MINISTERUL EDUCAŢIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare, Informatică şi Microelectronică

Departamentul Ingineria Software și Automatică

Admis la susţinere

Şef departament:

FIODOROV Ion dr., conf.univ.

„___” _____ 2024

CLASIFICAREA AUTOMATIZATĂ A TICHETELOR FOLOSIND ALGORITMI DE PROCESARE A LIMBAJULUI NATURAL ȘI METODE DE ÎNVĂȚARE PROFUNDĂ

Proiect de master

Autor: _____ Țurcan Cătălin

Coordonator: _____ Beşliu Corina, dr.

Consultant: _____ Catruc Mariana, lect. univ.

ABSTRACT

In short, this paper discusses the real world advantages of implementing an automated ticket classification. Before reaching the final solution, the required data must be transformed from its original state, in one better suited for the machine learning model. Along the way, multiple challenges were found. Further down, it will be discussed how these problems were identified, what is their impact on the performance of the model, and how the problems can be fixed. Such problems include the presence of multiple languages. There is a lot of text, all in different languages and they are also very unbalanced. The common approaches are either balancing them using oversampling and undersampling methods, or discarding the uncommon languages altogether. Another problem is the unbalanced classes inside the data. The presence of a hierarchical structure for the labels is not a problem in itself, but it can be challenging to correctly implement a model which takes advantage of this extra data. The scope of this paper is to look into different encoding and classification techniques and observe if the recent advancements in pre-trained models give an obvious advantage in the ticket classification problem. For this purpose the traditional encoding and classification methods like naive bayes, linear regression and decision tree are compared to using a large pre-trained model, in this case Distil-Bert-Multilingual. The observed challenges in this specific case is identifying the performance metrics for all these methods against a dataset that has a high degree of class imbalance as well as the presence of multiple languages in the dataset. The approach is simple and consists of three parts. The first part is exploring the data and observing all the particularities of the dataset. In this case it was observed that there are a lot of classes present for classification and there are a lot of distinct languages like German, Italian, English, Maltese, etc. The next part is cleaning the data, balancing the classes using different techniques like: SMOTE, random oversampling, random undersampling, and making the choice of a model, which can perform well based on the particularities found in the first step. In this case DistilBert-Multilingual was chosen, for its encoding support of multiple languages as well as being lightweight, therefore easier computationally. The final step is obtaining and interpreting the results, which in this case, show that using a large, easily obtainable, open-source model yields better results. In short, the worst result using a pre-trained model, returned an accuracy of 59%, while the best of the traditional methods yielded 55.3%, with the worst being 37.66% using naive bayes. The contribution of this paper is to emphasize the fact that using a large pre-trained model, capable of understanding general knowledge, significantly improves the accuracy of the system, without a significant increase in computing requirements.

REZUMAT

Pe scurt, această lucrare discută avantajele din lumea reală ale implementării unei clasificări automate a biletelor. Înainte de a ajunge la soluția finală, datele necesare trebuie transformate din starea inițială, într-una mai potrivită pentru modelul de învățare automată. Pe parcurs, au fost găsite multiple provocări. Mai jos, se va discuta cum au fost identificate aceste probleme, care este impactul lor asupra performanței modelului și cum pot fi rezolvate problemele. Astfel de probleme includ prezența mai multor limbi. Există o mulțime de text, toate în limbi diferite și, de asemenea, sunt foarte dezechilibrate. Abordările comune sunt fie echilibrarea lor folosind metode de supraeșantionare și subeșantionare, fie eliminarea totală a limbilor neobișnuite. O altă problemă o reprezintă clasele dezechilibrate din interiorul datelor. Prezența unei structuri ierarhice pentru etichete nu este o problemă în sine, dar poate fi o provocare să implementezi corect un model care profită de aceste date suplimentare. Scopul acestei lucrări este de a analiza diferite tehnici de codificare și clasificare și de a observa dacă progresele recente în modelele pre-antrenate oferă un avantaj evident în problema clasificării biletelor. În acest scop, metodele tradiționale de codificare și clasificare precum bayes naive, regresie liniară și arbore de decizie sunt comparate cu utilizarea unui model mare pre-antrenat, în acest caz Distil-Bert-Multilingual. Provocările observate în acest caz specific este identificarea valorilor de performanță pentru toate aceste metode în raport cu un set de date care are un grad ridicat de dezechilibru de clasă, precum și prezența mai multor limbi în setul de date. Abordarea este simplă și constă din trei părți. Prima parte este explorarea datelor și observarea tuturor particularităților setului de date. În acest caz s-a observat că există o mulțime de clase prezente pentru clasificare și există o mulțime de limbi distincte precum germană, italiană, engleză, malteză etc. Următoarea parte este curățarea datelor, echilibrarea claselor folosind diferite tehnici precum : SMOTE, supraeșantionare aleatoare, subeșantionare aleatoare și alegerea unui model, care poate funcționa bine pe baza particularităților găsite în primul pas. În acest caz, a fost ales DistilBert-Multilingual, pentru suportul său de codificare pentru mai multe limbi, precum și pentru faptul că este ușor, deci mai ușor din punct de vedere computațional. Pasul final este obținerea și interpretarea rezultatelor, care în acest caz arată că utilizarea unui model open-source mare, ușor de obținut, dă rezultate mai bune. Pe scurt, cel mai prost rezultat folosind un model pre-antrenat, a returnat o precizie de 59%, în timp ce cele mai bune metode tradiționale au dat 55,3%, cel mai rău fiind 37,66% folosind bayes naive. Contribuția acestei lucrări este de a sublinia faptul că utilizarea unui model pre instruit mare, capabil să înțeleagă cunoștințele generale, îmbunătățește semnificativ precizia sistemului, fără o creștere semnificativă a cerințelor de calcul.

Table Of Contents

INTRODUCTION

In today's environment, the amount of data is the largest it has ever been, and moreover, it is ever increasing. This is a big challenge in itself, because manually sorting through it is very time consuming, requires a lot of manpower and moreover, the accuracy of the labor is decreasing. Ticket classification is part of the support a business must provide, to maintain a high degree of customer satisfaction. The efficient management of customer satisfaction is very important for a B2C (business to customer) company. One critical aspect of this challenge is to effectively handle the customer tickets and solve them in a timely, reasonable manner. Part of this problem has been solved using machine learning algorithms.

For a ticket/problem to be solved efficiently, you must make sure that first, it has been correctly classified. Once classified correctly, you can be sure that the system will route the ticket to the person that is responsible for that specific problem.

If a ticket is classified slowly, or incorrectly classified, it can reach a person that does not know how to solve that particular problem. This creates further delays and for sure, will inevitably dissatisfy the customer. In the realm of customer service and technical support, the weight of a timely and accurate classified ticket can’t be overstated. There comes a point in any business’s life, when the volume of customers, their data, and their support requirements, becomes so large that it is impossible even for a large team to manage it well enough. This is why an automated ticket classification system is important and why it can help maintain your company's growth. Such a system will introduce advantages like enhanced efficiency, improved response time, optimized resource allocation and an enhanced customer experience.

REFERENCES

[1] J. Brownlee, ‘A Gentle Introduction to Imbalanced Classification’, MachineLearningMastery.com. Accessed: Dec. 10, 2023. [Online]. Available: https://machinelearningmastery.com/what-is-imbalanced-classification/

[2] J. K, ‘SMOTE’, Medium. Accessed: Dec. 11, 2023. [Online]. Available: https://towardsdatascience.com/smote-fdce2f605729

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ‘SMOTE: Synthetic Minority Over-sampling Technique’, jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[4] B. Saji, ‘Language Detection Using Natural Language Processing’, Analytics Vidhya. Accessed: Sep. 15, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/language-detection-using-natural-language-processing/

[5] A. Zangari, M. Marcuzzo, M. Schiavinato, A. Gasparetto, and A. Albarelli, ‘Ticket automation: An insight into current research with applications to multi-level classification scenarios’, *Expert Systems with Applications*, vol. 225, p. 119984, Apr. 2023, doi: 10.1016/j.eswa.2023.119984.

[6] C. Silla and A. Freitas, ‘A survey of hierarchical classification across different application domains’, *Data Mining and Knowledge Discovery*, vol. 22, pp. 31–72, Jan. 2011, doi: 10.1007/s10618-010-0175-9.

[7] ‘Word embeddings in NLP: A Complete Guide’. Accessed: Oct. 17, 2023. [Online]. Available: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

[8] ‘Imbalanced Classes’. Accessed: Oct. 17, 2023. [Online]. Available: https://goldinlocks.github.io/Imbalanced-Classes/

[9] ‘Training, Validation, Test Split for Machine Learning Datasets’. Accessed: Oct. 15, 2023. [Online]. Available: https://encord.com/blog/train-val-test-split/

[10] ‘distilbert-base-multilingual-cased · Hugging Face’. Accessed: Oct. 15, 2023. [Online]. Available: https://huggingface.co/distilbert-base-multilingual-cased

[11] ‘Automatic Ticket Classification Case Study | NLP’. Accessed: Oct. 15, 2023. [Online]. Available: https://kaggle.com/code/abhishek14398/automatic-ticket-classification-case-study-nlp

[12] *GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION,* Tom M. Mitchell, [online] [accessed:11.12.2023], https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

[13]C. Molnar, *Interpretable Machine Learning*. Accessed: Dec. 10, 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[14] S. P. Paramesh, C. Ramya, and K. S. Shreedhara, ‘Classifying the Unstructured IT Service Desk Tickets Using Ensemble of Classifiers’, in *2018 3rd International Conference on Computational Systems*

and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India: IEEE, Dec. 2018, pp. 221–227. doi: 10.1109/CSITSS.2018.8768734.

[15]M. Marcuzzo, A. Zangari, M. Schiavinato, L. Giudice, A. Gasparetto, and A. Albarelli, 'A multi-level approach for hierarchical Ticket Classification', in *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 201–214. Accessed: Jan. 10, 2024. [Online]. Available: https://aclanthology.org/2022.wnut-1.22

[16] A. Revina, K. Buza, and V. G. Meister, 'IT Ticket Classification: The Simpler, the Better', *IEEE Access*, vol. 8, pp. 193380–193395, 2020, doi: 10.1109/ACCESS.2020.3032840.

[17] M. Altıntaş and A. C. Tantuğ, 'Machine Learning Based Ticket Classification In Issue Tracking Systems', in *The 2nd International Conference on Artificial Intelligence and Computer Science (AICS)*, 2014.

[18]'Classification: Accuracy | Machine Learning', Google for Developers. Accessed: Dec. 10, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy

[19] 'Classification: Precision and Recall | Machine Learning', Google for Developers. Accessed: Dec. 10, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall