Higher Education and Oriental Studies (HEOS) – Vol 3(1): 74 – 86
www.heos.asia/ojs
e-ISSN 2785-9118
http://doi.org/10.54435/heos.v3i1.95

Cyberbullying Detection on Social Media

HUANG HUANG*
University of Malaya, Malaysia
erichuangemail@163.com

DONGKAI QI
Illinois Institute of Technology
dqi2@hawk.iit.edu

**Corresponding Author*

Date Received: 2 January 2023 Date Accepted: 22 January 2023 Date Published: 28 February 2023

ABSTRACT

Cyberbullying is an act of bullying with the use of technology as a medium. The exponential growth of social media has exposed many people especially young people to the risk of being target of cyberbullying. With the help of machine learning, we can detect language patterns from the cyberbullying posts and develop a model to detect cyberbullying content automatically. This project uses data from Formspring.me, a question-and-answer formatted website. The data is labeled by Amazon's Mechanical Turk, a web service. After data preparation and pre-processing, Scikit-learn, a python library is used to train a model to classify if cyberbullying is present in the post by recognizing bullying content based on its insult words as features. Four machine learning techniques, namely, logistic regression, decision tree, random forest and support vector machine are used to train the model. Consequently, all four algorithms achieve more than 78% of accuracy in identifying the true positive where support vector machine achieves the best performance with a score of 87% of accuracy in identifying the true positive. Furthermore, the context of cyberbullying post is investigated and categorised into five categories, namely, swearing, abusive, sexism, vulgar, and racism. Lastly, a study on the severity level and the context of cyberbullying content analyzed swearing as the most frequent category of cyberbullying in this dataset.

Keywords: c*yberbully; machine learning; social media; natural language processing*

INTRODUCTION

BACKGROUND

Cyberbullying is the act of bullying that conducts on digital devices such as smart phones, computers, tablets, and so on, where internet is accessible. Common means of cyberbullying can occur are text, forum discussion and the mainstream social media with the condition that people are able to view, participate, or share their opinion. Given the rise of industry revolution, social media is getting more mainstream than ever, social media users are growing exponentially including people from all walks of life. Sending, posting, or sharing potentially harmful content, false information, or leaking private information of others that would cause embarrassment or humiliation are examples of cyberbullying that occur in our daily life, especially in the abstract world of social media. Although laws against cyberbullying have been established, criminals are way too numerous to tackle or being brought to justice to face charges. This phenomenon leaves people especially young people exposed to the risk of being the target of cyberbullying. Tragically, in more extreme circumstances where young victims are vulnerable and under protection of no guardians, they consequently commit suicide as a way of escaping from cyberbullying.

PROBLEM STATEMENT

The number of social media company has risen but approaches to prevent cyberbullying have not been well-equipped by the application, leaving cyberbullies to roam across the internet and attacking innocent people with impunity. With this being said, prevention measures and law protection are not keeping up with the criminal records of cyberbullying.

OBJECTIVE

There are numerous remarkable research conducted to overcome cyberbullying specifically on detection of cyberbullying with the help of machine learning to study the sentiment and contextual features from the conversation or medium. This project will tackle this problem by performing a classification on social media data to detect the act of cyberbullying. The objectives of this project are as follows: First, to classify posting susceptibility to cyberbullying with certain degree of confidence; Second, to investigate the context of discrimination on posting susceptible to cyberbullying; Finally, to study the severity of cyberbullying based on the context of cyberbullying.

DATA COLLECTION

This section will describe the details of data collection and data labelling in this project.

DATA ORIGIN

The origin of data in this project is a question and answer-based website, Formspring.me, where users openly invite others to ask and answer questions. The feature of anonymity option which allow users to post questions anonymously to other user's profile. This dataset is provided in Kaggle (Aman,2022) which represented 50 ids from Formspring.me that were crawled in Summer 2010. The columns contain the user id, post, question of post, answer of post, and response of labeling cyberbullying, bully words, and the severity level in a range of 10 from three different respondents (Cruz,2022).

DATA LABELING

The cyberbullying label work was determined by Amazon's Mechanical Turk service, an online marketplace that allows requestors to post tasks that are then completed by paid workers. The labeling work was conducted according to the following question: First, does this post contain cyberbullying? (Yes, 1 or No, 0) Second, on a scale of 1 (Mild) to 10 (Severe), how bad is the cyberbullying in this post? (0 for no cyberbullying) Finally, what words or phrases in the post(s) are indicative of cyberbullying? (NaN for no cyberbullying)

DATA PREPARATION

This section will describe the data preparation process before deploying the model.

DATA CLEANING

The first cleaning process comprises dropping any missing data from the post, questions of post, or answer of post. Then, the convert of "Yes" and "No" value to integer "1" and "0" for conveniences of further processing. Next, the dataset is filtered so that the response of '0' from all respondents agreed upon are accepted and at least two responses of "1" from respondents are accepted to be "1". This leads to the creation of a new column "overall ans" which conclude the filter step above aggregating three responses into one overall decision (Muguro,2022). The

severity from three respondents are averaged to a new column called 'overall severity' for aggregating into one decision value (Bhattacharjee, 2022).

DATA PREPROCESSING

The preprocessing process occurs with four common text preprocessing approaches. First, tokenizing the data. This is a process where strings are tokenized or split into a list of tokens or small parts in order for the machine to be able to understand and capture the pattern (Abdoun,2022). In this step, text are split based on whitespace and punctuation with the help of a machine learning module, punkt, which is a pre-trained model that tokenize words and sentences. Next, the text is normalized to convert to this canonical form. This process helps to group words presented in different forms but with similar meanings. One of the famous techniques of normalization, lemmatization is used to analyze the structure of the word and its context to convert to a normalized form with the help of a module that uses a lexical database for English language that helps the script determine the base word, wordnet (Cimiano,2022). Lastly, the noise is removed from the data with the help of regular expression and a stopwords resource from the python module. Examples of noise include hyperlink, numbers, stop words, and so on. After the pre-processing step, the data comprises 11609 values where only 776 values (nearly 7% of data) are positive values (cyberbullying, '1'). A severe imbalanced property of data is observed, while extra care is essentially needed for modeling and analyzing the reading of the result (Stefan,2022).

MODEL PREPARATION

This section describes the preparation for pre-processed text to be ready to split into training, validating, and testing purposes.

RESAMPLING

Due to the severe imbalanced property of the data, this project introduces the resampling technique to minimize this potential disadvantage of data (Kang Hong,2022). Resampling is a technique that consists of drawing repeated samples from the original data samples. The resampling approaches conducted are oversampling the minority class, which is a type of data augmentation for the minority class and is referred to as Synthetic Minority Oversampling technique (SMOTE) described by Nitesh Chawla (Fatima,2022). This paper on SMOTE also suggested combining SMOTE with random undersampling of the majority class, which this project takes as one of the approaches. With the help of a python library "imbalancedlearn," the minority class (positive value, "1" as cyberbullying) is oversampling to have 10 percent of the number of examples of the majority class, whereas the majority class (negative value, "0,", as no cyberbullying) is undersampling to reduce the number of examples in the majority class to have 50 percent more than the minority class (Vo Nhi, 2021).

FEATURE EXTRACTION

This project illustrates a bag-of-words model in which a way of extracting features from text for use in modelling where bag-of-words is a representation of text that describes the occurrence of words within a document (Park,2021). In this step, a collection of text posts is converted to matrix of token counts and a creation of words vocabulary. Due to the fact that large counts of certain words may not be meaningful in the encoded vectors, word frequencies are calculated by one of the popular approaches, Term Frequency – Inverse Document (TF-IDF), which summarizes how often a given word appears within a document and downscales

words that appear frequently across documents to capture meaningful patterns (Lathabai,2021). This process is conducted with the help of a python library, scikit-learn.

MODEL EVALUATION

This section discusses the machine learning techniques used, and analyzing the statistical results indicate the learning experiment. The Python language will be used to evaluate the model. The machine learning techniques or algorithms conducted are logistic regression, decision tree, random forest and support vector machine. The model evaluation is based on three set of data: training set (70%), validation set (20%), and testing set (10%). Due to the fact that the dataset is imbalanced, the model evaluation will focus heavily on the reading of true positives, false negatives, and recalls, as the accuracy is initially high even if the prediction of all negatives is made. As shown in code file, the accuracy with all negative predictions (as no cyberbullying), the model will still achieve a high accuracy of 94.14%. The usual way of interpretation will be discarded and focus on how well the algorithm may capture the positive value (as cyberbullying) and how bad the model missed out to capture the positive value(Schick,2021).

LOGISTIC REGRESSION

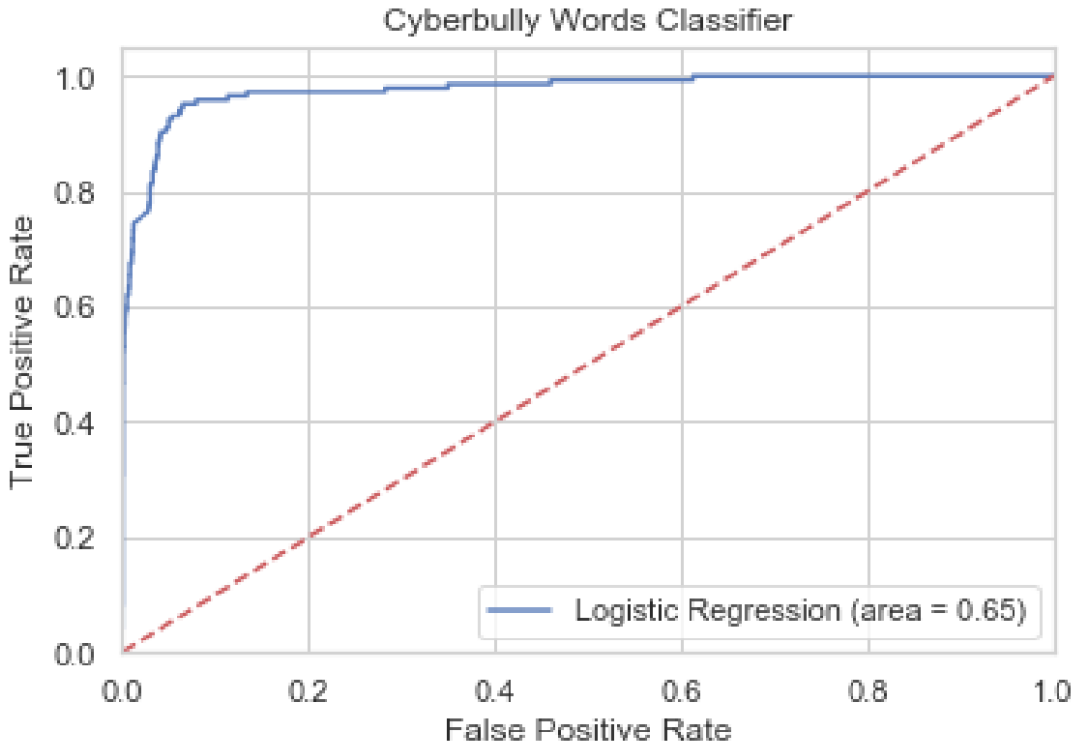
Logistic regression is a supervised learning classification algorithm that can be used to predict the probability of3 a target variable. The nature of the target or dependent variable is dichotomous, which means that there would only be two possible classes. The dependent variable is binary in nature, having data coded as either 1 (success/yes) or 0 (failure/no). To achieve the main objective of this project, a logistic regression is used as one of the models to predict the if cyberbullying is presented in social media posts. Logistic regression is refined with a threshold of 0.4 probability in a way of saying, a post has 0.4 probability of classifying as cyberbullying (low threshold). This is due to the fact that, classifying a cyberbullying post as negative (as no cyberbullying) is more costly that classifying a non-cyberbullying post as positive (as cyberbullying). The algorithm will be less probable to miss positive values (as cyberbullying). The following results are running the model:

FIGURE 1. Report of LG

```
Accuracy:
 0.9758828596037898

Confusion Matrix:
 [[1077   16]
 [  12   56]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.78      0.82      0.80        68

    accuracy                           0.98      1161
   macro avg       0.88      0.90      0.89      1161
weighted avg       0.98      0.98      0.98      1161
```

The result from the testing set shows that the model achieved a high recall of 82%, F1-score of 80%, and precision of 78% for positive value (‘1’, as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value (‘0’, as no cyberbullying). For reporting purposes, the overall accuracy is 97.59%. As for interpretations of these results, this model did churn out an average result even with good accuracy because the recall is 0.61. A good model should have a recall near to 1. We managed to classify 2165 words correctly, which is true positive and 89 words incorrectly, which is false negative. Logistic regression achieves a good performance in classifying post with cyberbullying.

FIGURE 2. ROC Curve



We managed to carve out the AUC score for this model, which turned out to be 0.65. Having it close to 0.5 making this classifier as slightly unsuitable in achieving our target as the closer the score to 1, the better the classifier. We would decide to move on with another classifier that possesses better results and quality than the logistic regression model does.

SUPPORT VECTOR MACHINE

The support vector machine (SVM) is a technique that can be used for classification and regression problems. To achieve the main objective, SVM is used to perform classification by finding the hyper-plane that differentiates the two classes. The motivation to consider SVM as one of the approaches is due to its property of being an optimal margin classifier. This algorithm will try to find a decision boundary that maximizes the geometric margin and results in a very confident set of predictions on modelling. In particular, SVM will capture the maximum distance between post with cyberbullying and post without cyberbullying in order to form the optimal decision boundary line. Therefore, a linear SVM is chosen for modeling to classify if cyberbullying is present in posts. The result from the testing set shows that the model achieves a relatively high recall of 85%, F1-score of 86%, and precision of 87% for positive value (‘1’, as cyberbullying); as well as an expected high reading precision of 99%, recall of 99%, and F1-score of 99% for negative value (‘0’, as no cyberbullying), as well as an overall accuracy of 98.36%. SVM achieves great performance in classifying posts with cyberbullying and trained an extremely reliable model.

FIGURE 3. Report of SVM

```
Accuracy:
 0.983634797588286

Confusion Matrix:
 [[1084    9]
 [  10   58]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.87      0.85      0.86        68

    accuracy                           0.98      1161
   macro avg       0.93      0.92      0.93      1161
weighted avg       0.98      0.98      0.98      1161
```

DECISION TREE

The decision tree is a decision support tool that uses a tree-like graph, which comprises the possibilities of event outcomes, information gain, and resource costs (Mitha,2021) It is one of the algorithms that contain conditional control statements. Tree-based methods empower predictive models with advantages of high accuracy and are easily interpreted. Therefore, a decision tree classifier is one of the approaches to classify if cyberbullying is present in a post based on its conditional properties. The result from the testing set shows that the model achieves a relatively high recall of 79%, F1-score of 81%, and precision of 82% for positive value (‘1’, as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value (‘0’, as no cyberbullying), as well as an overall accuracy of 97.76%. The decision tree classifier achieves a relatively good performance in classifying post with cyberbullying.

FIGURE 4. Report of decision tree

```
Accuracy:
 0.9776055124892334

Confusion Matrix:
 [[1081   12]
 [  14   54]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.82      0.79      0.81        68

    accuracy                           0.98      1161
   macro avg       0.90      0.89      0.90      1161
weighted avg       0.98      0.98      0.98      1161
```

RANDOM FOREST

The Random Forest Classifier is used to train our model to classify if posts contain cyberbullying. Random Forest is one of the most known and effective machine learning algorithms in data mining and particularly in text classification. A difference between Random Forest and Decision Tree is that the decision tree is built on an entire dataset using all features, whereas Random Forest randomly selects specific features to build multiple decision trees and obtain the average result. Each of the decision tree consist class prediction and the most votes of the prediction will be selected as algorithm model. An additional feature is it uses bootstrap sampling to extract a number of training samples and group the features from the original training set, establishes a plurality of unpruned decision trees (Lee Joo Yun,2021), then combining the decision trees to form a random forest model. Consequently, the randomness increases the diversity of decision trees and makes the resulting integrated model have better classification performance. Due to the feature of random forest, it is considered as one of the approaches to capture the pattern of post with cyberbullying. The result from the testing set shows that the model achieves a relatively high recall of 78%, F1-score of 82%, and precision of 87% for positive value (‘1’, as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value (“0”, as no cyberbullying), and an overall accuracy of 98.02%. The decision tree classifier achieves a relatively good performance in classifying posts with cyberbullying(Du Jin,2022). Thus, the Random Forest Classifier is pretty accurate in classifying cyberbullying posts.

FIGURE 5. Report of random forest

```
Accuracy:
 0.9801894918173988

Confusion Matrix:
 [[1085    8]
 [  15   53]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.87      0.78      0.82        68

    accuracy                           0.98      1161
   macro avg       0.93      0.89      0.91      1161
weighted avg       0.98      0.98      0.98      1161
```

SUMMARY

Last but not least, a conclusion to model evaluation is SVM achieves the best performance overall on this dataset with the highest recall which considered as one of the main performance metrics as this metric captures how bad the post with cyberbullying is not captured, and the highest overall accuracy(Yin,2022). This means that it correctly classifies the most post with cyberbullying and post with no cyberbullying.

DISSECTION: CYBERBULLYING

This section will discuss the result of investigation of the context of discrimination in cyberbullying post (second objective) and the level of severity of cyberbullying content based on its context (third objective).

CONTEXT OF DISCRIMINATION

In this step, the feature of cyberbullying is extracted based on the potential bully words in cyberbullying post. Furthermore, the bully words are categorized into five categories and are selected based on its top five most frequent word counts. The categories are swearing, abusive, vulgar, sexism, and racism.

1. Swearing is an expression of strong feelings toward something, generally considered to be language that is strongly impolite, rude, or offensive.

2. Abusive is the use of remarks intended to demean, humiliate, mock, insult, or belittle people.

3. Vulgar is a practice that makes explicit and offensive references to sex or bodily functions.

4. Sexism is also known as prejudice, stereotyping, or discrimination on the basis of sex, typically against woman.

5. Racism is also known as prejudice, discrimination, or antagonism directed against a person or people on the basis of their membership of a particular racial or ethnic group. Example of bully words are categorized into a table for better alignment, as shown below:

FIGURE 6. Bully words categories

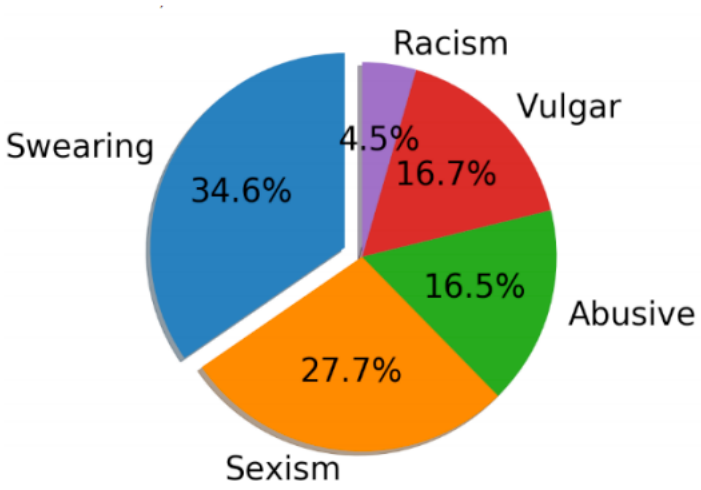
No.	Category									
	Swearing		Abusive		Vulgar		Sexism		Racism	
	word	count	word	count	word	count	word	count	word	count
1	fuck	156	stupid	33	ugly	46	bitch	158	nigga	27
2	shit	53	fake	30	dick	28	whore	15	racist	2
3	damn	15	dumb	24	pussy	27	faggot	12	aussie	1
4	wtf	15	stfu	15	fat	13	gay	8	australian	1
5	asshole	8	retard	16	virgin	5	slut	5	chinese	1

A wordcloud visualization technique is used to observe the result better:

FIGURE 7. Word cloud of bully words

Overall, swearing comprises 34.6%, sexism consists of 27.7%, abusive with 16.5%, vulgar with 16.7%, and racism with 4.5% out of all cyberbullying posts. It appears that swearing and sexism are two most common practices in cyberbullying in the dataset. A pie chart is used for better visualization, as shown below:

FIGURE 8. Pie chart of bully words



SEVERITY OF CYBER BULLYING

This step will examine the severity level of cyberbullying based on the context of cyberbullying. Investigation is conducted to examine the most frequent category of bully words based on its severity level. There are 10 severity level where ranging from 1 (mild) to 10 (severe). For better visualization, the result is tabulated below:

FIGURE 9. Severities of cyberbullying

Severity	Top Category
10	Swearing (fuck, shit, badass)
9	Swearing (fuck, shit, motherfucker)
8	Sexism (bitch, hoe, whore)
7	Sexism (bitch, hoe, sexy)
6	Sexism (bitch, hoe, faggot)
5	Sexism (bitch, faggot, gay)
4	Sexism (bitch, gay, lesbian)
3	Sexism (bitch, whore, slut)
2	Sexism (bitch, gay, hoe)
1	Abusive (fake, stupid, hate)

According to the result, swearing is most widely used in high severity, then followed by sexism and lastly abusive in low severity in this dataset. Swearing is the most common practice in cyberbullying and ranks as the highest severity level according to the dataset.

CONCLUSION

This project uses a language-based method of detecting cyberbullying by classification. By capturing the pattern of cyberbullying based on the insult words as features, we are able to identify 98.4 correctly.

REFERENCES

Abdoun, N., & Chami, M. (2022). Automatic Text Classification of PDF Documents using NLP Techniques. INCOSE International Symposium,2022(12).

Aman, A., & Reji, D., J. (2022). EnvBert: An NLP model for Environmental Due Diligence data classification. *Software Impacts, 10*(04).

Bhattacharjee, S., Delen, D., Ghasemaghaei, M., Kumar, A., & Ngai, W., T. (2022). Business and government applications of text mining & Natural Language Processing (NLP) for societal benefit: Introduction to the special issue on "text mining & NLP". *Decision Support Systems, 2022*(11).

Cimiano, P., Armaselu, F., Apostol, E. S., Khan, A. F., Liebeskind, C., McGillivray, B., & Dojchinovski, M. (2022). LL(O)D and NLP perspectives on semantic change for humanities research. *Semantic Web, 2022*(6).

Cruz, L. A., Coelho da Silva, T. L., Magalhães, R. P., Melo, W. C. D., Cordeiro, M., de Macedo J. A. F., & Zeitouni, K. (2022). Modeling Trajectories Obtained from External Sensors for Location Prediction via NLP Approaches. *Sensors, 2022*(19).

Du, J. (2022). Social Networking Media in Higher Education: A Review. *Higher Education and Oriental Studies, 2022*(4).

Fatima, R., Samad, S. N., Riaz, A., Ahmad, S., El Affendi, M. A., Alyamani, K. A. Z., & Latif, R. M. Amir. (2022). A Natural Language Processing (NLP) Evaluation on COVID-19 Rumour Dataset Using Deep Learning Techniques. *Computational Intelligence and Neuroscience, 2022*(4).

Kang, H., Zhang, J., & Kang, J. (2022). Analysis of Online Education Reviews of Universities Using NLP Techniques and Statistical Methods. *Wireless Communications and Mobile Computing, 2022*(12).

Lathabai, H. H., Nandy, A., & Singh, V. K. (2022). Institutional collaboration recommendation: An expertise-based framework using NLP and network analysis. *Expert Systems with Applications, 2022*(11).

Lee, J. Y. (2021). Ontology-Based Natural Language Processing of Social Media Data in the Assessment of Health Information Sought During Pregnancy. *Studies in health technology and informatics, 2021*(2).

Mitha, S., Schwartz, J., Cato, K., Woo, K., Smaldone, A., & Topaz, M. (2021). Natural Language Processing of Nursing Notes: A Systematic Review. *Studies in health technology and informatics, 2021*(3).

Muguro, J., Njeri, W., Matsushita, K., & Sasaki, M. (2022). Road traffic conditions in Kenya: Exploring the policies and traffic cultures from unstructured user-generated data using NLP. *IATSS Research, 2022*(3).

Park, W. H., Shin, D. R., & Qureshi, N. M. F. (2021). Effective Emotion Recognition Technique in NLP Task over Nonlinear Big Data Cluster. *Wireless Communications and Mobile Computing, 2021*(7).

Schick, T., Udupa, S. & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics, 2021*(2).

Stefan, R. & Michael, H. (2021). Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science. MORGAN & CLAYPOOL.

Vo Nhi, N. Y., Vu Quang, T., Vu Nam, H., Vu Tu, A., Mach Bang, D., & Xu, G. (2021). Domain-specific NLP system to support learning path and curriculum design at tech universities. *Computers and Education: Artificial Intelligence(prepublish), 2021*(6).

Yin, Z. (2022). A Review on the Cause, Advantages, Challenges, and Future of Media Convergence. *Higher Education and Oriental Studies,2022*(5).

ABOUT THE AUTHORS

Huang Huang received BSc. in Software engineering from WuHan Textile University and MSc. In Data Science from University Malaya (UM) in 2021. He is a PhD candidate in University Malaya (UM) and his major research direction includes natural language processing and fuzzy inference system. He has 2-year experience in software testing.

Qi Dongkai received a Master of Science degree in Artificial Intelligence from the Illinois Institute of Technology. He is a PhD student at the University Malaya (UM) since 2019. His research areas of interest are computer vision, natural language processing and artificial intelligence project management. He has 13 years of experience in financial data project management in commercial banks and the People's Bank of China, and 3 years of experience in artificial intelligence software project development.