2020 7th Intl. Conference on Soft Computing & Machine Intelligence

Mining Campus Transfer Request Data

Nkosikhona Dlamini, Senyeki Marebane, Jabulani Makhubela
Department of Computer Science
Tshwane University of Technology
Emalahleni, South Africa
e-mail: dlaminin3@tut.ac.za, marebanesm@tut.ac.za, makhubelajk@tut.ac.za

***Abstract*—Multi-campus system in a university occurs when several campuses are separated geographically within a university. This allows students to have choices in terms of where to pursue their education and also enables the student to have quality comparable education irrespective of location. However, this presents challenges to University's faculty management in terms of processing the transfer of students between campuses. Large amount of data campus transfer requests need to be analyzed manually. Furthermore, such transfer requests do impact on university student enrolment plan and resource allocation including teaching human resources, lecture venues, allocation of residential accommodation amongst others. This study has adopted Machine learning topic modelling to analyze the campus transfer data in order give the faculty management an overview of what is contained in the data, and classification models to automate the process of accepting or rejecting a transfer request. The results show that the manual process of approving campus transfer requests can be automated using Multinomial Na¨ive Bayes and it achieves a 74% AUC performance.**

Keywords-text mining; classification; topic modeling; campus transfare

I. INTRODUCTION

Multi-campus university model offers opportunities for enhanced student choices in terms of location of study, convenience of flexibility and experience, increases access to education amongst other benefits. It further ensures quality comparable education irrespective of location. Nicolson 2004 [1] defines multi-campus system as: "campuses significantly separated by geography which are combined into a single system". Lee and Bowen 1971 [2] affirms that such structures are organized under one management frame-work. Whilst the multi-campus approach can take various contextual forms, it is aimed to achieve multiple objectives and improved coordination [3] . In South Africa universities running on this model are a product of higher education institutions reorganization by the government [4]. Although literature offers variable forms of multi-campus university, we choose the above definition as it embraces the structure of the university under consideration in this study. The structure of this university presenteds an enhanced opportunity for choosing which campus to study at, allows existing students to complete qualifications by switching between campuses by considering relevant factors that will enhance their studies and improve their social experience during studies. The advent of" Fees Must Fall" movement [5], [6] which resulted in fee free higher education has even expanded the palatability of the campus transfer option.

However, these present challenges to university faculty management in terms of processing the transfer of students between campuses. For example, students at one university in South Africa which has major campuses mainly across three provinces, have the liberty to exercise such options within a faculty. Such campus transfers present challenges because university policy requires the processing of the transfers to be handled by faculties, amongst other reasons to ensure proper control and coordination. Students wanting to change campus submit requests manually or by using online forms. The Faculty then convenes a seating of affected managers and heads of departments to process the transfer request forms which are usually accompanied by volumes of supporting documents. Such processes can be mundane and time consuming.

The nature of this process has the potential of being inconsistent as a result of human intervention and inability to predict the outcomes of a decision for either allowing or declining the transfer request. Furthermore, such transfer requests do impact on university's student enrolment plan and resource allocation including teaching human resources, lecture venues, allocation of residential accommodation.

Apart from the potential adverse effect on proper allocation of resources and process, campus transfer has the potential to affect the progress of a transferring student and subsequently the student success rates and graduation rates of the Faculties amongst others. Since there is a wide range of reasons for campus transfer requests observed over the years, there is a need to have general overview of such reasons to develop capacity of expediting decision making. However, these are embedded in heaps of data in the student transfer records accumulated annually.

Searching and identifying such reasons manually is time consuming and can delay decision making. It should be noted that most prevalent works in literature regarding multi-campus university studies are focused on leadership, governance and academic issues, with less attention to the administration. This study derives its relevance from the proposed improvement of processing campus transfer requests by using machine learning algorithms to address the administrative element of a multi-campus university. In this paper, we propose the application of machine learning algorithms to assist with the analysis of huge amount of data

978-1-7281-7559-1/20/$31.00 ©2020 IEEE

received through student transfer requests in order to improve universities decision making and other administration strategies.

II. Related Works

Educational data mining involves the application of automated statistical, machine learning and data mining techniques to extract useful information from voluminous data sets in educational systems [7]-[9] to resolve educational research challenges [10]. The application of educational datamining in higher education environments provides an opportunity to improve quality of education activities [8]. It improves on the quality of decision making processes by providing new and useful knowledge that is hidden from managers in conventional processes across various operations [11]. Educational data mining also helps institutions of higher learning to address knowledge gaps, better decision making, and higher accuracy predictions, advanced planning, maximize educational system efficiency, efficient resource allocation, amongst others [8]. Successes in applications of data mining were achieved in various industries such as biomedical [12], computer auditing [13], climatology [14], sales [15] amongst others. Hence [9] asserts that the higher education domain has the potential of taking advantage of educational data mining techniques to improve the quality of education activities. Romero and Ventura 2007 [9] reviews studies provides in educational data mining studies from 1995 to 2005 showing areas of application and techniques used. Their results are more oriented towards supporting academic activities that involve students and instructors in higher education, than administrative decision making activities.

Although Pe˜na-Ayala 2014 [16]'s work extended on [8] 's work, the study shows most educational data mining approaches implement data mining techniques than extending the data mining field. The need to support the category of administrative processes in education brings a different frontier of problems to be solved. The authors in this study are interested in using educational data mining techniques to improve decision making process, through automated analysis of the large amount of data that faculties encounter when processing campus transfer requests within an educational environment. Manual processing of student requests to change campuses in a multi-campus university as alluded above can be mundane and time consuming, and possibly result in inconsistent decisions. Although literature search by the authors provides no evidence of using data mining on this process and related administrative processes in multi-campus university, we use Latent Dirichlet Allocation (LDA) for topic modelling by extracting latent topics of the reasons submitted with the student transfer requests; and Multinomial Na¨ive Bayes algorithm for prediction in educational data mining.

A. Latent Dirchlet Allocation (LDA) for Topic Modelling

LDA is a generative probabilistic topic modelling approach [17]. It extracts latent semantic topics from large collections of text documents [18]. The key in the LDA model is a premise that words contain strong semantic information from document text [21]. Blei, Carin and Dunson (2010) [19] state that topic modelling algorithms have been successfully used in many areas such as: tourist satisfaction analysis [20] and consumer complaints analysis [21] amongst others.

Research study by Bashri and Kusumaningrum 2017 [22] using LDA and topic polarity world cloud visualization on student's comments by using the best combination of parameters for sentiment analysis and opinion mining proved to outperform Na¨ive Bayes and Logistic Regression in terms of F-Measure by 61%, 54%, and 56% respectively. Shams, Shakey and Faili (2012) [23] experimented on sentiment analysis in Persian language and LDA-based sentiment analysis on the classification of Indonesian News Articles.

Another effort by Kusumaningrum et al. 2016 [15] on semantics and opinion mining in a Bahasa Indonesian language showed a document classification best overall accuracy of 70%. In the current work we use the LDA model to depict topics that exist in the campus transfer data. This data visualization is aimed at assisting management to see what discussions are happening in campus transfer data and get the general idea of the main reasons behind requests for campus transfer. This alleviates the need to go through large datasets manually.

B. Multinomial Na¨ive Bayes for Text Classification

Recent trends have been moving towards the use of neural networks to solve classification problems, harnessing the feature extraction power that comes with neural networks [24]. However, other traditional classification approaches are still relevant because training neural networks requires huge amount of data compared to other classification models like Na¨ive Bayes models. Multinomial Na¨ive Bayes is one of the popularly used text classification techniques because it is easy to implement.

A study conducted by [25] in 2019 compared the performance between two algorithms: Multinomial Na¨ive and Bernoulli Na¨ive Bayes on prediction of sentiments in news articles. Multinomial Na¨ive Bayes proved superior performance compared to Bernoulli Na¨ive Bayes. The uniformly better performance of Multinomial Na¨ive Bayes as compared to Bernoulli Na¨ive Bayes was also observed in an earlier study by (McCallum and Nigam 1998) [26]. The Multinomial model works to determine the frequency of occurences of a term in a document. According to [25] Multinomial Na¨ive Bayes Classifier is given by (1):

$$P(t_k|p) \propto P(p) \prod_{1 \le k \le nd} P(t_k|p), \quad (1)$$

where $P(t_k|p)$ represents the conditional probability that the term t_k occurs in a document of class p. Probability $P(t_k|p)$ is given by (2):

$$P(t_k|p) = \frac{count(t_k|p)+1}{count(t_p)+|V|}. \quad (2)$$

In (2) $count(t_k|p)$ is the number of times the term t_k occurs in the documents which have class, and $count(t_p)$ refers to the total number of tokens contained in documents that have class p. The constants 1 and $|V|$ (total vocabulary in the document) are included as smoothing constants in order to avoid the mishaps in the calculation when the term does not

occur at all in the dataset, $P(p)$ is the prior probability of a document being of class p which is calculated as show in (3):

$$P(p) = \frac{number\ of\ Documents\ of\ class\ p}{total\ number\ of\ documents}. \quad (3)$$

The current work adopted the same approach, features coming from student's written motivation and the class in which a motivation belongs is either a "accept" or "reject". The studies, [27] and [30] presented remarkable results, this has motivated us to utilise two models: Latent Dirichlet allocation model and Multinomial Na¨ıve Bayes model in our experiments.

III. METHOD

We had a two pronged approach to the problem. The first was to do topic modelling on the data. The aim of doing topic modelling was to give management an overview of what is contained in the dataset. For purposes of the current work Latent Dirichlet Allocation model [17] was applied over the campus transfer data.

The second approach was to do prediction. The aim was to establish if the process of accepting or rejecting a transfer request can be automated. Even though Singh, Kumar and Gaur 2019 [25] observed that Multinomial Na¨ıve Bayes outperforms the other Na¨ıve Bayes classifiers in the Indian news articles dataset, in our study we chose to use Multinomial Na¨ıve Bayes because it requires less computation resources compared to Neural networks and support vector machines (SVM) [27]. A separate study would need to be conducted to compare the performance of different models on the campus transfer dataset.

IV. EXPERIMENTS

Our classification experiments were guided by a text classification framework adopted from state-of-the-art in text mining discussed in a survey conducted by Miro´nczuk and Protasiewicz 2018 [28]. The following subsections give details on how we incorporated the state-of-the-art elements in text classification in the current work.

A. Data Preparation for Classification

The campus transfer data contains a motivation that a student submits, and the outcome of the panel whether to accept or reject campus transfer request. The motivation forms part of our features and the outcome is a class. Motivation is text-data written in English language. Silva and Ribeiro 2003 [29] reflects that not all the words in text carry meaning. In the natural language processing field these words are referred to as stop-words. Algorithms were built to remove stop-words from written text [30]. Words like "o", "but" are regarded as stop-words in English language.

The algorithms remove generic stop-words, however, there are domain specific stop-words that would not be removed by the algorithms. We extended the stop-words so that we can remove other words that do not carry meaning as applicable on the campus transfers dataset. These words included names of the university campuses, all South African place names, "academic", and "transfer".

B. Data Segmentation

We obtained the data from university campus transfer records. There were 1391 campus transfer requests made for the period 2013 to 2018. Our data was segmented into 80% training-set and 20% test-set. During training the 80% data was segmented using ten-fold cross-validation scheme.

C. Model Evaluation

The metrics we measured when evaluating the performance of our model were precision, recall, accuracy (4) and area under the curve.

$$accuracy = \frac{\sum(true\ positive)+\sum(true\ negative)}{\sum(Query\ documents)}. \quad (4)$$

where true positive are motivations predicted as 'accept' and their label from the dataset is actually 'accept', and true negative are motivation predicted as 'reject' and their labels on dataset are actually 'reject'.

We chose a threshold range of 0.0 up to 1.0 with 0.1 increments. At each threshold value, we calculated the recall and specificity. Using these values at each threshold we plotted the ROC curve depicted in Fig. 1.

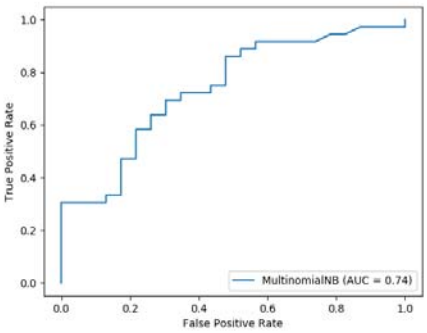

Figure 1. ROC.

V. RESULTS

Our results depict two major insights from the experiments: the topics that are observed in the student data and how the classifier model performs on automating decision making about accept or reject on campus transfer requests.

A. Topics

The few topics observed in the data set were:

- Family support structure: Parents or guardians changing jobs to other provinces requiring the dependant student to change to a nearest campus.
- Change in programme offering, a specialization not found in one campus.
- A student found employment at a different city and would like to continue with studies whilst working.

These topics inform management about academic and socioeconomic reasons that make students to request campus transfers. Fig. 2 shows some of the topics that are observed in the dataset.

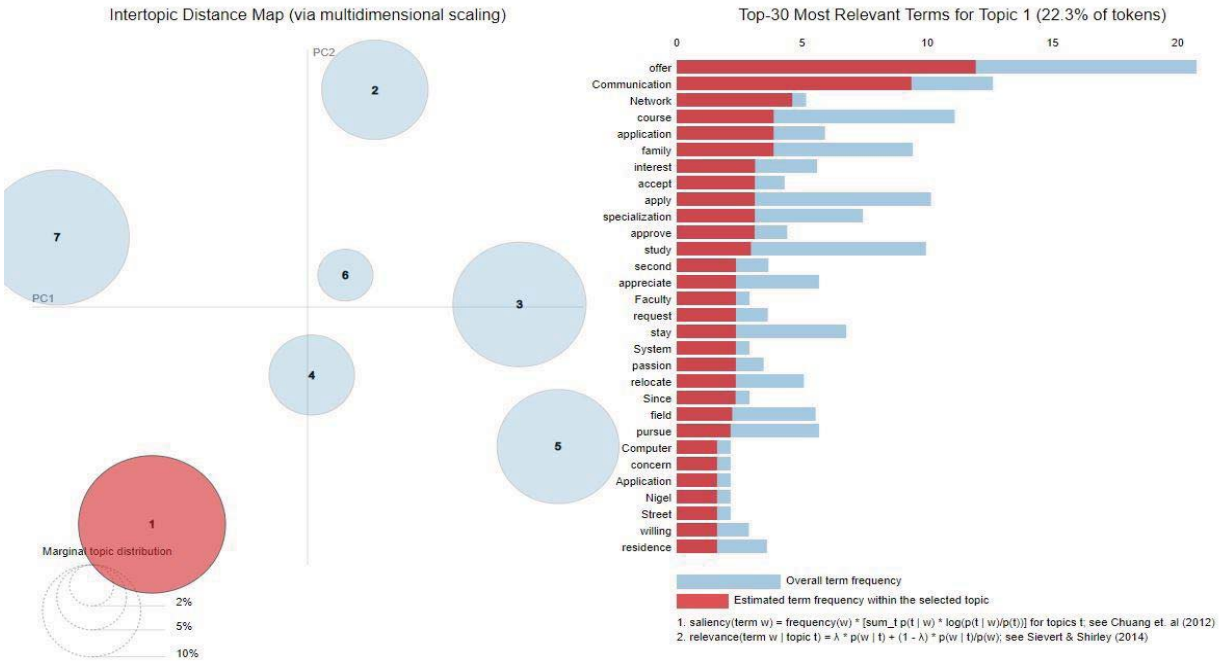

Figure 2. Topics.

B. *Classification Model Performance*

We obtained average accuracy of 66%, average precision 68% and recall 64%. We plotted a ROC curve as depicted in Fig. 1 with AUC of 74%, and a confusion matrix depicted in Fig 3.

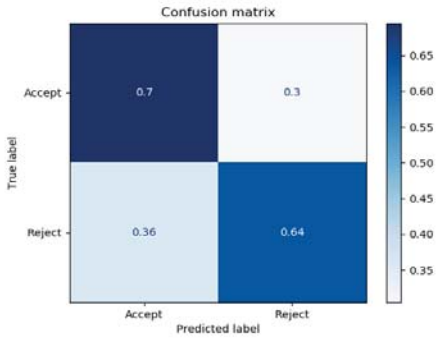

Figure 3. Confusion matrix.

VI. Discussion

Our results show that the manual process of approving campus transfer requests can be automated. A simple algorithm like Na¨ıve Bayes produced better than guessing results. It has a 74% AUC. As reflected in related works, scholars have investigated educational data mining to get insights on things that affect pass rate like submission of assignments and class attendance. We have investigated how to support management by automating decision making on campus transfer requests. Automated process removes human biases to ensure a relatively fair outcome. It also removes the resource burden on institutions where staff need to spend time manually going through each campus transfer request.

The topic modelling depicted a visual presentation of student motivation giving a faster overview of the data. Manual viewing of the data by personnel may otherwise be a difficult task, and important insights may be missed.

VII. Conclusion

A simple Multinomial Na¨ıve Bayes classifier can produce plausible performance in classifying motivation for campus transfer request. This work affirmed that automation of accept and or reject can be achieved on campus transfer request data.

Future works could be on seeking ways in which the accuracy of 66% can be improved by using different classifier models such as SVM. Other improvements can come from harvesting more campus transfer data and train models on a larger dataset than used in the current work.

References

[1] R. Nicolson, "The management of multicampus systems: the practice of higher education," South African Journal of Higher Education, vol. 18, p. 346–358, 2004.

[2] E. C. Lee and F. M. Bowen, "The Multicampus University: A Study of Academic Governance.," 1971.

[3] R. Pinheiro and L. N. Berg, "Categorizing and assessing multi-campus universities in contemporary higher education," Tertiary Education and Management, vol. 23, p. 5–22, 2017.

[4] R. of South Africa, "The Higher Education Act," Act 101 of 1997, 1997.

[5] S. Booysen, G. Godsell, R. Chikane, S. Mpofu-Walsh, O. Ntshingila and R. Lepere, Fees must fall: Student revolt, decolonisation and governance in South Africa, Wits University Press, 2010.

[6] T. Luescher, L. Loader and T. Mugume, "# FeesMustFall: An Internet-age student movement in South Africa and the case of the University of the Free State," Politikon, vol. 44, p. 231–245, 2017.

[7] A. A. Saa, "Educational data mining & students' performance prediction," International Journal of Advanced Computer Science and Applications, vol. 7, p. 212–220, 2016.

[8] N. Delavari, S. Phon-Amnuaisuk and M. R. Beikzadeh, "Data mining application in higher learning institutions," 2008.

[9] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert systems with applications, vol. 33, p. 135–146, 2007.

[10] C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, p. 12–27, 2013.

[11] M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou and A. Tsolakidis, "Improving quality of educational processes providing new knowledge using data mining techniques," Procedia-Social and Behavioral Sciences, vol. 147, p. 390–397, 2014.

[12] J. Han, "How can data mining help bio-data analysis?," in Proceedings of the 2nd International Conference on Data Mining in Bioinformatics, 2002.

[13] T. Y. Wah, M. K. M. Nor, Z. A. Bakar and L. S. Peck, "Data Mining in Computer Auditing," InSITE-"Where Parallels Intersect, 2002.

[14] S. Lee, S. Cho and P. M. Wong, "Rainfall prediction using artificial neural networks," journal of geographic information and Decision Analysis, vol. 2, p. 233–242, 1998.

[15] R. Kusumaningrum, M. I. A. Wiedjayanto, S. Adhy and others, "Classification of Indonesian news articles based on Latent Dirichlet Allocation," in 2016 International Conference on Data and Software Engineering (ICoDSE), 2016.

[16] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert systems with applications, vol. 41, p. 1432–1462, 2014.

[17] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, p. 993–1022, 2003.

[18] M. A. Safi'ie, E. Utami and H. A. Fatta, "Latent Dirichlet Allocation (LDA) Model and kNN Algorithm to Classify Research Project Selection," in IOP Conference Series: Materials Science and Engineering, 2018.

[19] D. Blei, L. Carin and D. Dunson, "Probabilistic topic models," IEEE signal processing magazine, vol. 27, p. 55–65, 2010.

[20] B. Guo, R. Zhang, G. Xu, C. Shi and L. Yang, "Predicting students performance in educational data mining," in 2015 International Symposium on Educational Technology (ISET), 2015.

[21] K. Bastani, H. Namavari and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," Expert Systems with Applications, vol. 127, p. 256–271, 2019.

[22] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization," in 2017 5th International Conference on Information and Communication Technology (ICoIC7), 2017.

[23] M. Shams, A. Shakery and H. Faili, "A non-parametric LDA-based induction method for sentiment analysis," in The 16th CSI international symposium on artificial intelligence and signal processing (AISP 2012), 2012.

[24] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent convolutional neural networks for text classification," in Twenty-ninth AAAI conference on artificial intelligence, 2015.

[25] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," in 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019.

[26] A. McCallum, K. Nigam and others, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, 1998.

[27] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in European conference on machine learning, 1998.

[28] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," Expert Systems with Applications, vol. 106, p. 36–54, 2018.

[29] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Proceedings of the International Joint Conference on Neural Networks, 2003., 2003.

[30] S. Bird, E. Klein and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.