing and routing tickets to its final solver. This work is part of a bigger effort to automate the incident management resolution process: by correctly predicting the solution category for a new incident it is possible to improve the entire tickets routing process, reducing the incident life and saving organizations resources.

For the future, we plan to extend this study to a system that automatically suggest a possible resolution to an incoming incident ticket using knowledge obtained from a historic dataset.

Acknowledgment

This work is financed by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., under the project UID/Multi/04466/2019.

References

Acock, A. C. and Stavig, G. R. (1979) ‘A Measure of Association for Nonparametric Statistics’, *Social Forces*, 57(4), p. 1381. doi: 10.2307/2577276.

Agarwal, S., Sindhgatta, R. and Sengupta, B. (2012) ‘SmartDispatch’, in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12*. New York, New York, USA: ACM Press, p. 1393. doi: 10.1145/2339530.2339744.

Aggarwal, C. C. and Zhai, C. (2012) ‘A Survey of Text Classification Algorithms’, in *Mining Text Data*. Boston, MA: Springer US, pp. 163–222. doi: 10.1007/978-1-4614-3223-4_6.

Altintas, M. *et al.* (2014) *Machine Learning Based Ticket Classification In Issue Tracking Systems*. Available at: http://worldconferences.net (Accessed: 17 February 2019).

Cawley, G. C. and Talbot, N. L. C. (2010) ‘On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation’, *Journal of Machine Learning Research*, 11(Jul), pp. 2079–2107. Available at: http://www.jmlr.org/papers/v11/cawley10a.html (Accessed: 17 February 2019).

Feldman, R. and Sanger, J. (2007) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, *Saudi Med J*. doi: 10.1179/1465312512Z.00000000017.

Gupta, R., Prasad, K. H. and Mohania, M. (2008) 'Automating ITSM incident management process', *5th International Conference on Autonomic Computing, ICAC 2008*, 1, pp. 141–150. doi: 10.1109/ICAC.2008.22.

Kang, Y.-B. *et al.* (2010) 'A knowledge-rich similarity measure for improving IT incident resolution process', in *Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10*. New York, New York, USA: ACM Press, p. 1781. doi: 10.1145/1774088.1774466.

Kosasi, S., Prabowo, H. and Budiastuti, D. (2017) 'Improving information performance of schools in higher education through IT service management', *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*. doi: 10.1109/CITSM.2017.8089246.

Kotu, V. and Deshpande, B. (2015) 'Text Mining', in *Predictive Analytics and Data Mining*. Elsevier, pp. 275–303. doi: 10.1016/B978-0-12-801460-8.00009-4.

Marcu, P. *et al.* (2009) 'Towards an optimized model of incident ticket correlation', *2009 IFIP/IEEE International Symposium on Integrated Network Management, IM 2009*, pp. 569–576. doi: 10.1109/INM.2009.5188863.

OGC (2007) *The Official Introduction to the ITIL Service Lifecycle*, *Office of Government Commerce*. doi: citeulike-article-id:2420971.

Patil, L. H. and Atique, M. (2013) 'A novel approach for feature selection method TF-IDF in document clustering', in *2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, pp. 858–862. doi: 10.1109/IAdCC.2013.6514339.

Porter, M. F. (1980) 'An algorithm for suffix stripping', *Program*. doi: 10.1108/eb046814.

Salah, S. *et al.* (2016) 'A Model for Incident Tickets Correlation in Network Management', *Journal of Network and Systems Management*, 24(1), pp. 57–91. doi: 10.1007/s10922-014-9340-6.

Silva, S., Ribeiro, R. and Pereira, R. (2018) 'Less is more in incident categorization (Short Paper)', *DROPS-IDN/9275*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik. doi: 10.4230/OASICS.SLATE.2018.17.

Silva, S., Ribeiro, R. and Perreira, R. (2018) 'Less is more in incident categorization', *Scientific American*, 285(3), pp. 66–73. doi: 10.4230/OASIcs.SLATE.2018.17.

Wan, S. H. C. and Chan, Y. H. (2007) 'IT Service Management for campus environment -practical concerns in implementation', in *10th IFIP/IEEE International Symposium on Integrated Network Management 2007, IM '07*. doi: 10.1109/INM.2007.374833.

Zhao, G. and Yang, S. (2013) 'IT service incident management model decision based on ELECTRE III', in *Proceedings of 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2013*. doi: 10.1109/ICIII.2013.6702987.

Zinner, T. *et al.* (2015) 'Text Categorization for Deriving the Application Quality in Enterprises Using Ticketing Systems', in. Springer, Cham, pp. 325–336. doi: 10.1007/978-3-319-22729-0_25.