Classification of Research Topics based on Text Classification Algorithm

Aileen U. BALBIDO, Adrianne N. DE GUZMAN, Darien Chelsey P. GALURA, Alfonso Miguel A. ALVARAN, Jason Paul L. VILLANUEVA, Timothy M. AMADO, Aaron U. AQUINO, Cherry G. PASCION, John Carlo V. PUNO, Ira C. VALENZUELA

Technological University of the Philippines-Manila
Manila, Philippines

*ira_valenzuela@tup.edu.ph

Abstract*— This study intends to automatically classify the researches to their respective category using Multinomial Naïve-Bayes algorithm. The datasets are obtained from the University Research and Development Services Office of the Technological University of the Philippines – Manila and numerous journals are also collected from www.ieeexplore.ieee.org. The research titles will first undergo Text Preprocessing method in the form of Stopwords removal, to remove common conjunctions, prepositions, and words such as there, an, the, is, because, by, etc. Different related keywords are encoded for each category of "Education for STEAM", "Environment, Disaster Risk Reduction, Climate Change and Energy", "Food Production and Security", "Health System", "Smart Analytics and Engineering Innovations", "Social Sciences", and "Terrestrial and Marine Resources: Economy, Biodiversity and Conservation". These categories are decided based on the memorandum released by the Commission on Higher Education (CHED) on the Sustainable Development Goals. It has been concluded from this research that Health Systems provide the best output in both precision and recall in predicting titles and by increasing the support in other categories, especially Education for STEAM, they can further improve their categorization.

Keywords*— confusion matrix, multicast classification, multinomial Naïve-Bayes, text classification, text tagging

1. INTRODUCTION

Researches and innovations must be aligned relatively and correspondingly on the nation's goals. The Philippines' Commission on Higher education (CHED) under the pathways of relevance had studied the important developments that the country must focus on and condense the research topics into six platforms namely: "Education for STEAM", "Environment, Disaster Risk Reduction, Climate Change and Energy", "Food Production and Security", "Health System", "Smart Analytics and Engineering Innovations" and "Terrestrial and Marine Resources: Economy, Biodiversity and Conservation" with the addition of "Social Science". Hence, scholars and researchers can have guides for significant topics to develop [1].

However, there are several problems for the platform in terms of classification. The platforms have broad definition and some research topics tends to fall under two or more platforms. For example, the study of behavior of students under STEAM program. The study can be categorized under "Education for STEAM" since it focuses on students undertaking STEAM program, also it can be categorized into "Social Sciences" which focuses on psychological impact to students which is also under the platform of "Health Systems". The vast topics can overlap the other platforms which affects the project granting decision of CHED in approving researches and allocating funds suited for each platform.

To solve the categorization problem, there are numerous solutions that had been studied by various experts. Machine learning have been widely used on text classification [2], [3], [24], [25]. Supervised learning have been used primarily for training labeled data [4]. Under Supervised Learning, Logistic Regression [5], Multinomial Naïve-Bayes [6], Linear Support Vector Machine [7] and Random Forest [8] models are commonly used.

The system to be built will apply text classification using Multinomial Naïve-Bayes algorithm. There are many variations of Naïve-Bayes algorithm but in the study conducted in India, Multinomial Naïve Bayes performs slightly better than Bernoulli Naïve-Bayes for text classification [6]. Also on the study conducted in Germany, Multinomial Naïve-Bayes classifier is more accurate than Random Forest/Decision Tree in topic classification [8]. On the study conducted in Canada, Bayesian models and Support Vector Machine classifier have been compared in the classification of English phrases and SMS text messages and have been evaluated with slight differences in performance [7]. On the study conducted in Nepal, Logistic Regression, Multinomial Naïve-Bayes and Support Vector Machine algorithm have been used and compared in analyzing facts and opinions with insignificant differences in the overall results [5]. Also, there are numerous studies conducted in Multinomial Naïve-Bayes with great accuracy [9]-[12].

The result of this study will be of great benefit to all of the

researchers in an institution for them to easily classify their research titles in their respective category. Also, it will help them to sort their research output and be able to find related studies in fastest way. Since this study will classify the research titles of the faculty researchers automatically, it will also give the field of the researcher they excel at.

2. RELATED STUDIES

R. Wongso, et al. developed a study to find an effective method on how to identify a news report article in the Indonesian language automatically [13]. The study conducted different algorithms such as Singular Value Decomposition (SVD), Term Frequency- Inverse Document Frequency, and compared the result of it on what is best for the selection of the feature. Also, the researchers compared some methods for the classifier such as Support Vector Machine, Multinomial Naïve Bayes, and Multivariate Bernoulli Naïve Bayes. And based on the result of the study, Term Frequency- Inverse Document Frequency and Multinomial Naïve Bayes Algorithm have the best result with the preciseness of 85%.

Since the existing method of identification of news text which is manually done, it is very time consuming and the process of classification has not been accomplished successfully. Thus, a framework to organize a news text and automatically identify it was created by Z. Li [14]. The model of this study was based on Latent Dirichlet Allocation (LDA) and used a topic model that helps to downscale the dimension of a text and get the images of it. Also, the researchers conduct a study about some algorithms to classify multiclass of text, and the method that has the best result to make this study as a text classifier was Softmax Regression. The framework obtained a great performance result in classifying the text and successfully downscaled the dimension of it.

H. Lu, X. Liu and Z. Chen developed a framework based on Multivariate Neural Network Fusion that will automatically classify patent texts with high performance and precision tests [15]. The process of the study starts in representing the segregated patent text by text representation and then, the images of the text will then be extracted through different layers such as Bidirectional-Gated Recurrent Unit (Bi-GRU), attention, word embedded and convolution layer, and softmax layer that will be used for categorizing of the texts. After conducting different layers, the results obtained high performance in recognizing patent text and the model shows high accuracy and preciseness of classifying large amounts of patent texts.

A study that improved the model of data mining for classifying text documents automatically was created by K. Nithya. Classifying such data has a great impact in many monitoring data platforms and a big help in reducing working time [16]. Text identification is an important technique that trains a set of text to acquire the classification framework and it will automatically classify data using the acquired method. This study developed a mining framework based on different abstract analyses such as sentences, documents, and corpus that will study the data acquired. The data will then be analyzed and will extract the image vector for feature selection performance and it will be classified using K-Nearest Neighbor. The model used in this study improves the preciseness of text classification.

Since one of the key technologies for extracting and processing digitized text is text classification which is the process of directly identifying forms of data based on content, F. Miao integrates machine learning models in classifying Chinese news text automatically [17]. Based on the model, it has four processes: text pretreatment, text representation, classifier training, and classification. The study conducted different models in the training part of the framework such as K-Nearest Neighbor, Support Vector Machine (SVM), and Naive Bayes. After comparing the test results of the different algorithms, the researcher concluded that the SVM machine learning model has a high result of performance and accuracy in classifying text.

3. METHODOLOGY

3.1 Conceptual Framework

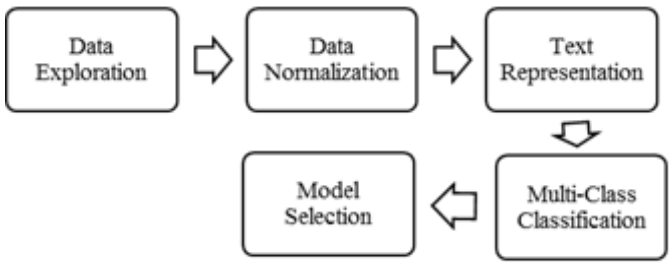

Figure 1: Conceptual Framework

Figure 1 shows the processes performed in this study. The data used as inputs are obtained from Technological University of the Philippines – Manila (TUP-M) and engineered for model selection. Evaluation of the classification is then used after the application of the multiclass Naïve-Bayes classification.

A. Data Exploration

The input data are from the actual records of research of University Research and Development Services (URDS) of TUP-M from year 2017 up to year 2019. Also, additional research titles from numerous journals are collected to maximize the ranges of words for each category. Journal titles that are publicized in www.ieeexplore.ieee.org are manually added in the database with the total of 35,206 research titles.

B. Data Normalization

Different related keywords are encoded for each category of "Education for STEAM", "Environment, Disaster Risk Reduction, Climate Change and Energy", "Food Production and Security", "Health System", "Smart Analytics and Engineering Innovations", "Social Sciences", and "Terrestrial and Marine Resources: Economy, Biodiversity and

Conservation". These categories are decided based on the memorandum released by the Commission on Higher Education (CHED) on the Sustainable Development Goals [1]. The system's database manually collects 5,000 titles for every category and then save them in the database. The titles are randomly splitted with the ratio of 66:33 for training and testing purposes.

C. Text Representation

The text classification model cannot process titles in raw form. It needs to be modified into a more readable representation through numerical features. Stopwords removal are then applied to remove common conjunctions, prepositions, and words such as there, an, the, is, because, by, etc. Then, TF-IDF is applied for feature extraction.

1. Term Frequency- Inverse Document Frequency

Using Python programming, tf-idf is used to extract the unique features of each category. Equation (1) shows the principle behind tf-idf feature extraction.

$$w_{i,j} = TF_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

TFi,j represents the number of occurrences of i in j, then, multiplied to the logarithmic of the quotient of the total number of documents (N) and the number of documents containing i (df_i). [18]

The parameters for the application of feature extraction using Python are as follows: *Sublinear_df* is set to True to apply the frequency's logarithmic form. *min_df* is set to 5, which represents the minimum of 5 titles the word appeared to be retained in the model feature. *norm* is set to 12 to guarantee that all the features had a Euclidian norm of 1. *ngram_range* is set to (1,2) to consider unigrams and bigrams. *stop_words* is set to liststop, a list of stopwords to be removed in the titles.

Out of all 35,206 research titles collected 11,966 features have been extracted. Then, the Chi-squared test is applied to identify the correlated words in each category using Python [19]. Equation (2) shows the principle formula for Chi-squared statistical test.

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{2}$$

Chi-squared is the summation of the ratio of squared difference of the observed frequencies (O) and expected frequencies (E) to the expected frequencies (E).

2. Multiclass Naïve-Bayes Classification

Based on previous works, Multiclass Naïve-Bayes Classification is the most commonly used for classifying research titles. Results from the feature extraction are then used for the Naïve-Bayes Classification [20], [21].

Naïve-Bayes classifier is based on the Bayes' theorem shown in (3). Based on the given formula, a category with the highest probability result will be assigned as the category of the word. There are many variations of the Naïve-Bayes classifier but in this research, Multinomial Naïve-Bayes Classifier is only applied.

$$P(X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_k = x_k) = \frac{n!}{\prod_i x_i} \prod_i p_i^{x_i} \tag{3}$$

This classifier is used with the assumption that the word extracted is independent or no relation to other categories.

D. Model Selection

After applying the Multinomial Naïve-Bayes classifier, the model is then evaluated for its accuracy to some other classifiers such as Linear Support Vector Machine (LSVM), Logistic Regression, and Random Forest Classifier. These models are under supervised learning algorithm to evaluate the system's performance compared to other models for future development of this study.

SVMs are generally used for pattern recognition and clustering [5]. Linear SVM is also discussed and applied in this study [7]. Logistic Regression is ideal classifier with numerical features [5], [22]. While, Random Forest comprised of multiple decision trees to produce predicted results [8], [23].

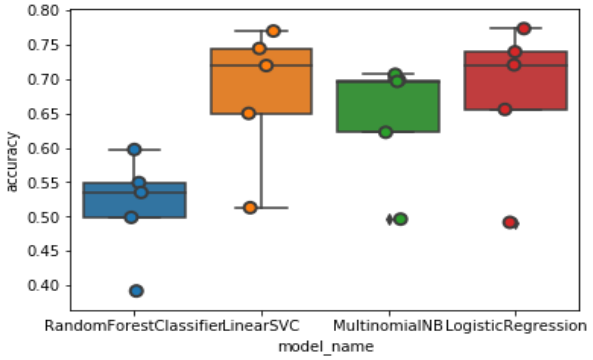

Figure 2: Model Accuracies

Figure 2 shows that the dataset model is more accurate in Linear Support Vector Machine and Logistic Regression by approximately 3%.

4. RESULTS AND DISCUSSION

After conducting the research using Multinomial Naïve-Bayes classifier, these are the obtained results:

Table 1: Comparison of Accuracies for each Category

Category	Precision	Recall	F1-score	Support
Food Production and Security	0.67	0.71	0.69	1648

Environment, Disaster Risk Reduction, Climate Change and Energy	0.70	0.68	0.69	1666
Terrestrial and Marine Resources: Economy, Biodiversity and Conservation	0.63	0.56	0.59	1617
Smart Analytics and Engineering Innovations	0.68	0.62	0.65	1687
Health Systems	0.89	0.91	0.90	1746
Education for STEAM	0.53	0.68	0.60	1588
Social Sciences	0.83	0.72	0.77	1666

As can be seen in Table 1, the Health Systems category provides the highest precision and recall, which when combined produces an F1-score of 90% followed by Social Sciences around 77%. These results are a breakthrough for the research categorization for the Commission on Higher Education (CHED) since there are no existing models that are designed based on their Sustainable Development Goals (SDGs). Education for STEAM gives the worst result with the lowest support of all the data in the experiment.

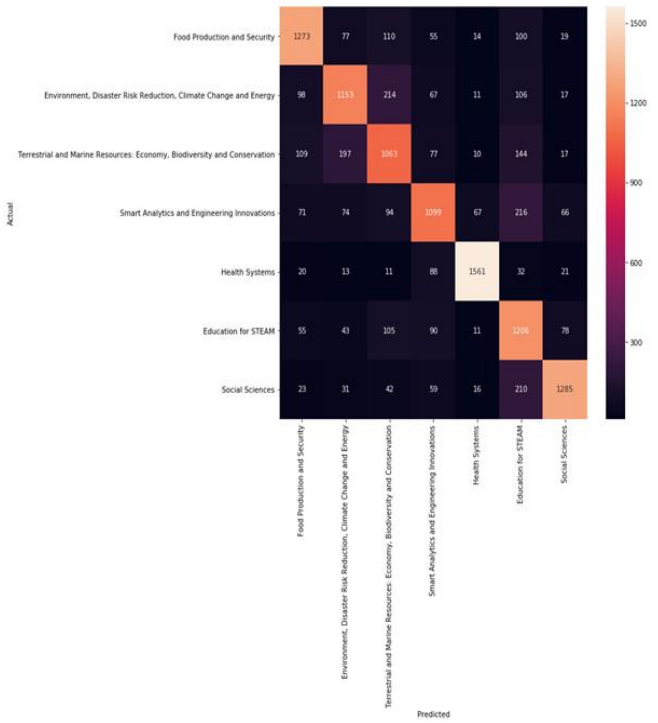

Figure 3: Confusion Matrix

Most of the titles are predicted correctly as shown in the confusion matrix in Figure 3. However, multiple misclassifications still occurred in the model due to the inconsistencies of words for each label that are common between two or more categories.

5. CONCLUSION

It has been concluded from this research that Health Systems provide the best output in both precision and recall in predicting titles and by increasing the support in other categories, especially Education for STEAM, they can further improve their categorization. Accuracy of different supervised machine learning composed of Multinomial NB, Linear SVC, Logistic Regression and Random Forest on research titles were analyzed and evaluated. Although, Linear SVC and Logistic Regression perform slightly better than Multinomial NB by approximately 3%. The difference in accuracies is insignificant as Multinomial NB provides an overall 70% accuracy which shows that this model does not varied much on the given dataset.

FUTURE WORK

For future work, the proponents intend to collect data directly on the CHED database to use in the dataset training in both English and Filipino language and correctly labeled with their respective category with the help of the research experts on that institution to ensure the results are accurate. The utilization of more classifier models will likewise give more credible results.

REFERENCES

1. P. B. Licuanan, **Pathways To Equity, Relevance and Advancement in Research, Innovation, and Extension in Philippine Higher Education**, *official memorandum*, *Commission on Higher Education*, Diliman, QC, Philippines, 2016. [Online]. Available: https://ched.gov.ph/wp-content/uploads/2017/10/CMO-52-s.-2016.pdf
2. F. Miao, P. Zhang, L. Jin and H. Wu, **Chinese News Text Classification Based on Machine Learning Algorithm**, *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, 2018, pp. 48-51, doi: 10.1109/IHMSC.2018.10117.
3. A. Poornima and K. S. Priya, **A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques**, *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2020, pp. 493-496, doi: 10.1109/ICACCS48705.2020.9074312.
4. A. A. Shah and K. Rana, **A Review on Supervised Machine Learning Text Categorization Approaches**, *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, Kottayam, India, 2018, pp. 1-6, doi: 10.1109/ICCSDET.2018.8821134.
5. S. Regmi, B. K. Bal and M. Kultsova, **Analyzing facts and opinions in Nepali subjective texts**, *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Larnaca, 2017, pp. 1-4, doi: 10.1109/IISA.2017.8316445.
6. G. Singh, B. Kumar, L. Gaur and A. Tyagi, **Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification**, *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, London, United Kingdom,

2019, pp. 593-596, doi: 10.1109/ICACTM.2019.8776800.

7. J. Maier and K. Ferens, **Classification of english phrases and SMS text messages using Bayes and Support Vector Machine classifiers**, *2009 Canadian Conference on Electrical and Computer Engineering*, St. John's, NL, 2009, pp. 415-418, doi: 10.1109/CCECE.2009.5090166.
8. M. A. Rahman and Y. A. Akter, **Topic Classification from Text Using Decision Tree, K-NN and Multinomial Naïve Bayes**, *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-4, doi: 10.1109/ICASERT.2019.8934502.
9. D. E. Cahyani and K. A. P. Nuzry, **Trending Topic Classification for Single-Label Using Multinomial Naive Bayes (MNB) and Multi-Label Using K-Nearest Neighbors (KNN)**, *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2019, pp. 547-552, doi: 10.1109/ICITISEE48480.2019.9003944.
10. Z. Tan, Y. Zhang, C. Zhang, R. Huang, P. Lei and X. Duan, **Research on The Text Emotion of Multinomial Naïve Bayes Integration Algorithm**, *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Chongqing, China, 2019, pp. 107-111, doi: 10.1109/IMCEC46724.2019.8984049.
11. U. Pujianto, M. F. Hidayat and H. A. Rosyid, **Text Difficulty Classification Based on Lexile Levels Using K-Means Clustering and Multinomial Naive Bayes**, *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, 2019, pp. 163-170, doi: 10.1109/ISEMANTIC.2019.8884317.
12. F. Gürcan, **Multi-Class Classification of Turkish Texts with Machine Learning Algorithms**, *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, 2018, pp. 1-5, doi: 10.1109/ISMSIT.2018.8567307.
13. R. Wongso, F.A. Luwinda, B.C. Teisnajaya, O. Rusli and Rudy , **News Article Text Classification in Indonesian Language**, *Procedia Computer Science*, vol. 116, pp. 137-143, 2017.
14. Z. Li, W. Shang and M. Yan, **News Text Classification Model Based on Topic Model**, *2016 International Conference on Computer and Information Science (ICIS)*, 2016.
15. H. Lu, X. Liu and Z. Chen, **A Patent Text Classification Model Based on Multivariate Neural Network Fusion**, *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)* , pp. 61-65, 2019.
16. L. Nithya, P.C.D. Kalaivaani and R. Thangarajan, **An enhanced data mining model for text classification**, *2012 International Conference on Computing, Communication and Applications*, pp. 1-4, 2012.
17. F. Miao, P. Zhang, L. Jin and H. Wu , **Chinese News Text Classification Based on Machine Learning Algorithm**, *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 1-4, 2018.
18. A. Barysevich. (October 25, 2019). **TF-IDF: Can it really help your SEO?**. Retrieved from https://www.searchenginejournal.com/tf-idf-can-it-really-help-your-seo/331075/
19. C. Hoang, C. Le and S. Pham, **Improving the Quality of Word Alignment by Integrating Pearson's Chi-Square Test Information**, *2012 International Conference on Asian Language Processing*, Hanoi, 2012, pp. 121-124, doi: 10.1109/IALP.2012.44.
20. G. Qiang, **Research and improvement for feature selection on naive bayes text classifier**, *2010 2nd International Conference on Future Computer and Communication*, 2010.
21. P. Liu, H. Yu, T. Xu and C. Lan, **Research on archives text classification based on Naive bayes**, *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2017.
22. Berwick, R. (2003). **An Idiot's guide to Support vector machines** (SVMs). *Retrieved on October*, *21*, 2011.
23. Goyal, K. (2020). **6 Types of Supervised Learning You Must Know About in 2020**. Retrieved from https://www.upgrad.com/blog/types-of-supervised-learning/
24. J. P. D. Delizo, M. B. Abisado and M. I. P. De Los Trinos, **Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes**, *International Journal of Advanced Trends in Computer Science and Engineering(IJATCSE)*, Vol. 9, no. 1.3, pp. 408-412, 2020.
25. M. P. Abraham and Udaya Kumar Reddy K R, **Feature Based Sentiment Analysis of Mobile Product Reviews using Machine Learning Techniques**, *International Journal of Advanced Trends in Computer Science and Engineering(IJATCSE)*, vol. 9, no. 2, pp. 2289 – 2296, March - April 2020.